Novel Approaches to Automated Personality Classification: Ideas and Their Potentials

Aleksandar Kartelj*, Vladimir Filipović* and Veljko Milutinović**
* School of Mathematics, University of Belgrade, Serbia
** School of Electrical Engineering, University of Belgrade, Serbia
aleksandar.kartelj@gmail.com

Abstract - In this paper, we propose several new research directions regarding the problem of Automated Personality Classification (APC). Firstly, we investigate possible improvements of the existing solutions to the problem of APC, for which we use different combinations of the APC corpora, psychological trait measurements, and learning algorithms. Afterwards, we consider extensions of the APC problem and the related tasks, such as dynamical APC and detecting personality inconsistency in a text. This entire research was performed in the context of social networks and the related datamining mechanisms.

I. INTRODUCTION

Personality classification is one of the problems considered by personality psychology, a branch of psychology. The focus of this field is the study of personality and individual differences. According to that study, personality can be defined as a dynamic and organized set of characteristics of a person, which have a unique influence on cognition, motivation and behavior of that person. In this paper the problem of automated personality classification is considered based on information from the following content: textual content that the person wrote and meta information about a person received on request, through social networks or other means. There are studies that also include speech, analysis of facial characteristics, gestures and other aspects of behavior, but they are not the subjects of our study. The standard approach to solving the APC problem based on the aforementioned content is described in the following steps: A. Gathering the corpus data, B. Determination of the personality characteristics of the participants, and C. Building the model.

A. Gathering the corpus data

The corpus includes a collection of content (both text and meta) from the participants on whom the building of the model is based. In previous research, this base usually consisted of student essays and accompanying meta information ([1], [16], [17] and [31]), originally collected and described in [24] and [23]. Other studies have considered Internet blogs ([9], [10], [11], [13], [18], [19], [20], [21], [34] and [36]). There were some efforts described in [22] and [27] that used emails and SMS, while the latest researches [3], [25] and [26] used information available on social networks, Twitter and Facebook.

B. Determination of the personality characteristics

Determination of personality characteristics is traditionally performed by subjecting participants to a personality test. The most common implementation of a personality test is The Big Five Model of traits that classifies the personalities to five separate characteristics, where each is evaluated on the real scale. Considered characteristics are: 1) openness to experience, 2) conscientiousness, 3) extraversion, 4) agreeableness and 5) neuroticism. Certain questionnaire based implementations of this model are described in [2], [7] and [14].

C. Building a model

After collecting the corpus and determining the personality characteristics for each of the participants, it is necessary to develop an appropriate classification model. This procedure involves the selection of independent variables, i.e. a set of relevant attributes that determine the dependent variable, where the dependent variable represents one or more personality traits. In [4] and [23] two sets of features were identified, LIWC and MRC, which turned out to be correlated with certain personality characteristics. The LIWC (Linguistic Inquiry and Word Count) represents a database of several hundred words, which have been found to contribute to the determination of personality characteristics. For example, words *hate* and *kill* are quite important in determining the level of neuroticism. The MRC is a psycholinguistic database consisted of words classified by various measures, such as, imagery, concreteness, frequency of usage, etc. For example, the word *ship* is highly rated on the scale of concreteness, whereas the word patience is very poorly rated. After a set of relevant attributes (features) is established, an appropriate classification model is built on top of it. These classification models are usually based on different linguistic, stylistic and statistical techniques.

In the next section we describe general ideas for the improvement of the solutions to the problem of APC. In the third section of the paper we give brief overview of a several solutions to the problem of APC, and ideas for

their improvement. The fourth section will hold the conclusions of our research.

II. General Ideas For Improvement

In this section we present several ideas for the improvement of APC. For clarity, the ideas are categorized as follows: A. New types of corpora, B. New ways of measuring the personality, C. New models and algorithms, and D. Extensions and related problems.

A. New types of corpora

A considerable number of researches on the topic of APC considered Internet blogs and student essays. In [28] user reviews from *tripadvisor.com* system for proposing tourist destinations are taken into account, but only to the extent of evaluation of the model formed based on the student essays corpus from [24]. Our view is that the user reviews and comments should be used, not only for the testing of the quality of the model, but also for its creation. News portals and comments made by the general population could be utilized for solving the problem of determination of certain personality traits, such as extroversion, neuroticism, and openness to experience. Comments from the Internet site *youtube*.com could be used in a similar way. The analysis of public data flows from social networks is yet another resource that is currently not used for solving the APC problem. In the following paragraph we consider the alternative, primarily indirect ways of measuring the personality traits, based on these new corpora.

B. New ways of measuring the personality

Measurement of personality traits of participants is a difficult task because it requires cooperation of persons who are the authors of the content. Our idea is to simplify this aspect of the research, which is usually based on the use of complex questionnaires, by simplifying the problem of the APC. In the context of news portals and Youtube, this would require labeling (tagging) of articles and video content with certain pervasive personality trait or combination of traits, for example, scientific content would clearly indicate towards openness to experience, extreme sports content towards a combination of extraversion and neuroticism, etc. In this way we would simultaneously get both the corpus of content and a set of information on certain personality traits of the authors of those contents. In the context of measurement of personality traits on social networks, such as *Facebook*, the idea would be to create an application, which would be appealing to the users. At the same time it should have a hidden function for the measurement of personality traits.

C. New models and algorithms

There are several potential algorithms that could be applied to solve the problem of the APC. Combining, or hybridization, of the existing methods is one of the ways of achieving better results. In [29] a regression model is considered, whose evaluation is based on ranking, instead of the standard approach, which is based on residuals. Reducing the problem of the APC to the problem of clustering is also one of the alternatives. That would require analysis of several personality traits simultaneously, as opposed to analysis of each trait separately. In our opinion, this approach would be more natural, as we believe that the traits are mutually conditioned, i.e. that they cannot be combined in a completely arbitrary way. In addition to the support vector machine (SVM), the use of other methods from the class of soft computing algorithms, such as neural networks and fuzzy logic methods could also provide valid results.

D. Extensions and related problems

The dynamical APC system refers to the system that is capable to adapt to the users and the input data, and to "learn" through use. It has been shown that psychological traits follow a normal distribution. However, the moments of distribution differ in terms of various age groups, demographic characteristics, level of education, and other information about the author of the content. The dynamical APC system would also learn the parameters of the distribution through use and this could be achieved by applying the Bayesian learning.

Algorithms that solve the problem of automated plagiarism detection ([6] and [15]) represent an interesting scientific and practical topic. A personality classification could be used to catch some inconsistency in texts. The idea is to pass different fragments of the same text through the algorithm that performs the APC and compare the resulting personality scores. If deviations are too high, then we could conclude that more than one person wrote the text. There are some issues related to this idea, for example, what if the observed text was legally written by more than one person (scientific papers usually have more than one author)?

III. Improvement Of The Existing Solutions

In this section we will give brief overview of several research papers that considered the problem of APC. After each overview we will propose one, the most prominent, idea for the improvement of the existing solution. This idea will be represented with a short description, diagram, and mathematical considerations. Note that in the previous section we proposed general ideas, which are applicable to almost every existing solution. Therefore, in this section, we will concentrate only on the specific aspects of ideas.

A. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text

In [17] three different statistical models are applied on the corpus composed of text and speech. The classification of traits was based on the five-factor model, where each of the five traits represented a binary variable (low or high). Apart from the classification algorithms, authors considered the regression and the ranking based models. The textual part of the corpus consisted of student essays previously collected and discussed in [24]. Beside the LIWC and the MRC features, authors utilized ratios between the usages of different types of sentences: commands, prompts, questions, and assertions. Testing was performed on the Weka statistical tool. The ranking model outperformed other two models in overall. The

results also suggested that the evaluation of openness to experience was the most accurate.

We think that utilization of a ranking based model combined with clustering algorithm is the right direction for improvement. In [17] the authors used the RankBoost ranking algorithm ([8] and [30]). This algorithm uses a set of training pairs $T_i = \{(x,y) \mid p_i(x) > p_i(y)\}$, where *x* and *y* are vectors of linguistic features associated with content of an author, and p_i corresponds to the score on the personality trait i $(i = 1..5)$ for that author. Each vector *x* is composed of *m* indicator functions $h_s(x)$ *for* $1 \le s \le m$, where each indicator takes value 0 or 1, depending on whether the feature at position *s* of vector *x* is lower or greater than some threshold value (in the case of binary classification, the threshold may be represented with the median). Ranking score is afterwards calculated as $F(x) = \sum_s \alpha_s h_s(x)$. The training process is then used to adjust values of vector α, such that the following loss function (1) is minimized:

$$Loss = \frac{1}{|T|} \sum_{(x,y) \in T} eval(F(x) \le F(y)), \qquad (1)$$

where the *eval* function returns 1 if the ranking scores of (*x, y*) pair are not ordered, 0 otherwise. In [32] authors propose a novel clustering framework called RankClus that directly generates clusters integrated with ranking. Based on the initial K clusters, ranking is applied separately, which serves as a good measure for each cluster. Their experiment results show that RankClus can generate more accurate clusters and in a more efficient way than the state-of-the-art link-based clustering methods. In Fig. 1, a schema of this new hybrid approach is represented. We believe that this approach could improve the existing solution to the problem of APC in qualitative manner, mainly because clustering may be a more appropriate class of algorithms for solving the problem of APC. On the other hand, the complexity of the implementation of such a hybrid algorithm would increase drastically.

B. Personality Based Latent Friendship Mining

In [34] authors proposed a method that solves the problem of finding compatible friends within community of web bloggers. Unlike other methods that considered this problem, where the main hypothesis was that compatible bloggers write about related topics, authors here assume that compatible bloggers have similar personality characteristics. This way, the problem of APC is considered as a subproblem of the main problem. The logistic regression model for binary classification was utilized in this research. The feature vector consisted of 20 elements: the number of comments, the number of images, readability, average size of sentence, open text, represented as a vector of the used words, and others.

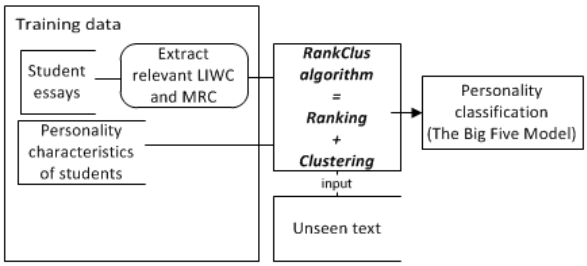

Figure 1. Ranking + clustering based algorithm

Let θ represent a similarity vector between a pair of bloggers w_i and w_j. The logistic function is defined as $Y = e^A / (1 + e^A)$, where the variable A is defined as $A = \lambda_0 + \lambda_1 X_1 + ... + \lambda_{20} X_{20}$. X_i is the i^{th} element of vector θ. Output *Y* ranges between 0 and 1. The threshold value of 0.5 is used to distinguish between two classes: non-potential friends (0) and potential friends (1). Articles of the bloggers, randomly selected from the MSN live space, were considered as an input in the regression. After that, based on their blogger friend lists, their friend's blogs were also downloaded. This way, blogger friends' relations were established. After that, the previously described regression model was trained and tested (80% of data for training and 20% for testing). Out of 20 features, those which proved to be significant are: plain text similarity, average number of comments, special symbols, font changes, words, hyperlinks, adjectives, nouns, average lexical density and readability. Authors concluded that the full-feature model gives solid precision and recall measures, 80% and 75%, respectively.

Our idea for improvement of this solution is based on the usage of a more sophisticated binary classifier, support vector machines ([5] and [33]). It would be interesting to check whether the SVM would give better results. Additionally, we suggest resizing of the 20^{th} feature, which represents the vector of all words occurring in text. Keeping in mind the results of previous researches, mainly [23] and [24], non-LIWC words should be removed from this feature. This modification (Fig. 2) would ensure stronger implementation of author's hypothesis, that compatible bloggers have similar personality characteristics.

C. Lexical Predictors of Personality Type

In [31] authors consider the problem of binary classification of neuroticism and extraversion. They used four different sets of features: function word list, conjunctive phrases, modality indicators, and appraisal adjectives and modifiers. Support vector machine is used as a binary classifier. This study is based on language psychology and computational stylistic. Computational stylistic is a discipline that threats the meaning of text through the following aspects: affect (what feeling is conveyed by the text?), genre (in what community of discourse does the text function?), register (what is the function of the text as a whole?), and personality (what sort of person, or who specifically wrote the text?). This discipline suggests that all these aspects can be extracted from the text's style of writing.

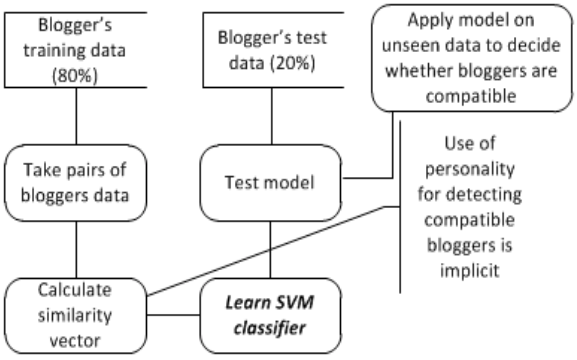

Figure 2. The SVM classifier instead of the logistic regression

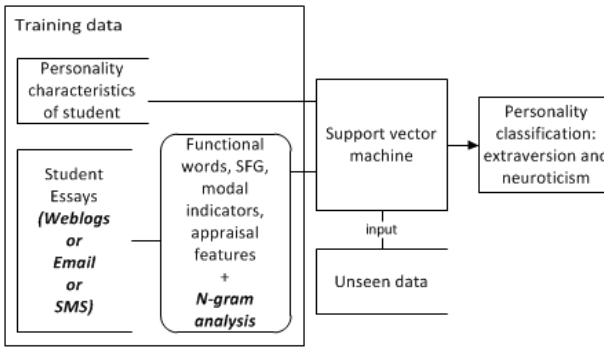

Figure 3. Utilization of the n-gram analysis and different types of corpora

Function words are frequent words that have a primarily grammatical function in the language (such as *and*, *for*, and *the*). The motivation behind using them as features is that they are not likely to be consciously controlled by the author of the text. Therefore, their frequencies can be useful in detecting author's style. Conjunctive phrases represent a concept in the theory of Systemic Functional Grammar (SFG), a functional approach to linguistic analysis ([12]). Modality indicators qualify events or entities in text to be classified according to their likelihood, typicality or necessity (modal verbs, adverbial adjunct, etc.). Finally, appraisal considers two attributes: attitude and orientation. Authors used corpus of essays written by students at the University of Texas at Austin between 1997 and 2003 ([23] and [24]). Appraisal features gave the highest accuracy when the neuroticism was in question. When extraversion was in question, only function words improved accuracy. Moreover, usage of other three types of features reduced the overall accuracy when they were added to function words.

Future research should consider different types of corpora. We are not convinced whether the approach that the authors proposed here would work for some less formal types of text, such as weblogs (web blogs), emails, as well as social networks feeds and conversations. Hybridization of this approach with some data-driven approaches, such as, the n-gram analysis should also be considered (Fig. 3).

D. Comparative Evaluation of Personality Estimation Algorithms for the TWIN Recommender System

In [28], a comparative evaluation of several algorithms for APC is presented. The goal of this research was finding the most adequate algorithm for use within the TWIN ("Tell me What I Need") system. The TWIN is a tourist recommender system used on the website *tripadvisor.com*. Recommender systems are usually based on evolutionary user profiling. The initial amount of profile information is collected during the user registration process. Afterwards, information is collected through surveillance of user activities. The key hypothesis of this research is that groups of people that share similar personality characteristics usually prefer to choose similar tourist destination and hotels.

Within TWIN systems, user reviews are gathered, and then a personality profile of user is constructed based on them. Finally, user profile is used to decide which destinations are suitable. Authors used previously formed corpus of student essays and their personality characteristics ([23] and [24]). Personality characteristics were measured using the five-factor model. The LIWC and the MRC were used as input features. Four models were utilized: linear regression, M5' classification tree, M5' regression tree, and support vector machine. Testing of quality was performed on reviews of 15 randomly chosen users of the TWIN system that had at least 30 reviews. For each personality trait and for each of the four models, a personality score was determined. Hypothesis is that a model is good if it gives similar scores to all reviews, written by the same user (standard deviation is as little as possible). M5' regression tree has been shown as the most prominent model, with respect to this criterion. Therefore, it was implemented inside the TWIN system.

In this work, authors used quality metric that is similar to one of those used in cluster analysis, the so-called intra-cluster distance. There are several ways to model this distance, but the essential common intuition behind all of them is that, if this distance is minimized, then elements inside clusters are more similar. Beside the intra-cluster distance, there is also inter-cluster distance, which measures the distance between centroids of different clusters. Note that this metric should be maximized. Our recommendation regarding this research is to use composite similarity metric that combines intra-cluster and inter-cluster distance (Fig. 4). David-Bouldin (2) and Dunn index (3) are examples of two such metrics:

$$DB = \frac{1}{n}\sum_{i=1}^{n}\max_{i\neq j}\left(\frac{\sigma_i+\sigma_j}{d(c_i,c_j)}\right), \tag{2}$$

$$D = \min_{1\le i\le n}\left\{\min_{1\le j\le n, i\neq j}\left\{\frac{d(c_i,c_j)}{\max_{1\le k\le n} d'(k)}\right\}\right\}. \tag{3}$$

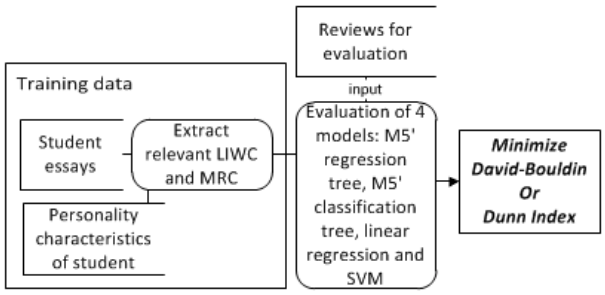

Figure 4. Composite cluster quaility metrics

n represents the number of clusters, c_x is the centroid of cluster x, σ_x is the standard deviation of elements inside cluster c_x and $d(c_i, c_j)$ is the distance between centroids.

E. *Our Twitter Profiles, Our Selves: Predicting Personality with Twitter*

In [25] authors analyze the relation between personality characteristics, based on the five-factor model and three publicly available Twitter stats. They used a corpus that consisted of 335 user's information. Proposed predictor of personality characteristics is based on the usage of: "following", "followers", and "listed" counts. Personality characteristics measurement was performed indirectly, through a Facebook application called *myPersonality*. Authors considered all users who specified their Twitter accounts on their Facebook profiles, and ended up having 335 Twitter users who took *myPersonality* questionnaire. The authors performed regression analysis based on the M5' Rules algorithm [35] for each of the five characteristics separately (Weka toolbox). Prediction error was 0.88 according to RMSE scale (root-mean-square error, takes values from interval [0, 5]). Authors claim that this error is low, and the main argument is the fact that the best collaborative filter algorithm for prediction of user movie ratings had RMSE of 0.8567 (company Netflix awarded the authors of this algorithm with the prize of $1M).

Regarding the possible improvement of personality prediction, we suggest two changes. First, the three aforementioned accounts should be normalized. We think that the usage of absolute figures could lead to obtainment of biased results. Given that all Twitter accounts used in this research are mapped with Facebook, $1/n_{facebook}$ could be utilized as a normalization factor, where $n_{facebook}$ represents the number of Facebook contacts. Another possibility for dealing with this problem would be to use ratios between Twitter counts. Second, implementation of some other regression-based algorithms should be considered. Keeping in mind that support vector machines had shown to be a high-quality solution for various classification and regression tasks, we would suggest the usage of support vector machine regression model (Fig. 5).

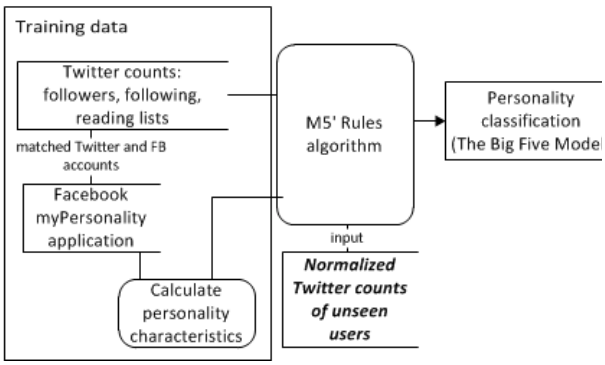

Figure 5. Normalization of Twitter stats

IV. CONLUSION

In this paper, new ideas regarding the problem of automated personality classification are presented. We proposed some general directions for improvement of all existing solutions. Afterwards, we gave a brief overview of some selected existing solutions and discussed about their possible specific enchantments. We also acknowledged extensions of APC and possible usages of APC for solving related problems. Solutions to the problem of APC could be used to solve other problems, whenever solving the problem of APC can draw some conclusions regarding those problems.

In our future work we will consider actual implementations of some of the proposed ideas. We also want to put efforts in finding and implementing new ideas related to the social network algorithms.

Our opinion is that, in the near future, the APC is going to be heavily exploited, directly or indirectly inside recommender systems, social networks and expert systems. Therefore, this paper, conceived as an initial analysis research, could announce new scientific and practical work related to the problem of APC.

ACKNOWLEDGMENT

This research was supported by the Ministry of Science, Technology and Development, Republic of Serbia, under the projects 174010 and III44006.

REFERENCES

[1] S. Argamon, P. Chase, S. Dhawle, S. Raj, H. Navendu, and G. S. Levitan, "Stylistic text classification using functional lexical features," Journal of the American Society of Information Science, Baayen 7, pp. 91–109, 2007.

[2] T. Buchanan, J. A. Johnson, and L. R. Goldberg, "Implementing a five-factor personality inventory for use on the internet," European Journal of Psychological Assessment 21, vol. 2, pp. 115–127, 2005.

[3] F. Celli, "Mining user personality in twitter," Tech. rep, 2011.

[4] M. Coltheart, "The MRC psycholinguistic database," Quarterly Journal of Experimental Psychology 33A, pp. 497–505, 1981.

[5] C. Cortes, and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, pp. 273–297, 1995.

[6] Sven Meyer zu Eissen, and Benno Stein, "Intrinsic Plagiarism Detection," In the Proceedings of the European Conference on Information Retrieval (ECIR-06), Springer, 2006.

[7] H. Eysenck, and S. Eysenck, Eysenck Personality Questionnaire-Revised. Hodder, London, 1991.

[8] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," In Proceedings of the 15th International Conference on Machine Learning, pp. 170–178, 1998.

[9] A. J. Gill, Personality and language: The projection and perception of personality in computer-mediated communication. Ph.D. thesis, University of Edinburgh, 2003.

[10] A. J. Gill, S. Nowson, and J. Oberlander, "What are they blogging about? personality, topic and motivation in blogs," unpublished.

[11] A. J. Gill, and J. Oberlander, "Taking care of the linguistic features of extraversion," In Proceedings of the 24th Annual Conference of the Cognitive Science Society, pp. 363–368, 2002.

[12] M. A. K. Halliday, Introduction to Functional Grammar, 2 ed. Edward Arnold, 1994.

[13] F. Iacobelli, A. J. Gill, S. Nowson, and J. Oberlander, "Large scale personality classification of bloggers," In Proceedings of the 4th international conference on Affective computing and intelligent interaction - Volume Part II. ACII'11. Springer-Verlag, Berlin, Heidelberg, pp. 568–577, 2011.

[14] O. P. John, E. M. Donahue, and R. L. Kentle, "The big five inventory: Versions 4a and 5b," Tech. rep., Berkeley: University of California, Institute of Personality and Social Research, 1991.

[15] C. Lyon, R. Barrett, J. Malcolm, "A Theoretical Basis to the Automated Detection of Copying Between Texts, and its Practical Implementation in the Ferret Plagiarism and Collusion Detector," In Plagiarism: Prevention, Practice and Policies Conference, Newcastle, UK, 2004.

[16] F. Mairesse, and M. A. Walker, "Words mark the nerds: Computational models of personality recognition through language," In Proceedings of the 28th Annual Conference of the Cognitive Science Society, 2006.

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

[18] A. Minamikawa, and H. Yokoyama, "Personality estimation based on weblog text classification," In Proceedings of the 24th international conference on Industrial engineering and other applications of applied intelligent systems conference on Modern approaches in applied intelligence - Volume Part II. IEA/AIE'11, Springer-Verlag, Berlin, Heidelberg, pp. 89–97, 2011.

[19] S. Nowson, The language of weblogs: A study of genre and individual differences. Ph.D. thesis, University of Edinburgh, 2006.

[20] S. Nowson, J. Oberlander, and A. J. Gill, "Weblogs, genres and individual differences," In Proceedings of the 27th Annual Conference of the Cognitive Science Society, Cognitive Science Society, pp. 1666–1671, 2005.

[21] J. Oberlander, "Whose thumb is it anyway? Classifying author personality from weblog text," In Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 627–634, 2006.

[22] K. Paul, Text messaging and personality. M.S. thesis, Ball State University, 2011.

[23] J. W. Pennebaker, M. E. Francis, and R. J. Booth, Inquiry and Word Count: LIWC. Lawrence Erlbaum, Mahwah, NJ, 2001.

[24] J. W. Pennebaker, and L. A. King, "Linguistic styles: Language use as an individual difference," Journal of Personality and Social Psychology, vol. 77, pp. 1296–1312, 1999.

[25] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft, "Our twitter profiles, our selves: Predicting personality with twitter," In Proceedings of the 3rd IEEE Conference on Social Computing (SocialCom), 2011.

[26] D. Quercia, R. Lambiotte, M. Kosinski, D. Stillwell, and J. Crowcroft, "The personality of popular facebook users," unpublished.

[27] P. C. Rigby, and A. E. Hassan, "What can oss mailing lists tell us? a preliminary psychometric text analysis of the apache developer mailing list," In Proceedings of the Fourth International Workshop on Mining Software Repositories. MSR '07. IEEE Computer Society, Washington, DC, USA, 2007.

[28] A. Roshchina, J. Cardiff, and P. Rosso, "A comparative evaluation of personality estimation algorithms for the twin recommender system," In Proceedings of the 3rd international workshop on Search and mining user-generated contents. SMUC '11. ACM, New York, NY, USA, pp. 11–18, 2011.

[29] S. Rosset, C. Perlich, and B. Zadrozny, "Ranking-based evaluation of regression models," Knowledge and Information Systems 12, vol. 3, pp. 331–353, 2007.

[30] R. Schapire, "A brief introduction to boosting," In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, vol. 2, pp. 1401–1406, 1999.

[31] S. A. Sushant, S. Argamon, S. Dhawle, and J. W. Pennebaker, "Lexical predictors of personality type," In Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America, 2005.

[32] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu, "Rankclus: Integrating clustering with ranking for heterogenous information network analysis," In Proc. 2009 Int. Conf. Extending Database Technology (EDBT'09), pp. 565–576, Saint Petersburg, Russia, March 2009.

[33] V. Vapnik, The Nature of Statistical Learning Theory. Springer, New York, 1995.

[34] F. Wang, Y. Hong, W. Zhang, and G. Agrawal, "Personality based latent friendship mining," In DMIN (2009-10-28), R. Stahlbock, S. F. Crone, and S. Lessmann, Eds. CSREA Press, pp. 427–433, 2009.

[35] I. H. Witten, E. and Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, 1999.

[36] T. Yarkoni, "Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers," Journal of Research in Personality 44, vol. 3 (June), pp. 363–373, 2010.