ISSN 2007-9737

The Big Five: Discovering Linguistic Characteristics that Typify Distinct Personality Traits across Yahoo! Answers Members

Nicolás Olivares[2], Luz María Vivanco[3], Alejandro Figueroa[1,2]

[1] Departamento de Ciencias de la Ingeniería, Facultad de Ingeniería, Universidad Andres Bello, Santiago, Chile

[2] Escuela de Ingeniería Informática, Universidad Diego Portales, Santiago, Chile

[3] Escuela de Ingeniería Industrial, Universidad Diego Portales, Santiago, Chile

{nicolivares, luzmavivanco}@gmail.com, alejandro.figueroa@unab.cl

Abstract. In psychology, it is widely believed that there are five big factors that determine the different personality traits: Extraversion, Agreeableness, Conscientiousness and Neuroticism as well as Openness. In the last years, researchers have started to examine how these factors are manifested across several social networks like Facebook and Twitter. However, to the best of our knowledge, other kinds of social networks such as social/informational question-answering communities (e.g., Yahoo! Answers) have been left unexplored. Therefore, this work explores several predictive models to automatically recognize these factors across Yahoo! Answers members. As a means of devising powerful generalizations, these models were combined with assorted linguistic features. Since we do not have access to ask community members to volunteer for taking the personality test, we built a study corpus by conducting a discourse analysis based on deconstructing the test into 112 adjectives. Our results reveal that it is plausible to lessen the dependency upon answered tests and that effective models across distinct factors are sharply different. Also, sentiment analysis and dependency parsing proven to be fundamental to deal with extraversion, agreeableness and conscientiousness. Furthermore, medium and low levels of neuroticism were found to be related to initial stages of depression and anxiety disorders.

Keywords. Big five, user analysis, personality analysis, natural language processing, community question answering.

1 Introduction

Community Question Answering platforms are social networks, where their members socialize and share their knowledge by posting and answering questions (e.g., opinions, word-of-mouth tips and facts). One of the primary motivations for utilizing these classes of services has to do with the fact that community members can ask personalized questions that will get answers tailored to their specific need.

By examining their activity, it becomes crystal clear that not all community fellows exhibit the same pattern of behaviour when participating in this sort of system. In fact, each member behaves and plays a distinct role in consonance with his/her interests, expertise, and personality. For instance, some users are more leaned to post questions than to provide answers, whereas other users compete to gain higher rewards and/or to be granted as many best answers as possible.

As a means of personalizing their service, thus enhancing user experience, it is critical for these sites to undertake a comprehensive assessment of the expertise, topic of interest and personality traits of their members. For example, this knowledge can cooperate on solving the “cold start problem”, that is to say on reducing the delay between posting

time and the arrival of good answers, via finding potential answerers that best match. By bridging this gap, these platforms keep their vibrancy and attractiveness as well as capture more attention [6, 24, 28, 40]. In fact, a profound understanding of community members is not only pertinent to route open questions to potential answerers, but also to personalize the display of content.

To the best of our knowledge, what characterizes the different personalities expressed in these communities has been largely unexplored, so far. More specifically, our work examines different linguistic characteristics that typify the distinct personality traits reflected in Yahoo! Answers. Namely, interpreted in the light of the Big Five factors, which are widely believed to define personality. More precisely, this is an interdisciplinary study, which contribution is three-fold:

1. Instead of asking Yahoo! Answers members to volunteer for answering the Big Five test[1], we conduct a discourse analysis based on a decomposition of the test into 112 descriptors, measuring each factor according to the five-likert scale.

2. This corpus is then utilized for building multi-class discriminant models to automatically identify the degree of each factor across community members. In so doing, we evaluated fifteen distinct supervised machine learning algorithms including Bayes, Maximum Entropy, Support Vector Machines, and several on-line learning strategies.

3. These learning approaches were coupled with a host of fine-grained linguistic characteristics. Put differently, high-dimensional feature spaces were constructed on top of assorted linguistically-motivated attributes extracted from natural language processing such as sentiment analysis, named entity recognition and dependency parsing. That is to say, we sought for linguistic features that characterize the presence of each factor.

[1] http://personality-testing.info/tests/IPIP-BFFM/

In a nutshell, our experiments unveil that a model effective in identifying the degree of one factor is unlikely to be effective in dealing with another factor, since a particular learning strategy and a specialized set of features are required. Since some of our findings are consistent with related studies on Facebook and Twitter, our results underscore that this decomposition is a feasible way of lessening the necessity of volunteers for answering the test. The reminder of this paper is organized as follows. Section 2 deals with the related work, next section 3 fleshes out our approach, section 4 breaks down our experiments and findings. Eventually, section 5 draws some conclusions and provides some future works.

2 Related Work

In the last decade, there has been an uptick in research into community question answering due to the wide variety of difficulties faced by this class of system [50]. In effect, enhancing their user experience entails several tough challenges, e.g., identifying high quality content [2, 40, 45, 59], bridging the gap between new questions and past good fitting answers [20, 25, 60], capitalizing on user search activity for enhancing the search across cQA archives [5, 28, 55], and finding potential experts that could readily answer new posted questions [41, 42, 43, 57]. Take for instance, the approach of [12] dug deeper into models capable of deciding whether an incoming question will be solved or it should be rerouted to an operator.

Recently, the focus have shifted towards understanding the behaviour of the community members [1, 26, 40]. In this vein, the work of [18] exploited multi-label learning for discriminating the multiple motivations an asker might have when publishing a question, and the study of [36] targeted at automatically identifying knowledge sharers from non-sharers. Further, [30] analysed the behavior of abusive users, and [44] estimated the reliability and expertise of Yahoo! Answers members.

Furthermore, [9] identified authorities, and [10] studied the behaviour of users associated with clarification questions. Lastly, [58] investigated the

impact and relation that exist between intrinsic (e.g. interest) and extrinsic motivations (e.g. reward) as well as expertise.

To the best of our knowledge, the Big Five personality factors have not been studied yet in the realm of Yahoo! Answers members. Needless to say, there have been recent studies concerning these factors, but in the sphere of Facebook [39, 46], and of several real world societies [27]. In particular, the work of [39] searched for patterns across Facebook statuses that typified each of these five personality factors. For this purpose, they asked some Facebook users to answer the Big Five test and to hand-in their status history. Along the same lines, the approach of [46] predicted each factor with an accuracy ranging between 65%-75% by taking advantage of their written messages.

In addition, they accounted for 75,000 volunteers that answered the test and provided their demographic information. They cast their prediction models as training binary classifiers on top of lexical, sentences and topic properties. Also, the work of [35] identified these five factors across Twitter users by analyzing their profile images. This study also had access to answered Big Five tests. In the same spirit, [51] distinguished the motivations and behavior that lead to posting selfies in Facebook. They additionally analyzed the roles of narcissism in predicting selfie-posting behavior. Furthermore, the work of [47] examined the motivations for posting pictures of oneself. More recently, [37] cast the detection of each Big Five factor across annotated essays as a binary (presence/absence) classification task.

In juxtaposition, this paper differs from earlier approaches in several aspects including: a) we make the first attempt to examine the Big Five factors in the restricted environment given by question-answering communities, namely in a space where people interact and express themselves in the form of questions and answers interchanges; b) we modelled each factor in consonance with the five-likert scale, and accordingly the automatic recognition of each factor is cast as a five-category classification task; c) as a means of discovering defining linguistic characteristics corresponding to each level/category vs. factor, we studied fifteen supervised multi-class models and checked the effectiveness of a plethora of fine-grained linguistic features (e.g., sentiment analysis and dependency parsing); and d) in order to acquired a study corpus, we benefited from a decomposition of the Big Five test in order to avoid the necessity of asking community fellows to answer the test.

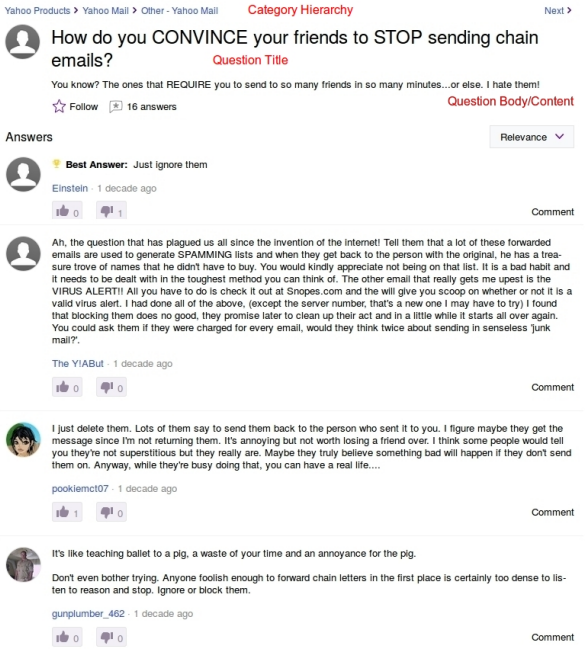

Fig. 1. A snapshot of an illustrative questions in Yahoo! Answers. The question title, body, category and its answers are indicated

3 Detecting Personality Traits Across Yahoo! Answers

For starters, we acquired a study corpus by crawling the Yahoo! Answers site from September 2015 to March 2016. The question taxonomy system supplied by Yahoo! Answers encompasses three-levels. Here, categories at the top level are broad, hence embracing a large amount of questions and constantly receiving newly posted questions and answers.

On the flip side, categories at the bottom level are very fine-grained, thus having a narrow coverage and seldom getting new questions. For

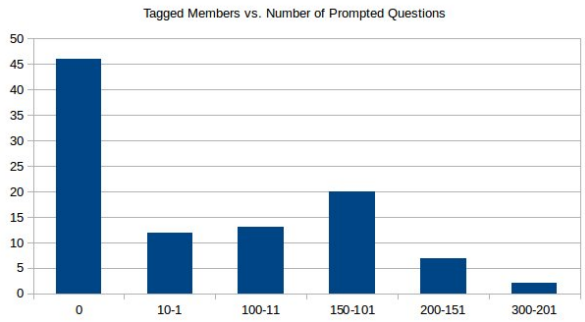

Fig. 2. The number of members (y-axis) and their corresponding amount of asked questions (x-axis) in the annotated material

this reason, this crawler navigated through the first two levels only, retrieving the top ten questions displayed when browsing the corresponding page. In order to grow the volume of fetched questions, this crawler visited several times each of these categories during this period of time. Accordingly, all question titles, bodies and answers were stored (see figure 1). Note that by crawling all first and second level categories, we aimed at accumulating questions targeting a wide variety of topics. In total, we acquired almost 370,000 questions and their respective answer sets.

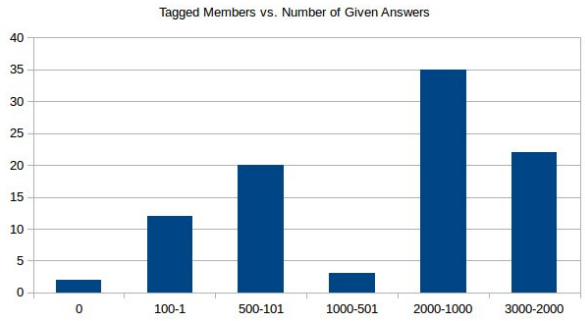

Fig. 3. The amount of members (y-axis) and their respective number of given answers (x-axis) in the labelled corpus

Since this crawler is not designed to filter the downloaded Yahoo! Answers pages by their language, we singled out the content conveyed in English by means of running a language detector[2] on every question and answer. Accordingly, the activity of each community fellow was assembled by searching for all his/her questions and answers across the fetched material (in English). Eventually, this corpus was reduced to the one hundred highest active members.

The assumption here is that the larger the amount of textual content a member provides, the higher the precision in determining the degree of each of the five factors, since there is a lower probability of missing pertinent pieces of information during the crawling phase. On average, a selected member was associated with 2,000 and 800 answers and questions (see figures 2 and 3), respectively.

Given the fact that we do not have access to ask these one hundred community members to volunteer for taking the Big Five test, we conducted a discourse analysis to quantify each factor. This analysis was carried out on the grounds of decomposing this test into a limited group of aspects, namely language descriptors (adjectives). The first list of descriptors was proposed by [3], and comprised more than 4.500 elements. In the course of time, the size of this list has been systematically shortened to 112 items [29].

All in all, using this small, but more precise, set of descriptors facilitates the discourse analysis. It is worth underscoring here that the array of adjectives corresponding to each factor is additionally divided into two groups: strongly and weakly related descriptors (see table **??**). The former signals a positive description of the respective factor, while the latter outlines what the factor is not. Also note that each question contained in the Big Five test is aimed at determining the degree of manifestation of each of these 112 adjectives in the subject of study.

In our case, we manually inspected the implicit and explicit presence of each of these 112 descriptors. As a means of quantifying the degree of each of these factors for each individual (I_f), we utilized the following equation:

[2] https://code.google.com/archive/p/language-detection/

ISSN 2007-9737

Table 1. Descriptors related to each of the five personality factors [29]

Extraversion		Agreeableness		Conscientiousness		Neuroticism		Openness	
Low	High	Low	High	Low	High	Low	High	Low	High
Quite	Talkative	Fault-finding	Symphatethic	Careless	Organized	Stable	Tense	Commonplace	Wide interests
Reserved	Assertive	Cold	Kind	Disorderly	Thorough	Calm	Anxious	Narrow interests	Imaginative
Shy	Active	Unfriendly	Appreciative	Frivolous	Planful	Contented	Nervous	Simple	Intelligent
Silent	Energetic	Quarrelsome	Affectionate	Irresponsible	Efficient	Unemotional	Moody	Shallow	Original
Withdrawn	Outgoing	Hard-hearted	Soft-hearted	Slipshot	Responsible		Worrying	Unintelligent	Insightful
Retiring	Outspoken	Unkind	Warm	Undependable	Reliable		Touchy		Curious
	Dominant	Cruel	Generous	Forgetful	Dependable		Fearful		Sophisticated
	Forceful	Stern	Trusting		Conscientious		High-strung		Artistic
	Enthusiastic	Thankless	Helpful		Precise		Self-pitying		Clever
	Show-off	Stingy	Forgiving		Practical		Temperamental		Inventive
	Sociable		Pleasant		Deliberate		Unstable		Sharp-witted
	Spunky		Good-natured		Painstaking		Self-punishing		Ingenious
	Adventurous		Friendly		Cautious		Despondent		Witty
	Noisy		Cooperative				Emotional		Resourceful
	Bossy		Gentle						Wise
			Unselfish						Logical
			Praising						Civilized
			Sensitive						Foresighted
									Polished
									Dignified

$$I_f = 5*(n_{srdf} - n_{srdnf}) + (n_{wrdf} - n_{wrdnf}) + b. \quad (1)$$

In this formula, n_{srdf} and n_{srdnf} stand for the number of highly related descriptors found and not found across the set of questions and answers posted by the member, respectively. On the other hand, n_{wrdf} and n_{wrdnf} denote the amount of weakly related descriptors found and not found in his/her activity, respectively. Lastly, b is a base factor computed as five times the number of all (found and not found) weak descriptors.

Eventually, the range method is utilized for transforming the I_f value into the interval between zero and one ($\bar{I}_f$). Thus, each of the five factors was labelled on an individual in conformity to the five-point likert scale as follows: a low degree (0) if $\bar{I}_f$ fall into the interval 0-0.2, while medium-low level in the event of 0.21-0.4 (1), medium 0.41-0.6 (2), medium-high 0.61-0.8 (3), and high 0.81-0.1 (4).

Now, we can cast the detection of each factor as a five-category classification task. For this reason, we tested the performance of several multi-class supervised learning techniques such as MaxEnt, Bayes and Passive Aggressive learning (see the full list on table **??**). To be more exact, we studied the effectiveness of the following learning approaches:

— **Support Vector Machines (SVMs):** Non-probabilistic linear classifiers aimed at separating categories by a gap that is as large as possible. We benefited from the multi-core implementation supplied by Liblinear[3] [11, 32]. More exactly, we capitalized on the Dual L2-regularized L1-loss SVMs and L2-loss SVMs.

— **Bayes:** Probabilistic classifiers based on the theorem of Bayes with a strong independence assumption between the features. We profited from the multinomial and bernoulli implementations supplied by OpenPR[4] [33]. Both combined with Laplace Smoothing.

— **Maximum Entropy Models (MaxEnt)[5]:** Probabilistic classifiers that belongs to the family of exponential models. MaxEnt does not assume that the features are conditionally independent [4].

— **Online learning:** Learning algorithms concerned with making decision with limited information [7]. We tested several approaches

[3] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/multicore-liblinear/

[4] http://www.openpr.org.cn/index.php/All/66-Naive-Bayes-EM-Algorithm/View-details.html

[5] http://http://www.nactem.ac.uk/tsuruoka/maxent/

provided Online Learning Library[6]: Log-Linear Models(SGD) [52], Winnow2 [34], AROW [15], several confidence weighted strategies [14, 17, 53, 54], dual averaging [56], and three passive aggressive methods [13].

Each of these fifteen models was tested with several combinations of fine-grained linguistically-oriented attributes computed by CoreNLP[7] [38]. All these properties were harvested from the questions and answers submitted by the user. To be more exact, the following sources of natural language processing were considered:

— **Part-of-Speech (POS):** We used frequency counts for each of the Penn Treebank POS tags. We also perceived as feature the size of their longest streak in the text.

— **Named-Entities (NER):** Like POS, we count the number of tokens labeled with each class of entity such as date, location, person and organization. Here, we benefited from the seven categories model supplied by CoreNLP. We additionally accounted for the number of entity and non-entity tokens. We also capitalized as attributes on the size of the longest sequence of each type.

— **Dependency Parsing (DP):** Similarly to [19, 21, 22, 23, 48, 49], we model frequencies for each type of relationship between pairs of terms. In addition, we computed the minimum, average and maximum (see sample in table **??**): a) depth of a tree; b) ramification per node; and c) number of nodes at each level of the tree.

— **Sentiment Analysis (SA):** CoreNLP indentifies the polarity of words and sentences according to a five-level scale (i.e., very positive, positive, neutral, negative and very negative). From this view, we extract several features: the amount of tokens of each polarity level, and the number of sentences tagged with each polarity level. The most common, minimum, maximum, and average sentiment level associated with terms and sentences.

[6] https://github.com/oiwah/classifier

[7] http://nlp.stanford.edu/software/corenlp.shtml

Table 2. Some attributes distilled from the lexicalised dependency view of the sentence "*click on start, click on accessories, click on system tools and finally click on disc clean up. Clean up c not d*"

Feature	Value
minimum ramification	1
maximum ramification	8
average ramification	1.5
minimum depth	3
maximum depth	5
average depth	4
frequency advmod	2
frequency conj	3
frequency prep	4
frequency punct	2
minimum breadth (level 1)	2
maximum breadth (level 1)	8
minimum breadth (level 4)	1

— **Bag-of-words (BoW):** We computed several versions comprising raw and lemmatized terms. We also considered alternatives with and without stop-words, and the combinations thereof. Furthermore, we computed a bag-of-words for each POS, NER and for every sort of sentiment polarity.

— **Others:** We exploit the number of tokens contained in the longest and shortest sentence as well as the average amount of terms across sentences.

4 Experiments

In all our experiments, we conducted a Leave-One-Out Cross-Validation (LOOCV), which is an ad-hoc methodology for small annotated corpora. With regard to an evaluation metric, we capitalized on the Mean Reciprocal Rank (MRR). The final MRR is the average of the reciprocal ranks of the predictions obtained for a sample of users U. In this case, for a given factor, the position of the correct degree in terms of the five-likert scale:

$$\frac{1}{|U|}\sum_{i=1}^{|U|}\frac{1}{rank_i}.$$

For each combination of a factor and a learner, the best set of features was determined by running

an SFS algorithm (a.k.a. Sequential Forward Selection) [8, 16, 31]. This process starts with an empty bag of properties and after each iteration adds the one that performs the best. In order to determine this feature, this procedure tests each non-selected attribute together with all the properties in the bag. The algorithm stops when no non-selected feature enhances the performance. Note that the outcome of this SFS is a subset of the 129 feature groups, thus of the 527,855 vector components, which this algorithm believes to perform the best.

In light of our empirical outcomes, we can draw the following conclusions (see table 3):

1. A bird's eye view shows that different classifiers reaped the best performance for different factors, meaning that there is a necessity of exploring a wide variety of learning approaches when devising models for identifying distinct personalities (see table **??**). More precisely, our experimental results point out to the fact that each factor is different from the others to the extent that their degrees are more efficiently recognized by distinct discriminant functions. Overall, on average terms, MaxEnt[8] and AROW [9] dominated their counterparts, hence holding a promise [4, 15, 52]. A closer look to our outcomes reveals that:
   (a) In the case of extraversion, the best configuration reaped an accuracy of 73%. Most of errors were due to medium-low (1) individuals seen as medium (2).
   (b) For agreeableness, the top system finished with 75% accuracy. The main source of misclassifications was also medium-low (1) members tagged as medium (2).
   (c) With regard to conscientiousness, the best model achieved an accuracy of 65%. Oppositely to the previous factors, a significant portion of the errors were due to medium (2) fellows labeled as medium-low (1).
   (d) Concerning neuroticism, the top classifier reached 69% accuracy. A substantial portion of the misclassifications were due to medium (2) users conceived as medium-low (1).
   (e) As for openness, the best system accomplished an accuracy of 63%. Similarly to the first two factors, a considerable fraction of errors comes from medium-low (1) individuals perceived as medium (2).

   All in all, our error analysis indicates that the major source of classifications comes from individuals starting to manifest the specific factor.

2. By and large, best classifiers incorporate features harvested from sentiment analysis and dependency parsing into their models (see table **??**). In particular, counts of positive/negative sentences and bags of positive/negative words. As for dependency trees[10], syntactic structures in their first levels were discriminative, i.e., average and maximum breadth. Overall, there is a sharp contrast in the kinds of features effective in dealing with neuroticism/openness and the other three factors.

3. More precisely, neutral words were found to be informative when dealing with extraversion. Words bearing the "anti" prefix were conspicuous in low levels of this factor (e.g., "*anti-semite*" and "*anti-christ*"). Conversely, politic-related nouns, identified as been emotionally charged, exhibited a stronger connection to higher levels (e.g., "*democrat*", "*hypocrisy*" and "*nationalism*"). Additionally, we observed that the average minimum depth of the dependency trees decreases as long as higher degrees of extraversion start to show up, this means that higher levels are more likely to be connected to a larger amount of easy to read sentences than lower levels.

[8] http://www.nactem.ac.uk/tsuruoka/maxent/

[9] https://github.com/oiwah/classifier

[10] http://nlp-ml.io/jg/software/pac/standep.html

Table 3. Best performance achieved by each combination of a classifier and a personality trait (MRR)

Classifier	Extraversion	Agreeableness	Conscient.	Neuroticism	Openness	Avg.
MaxEnt	0.8367	**0.8550**	0.7917	0.8217	0.7917	0.8193
Bayes (Bernoulli)	0.8283	0.8400	0.7867	0.8292	0.7750	0.8118
Bayes (Multinomial)	0.8217	0.8283	**0.8200**	0.7942	0.7650	0.8058
Dual Averaging	0.8225	0.8383	0.8067	0.7833	0.7550	0.8012
Passive Aggressive (PA)	0.8275	0.8250	0.8033	0.8292	0.7567	0.8083
Passive Aggressive I (PA-I)	0.8275	0.8250	0.8033	0.8292	0.7567	0.8083
Passive Aggressive II (PA-II)	0.8275	0.8250	0.8033	0.8292	0.7450	0.8060
Winnow2	0.8317	0.8333	0.7783	0.8242	**0.8000**	0.8135
Adaptive Regularization of Weight Vectors(AROW)	**0.8467**	0.8283	0.8117	0.8117	0.7733	0.8143
Confidence-Weighted (CW)	0.8308	0.8250	0.7950	0.8242	0.7767	0.8103
Soft Confidence-Weighted (SCW-1)	0.8300	0.8267	0.7950	**0.8333**	0.7700	0.8110
Squared Soft Confidence-Weighted (SCW-2)	0.8300	0.8267	0.7950	**0.8333**	0.7700	0.8110
Online SGD-L1 LogLinear	0.7842	0.8000	0.7600	0.7400	0.6933	0.7555
L2-regularized L1-loss SVM (dual)	0.8317	0.8333	0.7783	0.8242	0.7900	0.8115
L2-regularized L2-loss SVM (dual)	0.7450	0.7800	0.7467	0.7175	0.6733	0.7325
Max.	0.8467	0.8550	0.8200	0,8333	0.8000	
Avg.	0.8214	0.8260	0.7917	0.8083	0.7594	

4. Curious though it may seem, the count of negative sentences and the bag of negative words were instrumental in recognizing the degree of agreeableness. Words bearing the "anti" prefix were conspicuous in low levels of this factor (e.g., "*anti-semite*" and "*anti-christ*"). On the other hand, politic-related nouns, identified as been emotionally charged, exhibited a stronger connection to higher levels (e.g., "*democrat*", "*hypocrisy*" and "*nationalism*"). On average, people showing higher levels of agreeableness make reference to percentages ca. 20% more than other individuals. This group of people is likely to touch on topics related to taxes and insurances.

5. In the case of conscientiousness, selected attributes were dominated by dependency analysis. For instance, the average number of adverbial clauses (advcl) signal temporal, consequence, conditional clauses, etc. Furthermore, the best model profits from relations that link verbs and their dative objects, and it capitalizes on dependencies indicating number phrases that serve to modify the meaning of nouns with quantities. The common denominator of all these dependency relations is that they are used for providing specifics/details when asking/answering. Lastly, higher numbers of negative sentences were connected to lower degrees of this factor.

6. By inspecting the terminology, medium and low levels of neuroticism were found to be related to nouns linked to prescriptions for antidepressants and anti-anxiety agents as well as anti-psychotics. Given the fact that all these terms are related to depression and anxiety disorders, we conclude that they are people, at initial stages of these disorders, looking for help or information about their illness.

7. In juxtaposition, POS taggings were fundamental to the best model for openness. Specifically, the counts of existential there (EX), wh-pronouns (WP), and particles (RP). Overall, our outcomes point out to the fact that this is hardest factor to assess accurately across question-answering communities.

Interestingly enough, despite dealing with a linguistically different corpus, some of our empirical outcomes are in agreement with some of the findings discovered by researchers working on Facebook and Twitter material:

1. Analogously to Facebook [39], community fellows bearing a high level of conscientiousness are highly likely to share information

Table 4. Informative linguistic-characteristics vs. factors (best models only)

Feat. Group	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
SA	-Positive BoW -Neutral BoW	-No. of negative sentences -Negative BoW	-No. of negative sentences		
DP	-Minimum depth	-1st level maximum breadth -Open clausal complements	-Link from a noun to a number premodifier -Links between a verb and a verb heading a modifier -Links between a verb and an expletive 'there' subject -Links between a verb and its dative object -Breadth at level seven		
POS		-Existential there		-Symbols	-Existential there -Wh-pronouns -Particle
BoW				-Lemmatized w/stop-swords	-Lemmatized
NER		-Percentages		-Dates	

in children-related categories such as *Family & Relationships* and *Pregnancy & Parenting*. However, in Yahoo! Answers, we discovered that they also have a high level of activity in topics including *Health* and *Buisness & Electronics*.

2. In the same spirit of Facebook [39], community peers having a higher degree of openness are more likely to associate with intellectual topics. More precisely, we found out that they were involved in topics such as *Cars & Transportation* and *Society & Culture*, *Arts & Humanities* as well. We additionally noted that they very keen to write about current events in categories such as *Politics & Government*.

3. In Twitter, people with a high degree of agreeableness are probable to show positive emotions on their profile pictures [35]. In Yahoo! Answers, our experiments showed that sentiment analysis was informative to measure the degree of agreeableness.

In summary, our experiments unveil that a model effective in identifying the degree of one factor is unlikely to be effective in dealing with another factor, since a particular learning strategy and a specialized set of features are required. Our results also reveal that effective models targeted at neuroticism and openness require a deep word-level analysis (i.e., morphology), while agreeableness, conscientiousness and extraversion sentiment analysis. Additionally, our outcomes show that dependency parsing is instrumental to conscientiousness. Lastly, we found that our results were partially supported by previous research working on other kinds of data.

5 Conclusion and Future Work

This paper shows that it is possible to train effective multi-class discriminant models to identify different personality traits across question-answering communities. To be more exact, it shows that it is plausible to mitigate the need for conducting a psychological test to each user by using a deconstruction of the Big Five test into 112 descriptors, which can be manually inspected in the activity of each member.

Our outcomes highlights that effective models in tackling one factor are likely to be sharply different from effective models in coping with another factor. In terms of features, sentiment analysis and dependency parsing proven to be fundamental to deal with extraversion, agreeableness and conscientiousness. Furthermore, neuroticism was shown to be connected with the initial stages of depression and anxiety disorders.

As for future work, we envisage the use of semi-supervised learning to increase the power of generalization of our models, since annotated data is hard to obtain. In addition, multi-task learning holds a promise, since simultaneously learning the five factors might help to enhance each individual classification rate.

Acknowledgment

This work was partially supported by the project FONDECYT "*Bridging the Gap between Askers and Answers in Community Question Answering Services*" (11130094) funded by the Chilean Government.

References

1. **Adaji, I. & Vassileva, J. (2016).** Towards understanding user participation in stack overflow using profile data. *International Conference on Social Informatics*, Springer, pp. 3–13.
2. **Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G. (2008).** Finding high-quality content in social media. *Proceedings of the 2008 international conference on web search and data mining*, ACM, pp. 183–194.
3. **Allport, G. W. & Odbert, H. S. (1936).** Trait-names: A psycho-lexical study. *Psychological monographs*, Vol. 47, No. 1, pp. i.
4. **Andrew, G. & Gao, J. (2007).** Scalable training of l 1-regularized log-linear models. *Proceedings of the 24th international conference on Machine learning*, ACM, pp. 33–40.
5. **Arora, P., Ganguly, D., & Jones, G. J. (2015).** The good, the bad and their kins: Identifying questions with negative scores in stackoverflow. *ASONAM 2015*, ACM, pp. 1232–1239.
6. **Arora, P., Ganguly, D., & Jones, G. J. (2016).** Nearest neighbour based transformation functions for text classification: A case study with stackoverflow. *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ICTIR '16, ACM, New York, NY, USA, pp. 299–302.
7. **Blum, A. (1996).** On-line algorithms in machine learning. *In Proceedings of the Workshop on On-Line Algorithms, Dagstuhl*, Springer, pp. 306–325.
8. **Blum, A. L. & Langley, P. (1997).** Selection of relevant features and examples in machine learning. *Artificial Intelligence*, Vol. 97, No. 1?2, pp. 245–271. Relevance.
9. **Bouguessa, M. & Romdhane, L. B. (2015).** Identifying authorities in online communities. *ACM Trans. Intell. Syst. Technol.*, Vol. 6, No. 3, pp. 30:1–30:23.
10. **Braslavski, P., Savenkov, D., Agichtein, E., & Dubatovka, A. (2017).** What do you mean exactly?: Analyzing clarification questions in cqa. *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, ACM, New York, NY, USA, pp. 345–348.
11. **Chiang, W.-L., Lee, M.-C., & Lin, C.-J. (2016).** Parallel dual coordinate descent method for large-scale linear classification in multi-core environments. *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, ACM, New York, NY, USA, pp. 1485–1494.
12. **Convertino, G., Zancanaro, M., Piccardi, T., & Ortega, F. (2017).** Toward a mixed-initiative {QA} system: from studying predictors in stack exchange to building a mixed-initiative tool. *International Journal of Human-Computer Studies*, Vol. 99, pp. 1–20.
13. **Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., & Singer, Y. (2006).** Online passive-aggressive algorithms. *Journal of Machine Learning Research*, Vol. 7, No. Mar, pp. 551–585.
14. **Crammer, K., Dredze, M., & Pereira, F. (2012).** Confidence-weighted linear classification for text categorization. *J. Mach. Learn. Res.*, Vol. 13, No. 1, pp. 1891–1926.
15. **Crammer, K., Kulesza, A., & Dredze, M. (2009).** Adaptive regularization of weight vectors. *Advances in neural information processing systems*, pp. 414–422.
16. **Devijver, P. A. & Kittler, J. (1982).** *Pattern Recognition: A Statistical Approach*. Prentice-Hall, London, U.K.
17. **Dredze, M., Crammer, K., & Pereira, F. (2008).** Confidence-weighted linear classification. *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, ACM, New York, NY, USA, pp. 264–271.
18. **Espina, A. & Figueroa, A. (2017).** Why was this asked? automatically recognizing multiple motivations

behind community question-answering questions. *Expert Systems with Applications*, Vol. 80, No. 1, pp. 126–135.

19. **Figueroa, A. (2010).** Surface language models for discovering temporally anchored definitions on the web - producing chronologies as answers to definition questions. *Proceedings of the 6th International Conference on Web Information Systems and Technology*, pp. 269–275.
20. **Figueroa, A. (2017).** Automatically generating effective search queries directly from community question-answering questions for finding related questions. *Expert Systems with Applications*, Vol. 77, pp. 11–19.
21. **Figueroa, A. & Atkinson, J. (2009).** Answering definition questions: Dealing with data sparseness in lexicalised dependency trees-based language models. *WEBIST (Selected Papers)*, Springer, pp. 297–310.
22. **Figueroa, A. & Atkinson, J. (2009).** Using dependency paths for answering definition questions on the web. *Proceedings of the Fifth International Conference on Web Information Systems and Technologies - Volume 1: WEBIST,*, INSTICC, SciTePress, pp. 638–645.
23. **Figueroa, A. & Atkinson, J. (2012).** Contextual language models for ranking answers to natural language definition questions. *Computational Intelligence*, Vol. 28, No. 4, pp. 528–548.
24. **Figueroa, A., Gómez-Pantoja, C., & Herrera, I. (2016).** Search clicks analysis for discovering temporally anchored questions in community question answering. *Expert Systems with Applications*, Vol. 50, pp. 89–99.
25. **Figueroa, A. & Neumann, G. (2016).** Context-aware semantic classification of search queries for browsing community questionâ€“answering archives. *Knowledge-Based Systems*, Vol. 96, pp. 1–13.
26. **Gazan, R. (2015).** First-mover advantage in a social q&a community. *System Sciences (HICSS), 2015 48th Hawaii International Conference on*, IEEE, pp. 1616–1623.
27. **Gurven, M., Von Rueden, C., Massenkoff, M., Kaplan, H., & Lero Vie, M. (2013).** How universal is the big five? testing the five-factor model of personality variation among forager–farmers in the bolivian amazon. *Journal of personality and social psychology*, Vol. 104, No. 2, pp. 354.
28. **John, B. M., Goh, D. H. L., Chua, A. Y. K., & Wickramasinghe, N. (2016).** Graph-based cluster analysis to identify similar questions: A design science approach. *Journal of the Association for Information Systems*, Vol. 17, No. 9, pp. 590.
29. **John, O. P. & Srivastava, S. (1999).** The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, Vol. 2, No. 1999, pp. 102–138.
30. **Kayes, I., Kourtellis, N., Quercia, D., Iamnitchi, A., & Bonchi, F. (2015).** The social world of content abusers in community question answering. *Proceedings of WWW '15*, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pp. 570–580.
31. **Kohavi, R. & John, G. H. (1997).** Wrappers for feature subset selection. *Artificial Intelligence*, Vol. 97, No. 1/2, pp. 273–324. Relevance.
32. **Lee, M. C., Chiang, W. L., & Lin, C. J. (2015).** Fast matrix-vector multiplications for large-scale logistic regression on shared-memory systems. *2015 IEEE International Conference on Data Mining*, pp. 835–840.
33. **Lewis, D. D. (1998).** *Naive (Bayes) at forty: The independence assumption in information retrieval*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 4–15.
34. **Littlestone, N. (1988).** Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, Vol. 2, No. 4, pp. 285–318.
35. **Liu, L., Preotiuc-Pietro, D., Samani, Z. R., Moghaddam, M. E., & Ungar, L. H. (2016).** Analyzing personality through social media profile picture choice. *ICWSM*, pp. 211–220.
36. **Liu, Z. & Jansen, B. J. (2017).** Identifying and predicting the desire to help in social question and answering. *Information Processing & Management*, Vol. 53, No. 2, pp. 490–504.
37. **Majumder, N., Poria, S., Gelbukh, A., & Cambria, E. (2017).** Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, Vol. 32, No. 2, pp. 74–79.
38. **Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014).** The Stanford CoreNLP natural language processing toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60.

39. **Marshall, T. C., Lefringhausen, K., & Ferenczi, N. (2015).** The big five, self-esteem, and narcissism as predictors of the topics people write about in facebook status updates. *Personality and Individual Differences*, Vol. 85, pp. 35–40.

40. **Neshati, M. (2017).** On early detection of high voted q&a on stack overflow. *Information Processing & Management*, Vol. 53, No. 4, pp. 780–798.

41. **Pal, A., Farzan, R., Konstan, J. A., & Kraut, R. E. (2011).** Early detection of potential experts in question answering communities. *UMAP*, pp. 231–242.

42. **Pal, A., Harper, F. M., & Konstan, J. A. (2012).** Exploring question selection bias to identify experts and potential experts in community question answering. *ACM Trans. Inf. Syst.*, Vol. 30, No. 2, pp. 10.

43. **Pal, A. & Konstan, J. A. (2010).** Expert identification in community question answering: exploring question selection bias. *CIKM*, pp. 1505–1508.

44. **Pelechrinis, K., Zadorozhny, V., Kounev, V., Oleshchuk, V., Anwar, M., & Lin, Y. (2015).** Automatic evaluation of information provider reliability and expertise. *World Wide Web*, Vol. 18, No. 1, pp. 33–72.

45. **Ravi, S., Pang, B., Rastogi, V., & Kumar, R. (2014).** Great question! question quality in community q&a. *Proceedings of the International AAAI Conference on Web and Social Media*.

46. **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. (2013).** Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, Vol. 8, No. 9, pp. e73791.

47. **Seidman, G. (2013).** Self-presentation and belonging on facebook: How personality influences social media use and motivations. *Personality and Individual Differences*, Vol. 54, No. 3, pp. 402–407.

48. **Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., & Chanona-Hernández, L. (2013).** Syntactic dependency-based n-grams as classification features. **Batyrshin, I. & Mendoza, M. G.**, editors, *Advances in Computational Intelligence*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1–11.

49. **Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., & Chanona-Hernández, L. (2014).** Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications*, Vol. 41, No. 3, pp. 853–860. Methods and Applications of Artificial and Computational Intelligence.

50. **Srba, I. & Bielikova, M. (2016).** A comprehensive survey and classification of approaches for community question answering. *ACM Trans. Web*, Vol. 10, No. 3, pp. 18:1–18:63.

51. **Sung, Y., Lee, J.-A., Kim, E., & Choi, S. M. (2016).** Why we post selfies: Understanding motivations for posting pictures of oneself. *Personality and Individual Differences*, Vol. 97, pp. 260–265.

52. **Tsuruoka, Y., Tsujii, J., & Ananiadou, S. (2009).** Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, Association for Computational Linguistics, pp. 477–485.

53. **Wang, J., Zhao, P., & Hoi, S. C. H. (2012).** Exact soft confidence-weighted learning. *CoRR*, Vol. abs/1206.4612.

54. **Wang, J., Zhao, P., & Hoi, S. C. H. (2016).** Soft confidence-weighted learning. *ACM Trans. Intell. Syst. Technol.*, Vol. 8, No. 1, pp. 15:1–15:32.

55. **Wu, H., Wu, W., Zhou, M., Chen, E., Duan, L., & Shum, H.-Y. (2014).** Improving search relevance for short queries in community question answering. *WSDM '14*, ACM, New York, NY, USA, pp. 43–52.

56. **Xiao, L. (2010).** Dual averaging methods for regularized stochastic learning and online optimization. *J. Mach. Learn. Res.*, Vol. 11, pp. 2543–2596.

57. **Yan, Z. & Zhou, J. (2015).** Optimal answerer ranking for new questions in community question answering. *Information Processing & Management*, Vol. 51, No. 1, pp. 163–178.

58. **Yang, J., Bozzon, A., & Houben, G.-J. (2015).** Harnessing engagement for knowledge creation acceleration in collaborative q&a systems. **Ricci, F., Bontcheva, K., Conlan, O., & Lawless, S.**, editors, *User Modeling, Adaptation and Personalization*, Springer International Publishing, Cham, pp. 315–327.

59. **Yao, Y., Tong, H., Xie, T., Akoglu, L., Xu, F., & Lu, J. (2015).** Detecting high-quality posts in community question answering sites. *Information Sciences*, Vol. 302, pp. 70–82.

60. Zhou, G., Zhou, Y., He, T., & Wu, W. (2016). Learning semantic representation with neural networks for community question answering retrieval. *Knowledge-Based Systems*, Vol. 93, pp. 75–83.

Article received on 14/07/2017; accepted on 19/01/2018.
Corresponding author is Nicolás Olivares.