Cross-Cultural Personality Prediction based on Twitter Data

Cagatay Catal[1*], Min Song[2], Can Muratli[1], Erin Hea-Jin Kim[2], Mestan Ali Tosuner[1], Yusuf Kayikci[1]

[1] Istanbul Kultur University, Department of Computer Engineering, Bakirkoy, Istanbul, Turkey.
[2] Yonsei University, Department of Library and Information Science, Seoul, Republic of Korea.

* Corresponding author. Tel.: +902124984215; email: c.catal@iku.edu.tr
Manuscript submitted August 25, 2017; accepted October 13, 2017.
doi: 10.17706/jsw.12.11.882-891

Abstract: Social networking platforms such as Facebook, Twitter, YouTube, and Instagram, which generate a vast amount of data every second, emerged dramatically within the last ten years. This huge rich data provides crucial information about social interactions and human behaviour. Therefore, it is possible to identify the personality traits of a person by extracting and analysing relevant information from the social media. Recently, researchers demonstrated that personality prediction can be performed by using historical textual features and user profiles. While cross-cultural personality research is considered as a powerful tool to observe the differences in personality psychology, machine learning researchers analysing social media data did not perform research on the development of cross-cultural personality prediction models yet. In this paper, we propose a new research topic called cross-cultural personality prediction and discuss how these kinds of models can be built in practice. In the joint research project, we investigate whether there is a cultural or language difference in personality between people in two countries namely, South Korea and Turkey in Twitter. It will be particularly interesting because both of the countries are in Asia, but located in both ends of Asia and we'll investigate whether there is any meaningful personality difference in two countries or not. We argue that this kind of research will impact not only personality services such as IBM Personality Insight, but also cross-cultural personality psychology research area in the near future.

Key words: Personality prediction, regression analysis, social media mining, supervised learning.

1. Introduction

Personality is defined as "consistent behaviour patterns and intrapersonal processes originating within the individual" in psychology [1]. It is considered to be affected by environment and culture is the most influential factor for the environment [2]. Culture is defined as the knowledge network distributed among interconnected people [3]. Cultural studies are more difficult to perform as they require translations, new hypotheses, overseas trips, and networking with researchers from different countries [2]. These challenges provide exciting opportunities compared to the traditional research approaches [4]. Nowadays, social networking platforms produce a great amount of data every second. For instance, around 500 million tweets per day are sent on Twitter, which corresponds to 6,000 tweets per second and 200 billion tweets per year [5]. This data might be used for different purposes such as personality prediction.

Predicting personality of a user can contribute for many application areas such as personalized recommendations, marketing strategies, business intelligence, sociology, human resources management,

and mental diagnosis [6, 7]. Recently, researchers demonstrated that personality prediction can be performed by using historical textual features [8] and user profiles [9]. Most of the personality prediction studies used Facebook data [10]-[15] as it is the largest social networking platform in the world by the number of users (1,374,000,000 users) [16]. Some researchers focused on Twitter data [6], [8], [9], [18]-[20] to predict the personality traits as Twitter is one of the most popular social networks which has 289 million users [17] all around the world. A recent study [21] proposed a personality prediction model which uses activities of users in Twitter and Instagram. 62 users' tweets and Instagram images were used in conjunction with Random Forest Regression algorithm to build the model and this combination decreases the error for each personality dimension.

Most of these studies applied Big Five Model (a.k.a., Five Factor Model) [22] which has five dimensions (extraversion, agreeableness, conscientiousness, Neuroticism, openness) to represent the personality of a person. In these studies, a questionnaire having 44 questions [23] is filled by the user and then, values of each personality dimension are automatically calculated and recorded in the training datasets. Five-level Likert scale (Disagree strongly / Disagree a little / Neither agree nor disagree / Agree a little / Agree strongly) is used for the answers. Trait scores are calculated based on the answers for a question set. For example, extraversion is calculated based on the following questions: 1, 6R, 11, 16, 21R, 26, 31R, 36. R indicates that the question is reverse-scored. In addition to these values which will later be predicted, researchers collect several features such as the number of followers, number of following, number of retweets, and the frequency of words depending on the type of the social networking platform. Big Five Model [24] is used as a model of Trait Theory in psychology. Since many researchers confirmed that there are five personality traits based on the available data, this model is called Big Five Model. In addition to the Trait Theory, there are additional five theories called psychoanalytic, biological, humanistic, behavioural/social theory, and cognitive theory in psychology to explain different aspects of personality.

In this paper, we attempt to investigate whether cross-cultural prediction models based on Twitter data can be built or not and how these kinds of models can be implemented. We also conducted some preliminary experiments to create a personality prediction model for a country. While there are many software engineering papers which utilize from cross-company or cross-project data such as cross-project defect prediction [25]-[31] and cross-company effort estimation [32]-[38], there is no cross-cultural personality prediction study yet. Therefore, we see a big research potential on the analysis of cross-cultural personality prediction models.

2. Related Work

Most of the personality prediction studies aim to predict personality traits by using social media data. Results of the predictions are mostly given based on the Big Five Model as shown in Fig. 1. This field is very challenging for both academic researchers and companies in software industry. For example, IBM has a service called IBM Watson Personality Insights which provides an API for programmers to extract insights from social media accounts of users and predict personality traits of the users. IBM listed some potential applications of personality prediction as follows: Targeted marketing, customer acquisition, customer care, personal connections [39]. In addition to the Facebook and Twitter data usage for the personality prediction, some researchers developed personality prediction models based on mobile phone logs [40], [41]. Another research group [42] showed that some visual patterns are correlated with personality traits of users and personality of a user can be predicted from the favorite images on Flickr online photo-sharing platform. Data from a blogging site called Livejournal was used to estimate the influence and personality traits of a person in a research study [43]. In another study, YouTube data was used as the source domain to build the training model and extraversion trait of person was predicted in the small group meeting [44]. In

addition, researchers developed a personality prediction model by using speech data and Support Vector Machine (SVM) Regression algorithm [45]. As shown in the studies explained above, personality of a person can be predicted not only based on textual data, but also user profile, favorite pictures, Facebook likes, video, speech, and blogging site data. In our cross-cultural personality prediction project, we will first investigate the use of textual data.

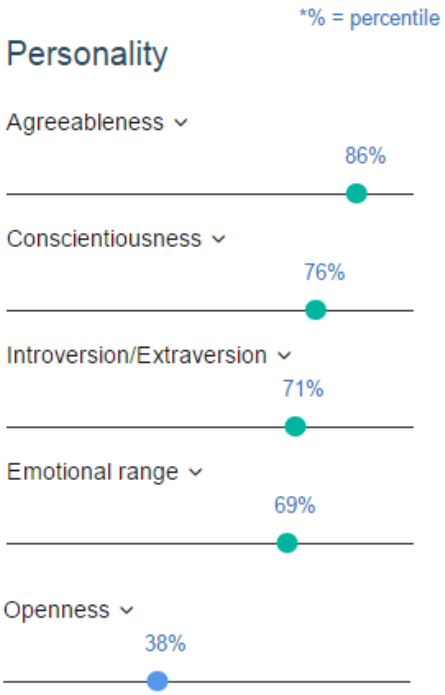

Fig. 1. Results of a personality prediction model.

3. Methodology

In this joint research project, we build two personality prediction models based on Korean and Turkish Twitter users' features. The first model which uses Korean users' dataset are used to predict the personality traits of Turkish Twitters users and the second model which uses Turkish users' dataset are used to predict the personality traits of Korean Twitter users. This research area can be called as Cross-Cultural Personality Prediction and there is no study in the literature on the use of cross-culture for predicting personality traits. While South Korea is in eastern sub region of the Asian continent, Turkey is in Western Asia. In addition to this analysis, we investigate the top k words (i.e., k = 20) for each personality dimension and check whether there are any cross-cultural words which can reflect a specific personality trait or not. We create two public personality datasets for Turkish and Korean Twitter users. We also work on several features such as the number of followers, number of retweets, number of hashtags, frequency of words, and several machine learning algorithms to build a high-performance regression model. We respond to the following research questions:

RQ-1: Is cross-cultural personality prediction feasible?

RQ-2: Are there any cross-cultural words which can determine personality traits in big five model?

RQ-3: Is there any effect of gender in Western Asia and Eastern Asia on the use of specific words?

RQ-4: How efficient is personality prediction models for Turkish and Korean Twitter users?

We translate the questionnaire [23] having 44 questions written in English into Turkish and Korean languages and distribute to the Twitter users via a tweet. We collect the latest 3000 tweets of users and the

other features of the users. All the tweets of users are written to MS SQL Server database and a dictionary is prepared to assign unique IDs for each specific word in the dictionary. After the best prediction models are built for Turkish and Korean users, we have to combine these two dictionaries so that we can investigate whether there are similar words between personality traits of users or not. This combined dictionary is called dictionary_with_two_cultures. The borders of the personality traits (lower bound analysis: 0-20% and upper bound analysis 80-100%) is analyzed to identify the common words for each dimension and then, we check if there are similar words in these intervals. We accept the Big Five Model and questionnaire [23] as universal to apply in two cultures.

The following regression algorithms which exist in Azure Machine Learning (ML) Studio [46] is analyzed to build the regression models: ordinal regression, Poisson regression, fast forest quantile regression, linear regression, Bayesian linear regression, neural network regression, decision forest regression, boosted decision tree regression. In addition to these regression algorithms, deep learning algorithms are applied to improve the performance of the models. In addition to the term frequency features, term weights such as TF-IDF and entropy is also considered. We evaluate models using N-fold cross-validation approach. The best model in terms of MAPE (mean absolute percentage error) and coefficient of determination parameters is used to produce the web service in Azure ML Studio environment. A mobile application is developed as the client application and the personality prediction results is sent to the client when the Twitter username is submitted to the prediction web service which is hosted on Azure cloud platform.

Some natural language processing libraries must be used to process the words extracted from tweets. Zemberek library [47] is used for this purpose in Turkish tweets. For example, the root of the words must be found to increment the frequency of a word. Otherwise, we'd have many similar words in the dictionary and we'd not be able to build a high-performance prediction model. A similar library is used for Korean tweets. After we build the best model for Turkish, we apply this model to predict the personality traits of Korean Twitter users and we investigate whether it works or not. Also, the reverse is performed. The following technical steps have been identified:

- *Implement 44-questions questionnaire form, which is applied to measure personality traits of users based on Big Five Model [23], by using Asp.net technology*

We provide 44 characteristics to the user shown as follows and the user selects one of the five categories (Disagree strongly-1, Disagree a little-2, Neither agree nor disagree-3, Agree a little-4, Agree Strongly-5) as the response. To save space, here we show only 10 characteristics. The complete characteristics can be accessed from the reference 23.

I see Myself as Someone Who...

____1. Is talkative
____2. Tends to find fault with others
____3. Does a thorough job
____4. Is depressed, blue
____5. Is original, comes up with new ideas

Scoring is performed based on the following related questions [23] for each dimension. R indicates that this is a reverse-scored item. The answers for each question are added numerically. If there is a R sign for a question, the result is subtracted from 6 and the value is added to the result. For example, for extraversion dimension, we add the results of questions 1, 11, 16, 26, 36; we subtract the results of questions 6, 21, 31 from the value 6, and add the values to the result. For example, if the answers for questions 1/11/16/26/36 are "agree strongly", and the answers for questions 6/21/31 are "agree a little", then the result will be

calculated as follows: [(5*5)+(6-4)+(6-4)+(6-4)] = 25 + 6=31. Minimum and maximum values for each dimension are represented at the right hand side of the dimensions.

Extraversion: 1, 6R, 11, 16, 21R, 26, 31R, 36 (min 8, max 40)

Agreeableness: 2R, 7, 12R, 17, 22, 27R, 32, 37R, 42 (min 9, max 45)

Conscientiousness: 3, 8R, 13, 18R, 23R, 28, 33, 38, 43R (min 9, max 45)

Neuroticism: 4, 9R, 14, 19, 24R, 29, 34R, 39 (min 8, max 40)

Openness: 5, 10, 15, 20, 25, 30, 35R, 40, 41R, 44 (min 10, max 50)

After this calculation, we have the values for each personality dimension of the user and they can be used as ground truth.

- *Extract the latest 3000 tweets and several features of voluntary users such as number of retweets*

Twitter API is used to retrieve the tweets and the other features (number of followers, number of following, number of friends, etc.) of the user. The following resource URL will be used to get them:
https://api.twitter.com/1.1/statuses/user_timeline.json
JSON (JavaScript Object Notation) is the response format. An example response is shown in the following link: https://dev.twitter.com/rest/reference/get/statuses/user_timeline

- *Insert data and responses of users into SQL Server database.*

In order to insert data into the MS SQL Server database, we need some tables. During the design phase, these tables are determined and data&features are written to the appropriate tables. The following tables are used during the design phase: question, answer, option, trait, score, user, tweets, tweet_feature. Tweet_feature table store the data about tweets such as hashtag, and mention. Tweets table store username and tweets of that user. User table is used to store the number of followers, number of following, number of tweets, and the favorite tweet. The remaining five tables are used for gathering the results of the questions.

- *Process all the participants' database records to prepare csv formatted training file.*

When the required number of participants for the study (i.e., 100) are satisfied, all the database records are processed and each unique word has an ID. Features and word frequencies of a user are written as a row in the csv formatted training file. Training file is ready once all the users' data is written to this file.

- *Design several experiments on Azure ML Studio to identify the best regression algorithm in terms of MAPE and coefficient of determination evaluation parameters.*

Training file created at step 4 is loaded into Azure Machine Learning Studio platform and then, several regression algorithms are evaluated. After the regression algorithm is added to the experiment screen, "train model", "score model" and "evaluate model" components are integrated into this screen. "Train model" is used to find the parameters of the algorithm and "score model" checks the model against a dataset, and "evaluate model" gets the evaluation parameters or evaluates the model against another model. The best model is identified based on MAPE and coefficient of determination values. MAPE must be minimum and the other parameter must be near to 1.

- *Publish the best model as a web service on Azure platform.*

"Web service input" and "Web service output" are added into the experiment and then, it is deployed as a web service.

- *Implement a mobile application for App Store to interact with the users who wish to know their personality traits.*

A mobile application is developed for iOS platform by using Swift languages. It sends the user name to the web service and gets the prediction results from the web service.

- *Connect the mobile app to the web service and inform the user about the data retrieved from web service.*

This mobile app consumes the personality prediction web service and informs the user with the prediction results. Graphical user interface is easy to use.

- *Identify the top k words at the two borders of personality traits in training file (0-20%) and (80-100%) are lower and upper bounds.*

We investigate the popular words preferred by a user group who has personality trait values at one border. For example, for extraversion dimension we analyze the words of users who have over 80% (percentile) extraversion values and users who have less than 20% values. These words might give interesting results to investigate.

These nine steps must be followed by two research groups (South Korea and Turkey). At the end of these steps, we have two training files (i.e., Turkish_train.csv and Korean_train.csv), two personality prediction web services, and two mobile applications.

In addition, we identified the following steps:

- *Retrieve two training files of two cultures.*

Turkish_train.csv and Korean_train.csv files are retrieved.

- *Perform step 9 for the training files of each culture and analyze the similarities between these words.*

We check whether there are universal words which can help to detect one personality trait or not. To do this, we convert our words into English in both of the training files and check the similarities.

- *Apply Turkish prediction web service to predict the data in Korean training file.*

We analyze whether Turkish personality model works for Korean Twitter users or not. MAPE and coefficient of determination parameters are evaluated for this decision.

- *Apply Korean prediction web service to predict the data in Turkish training file.*

We analyze whether Korean personality model works for Turkish Twitter users or not.

- *Analyze gender effect on two training files.*

We determine whether there are similarities in terms of word usage among a gender or not. Also, this is evaluated from the cultural perspective to know whether there are similar words which can determine a personality trait among a gender across cultures or not.

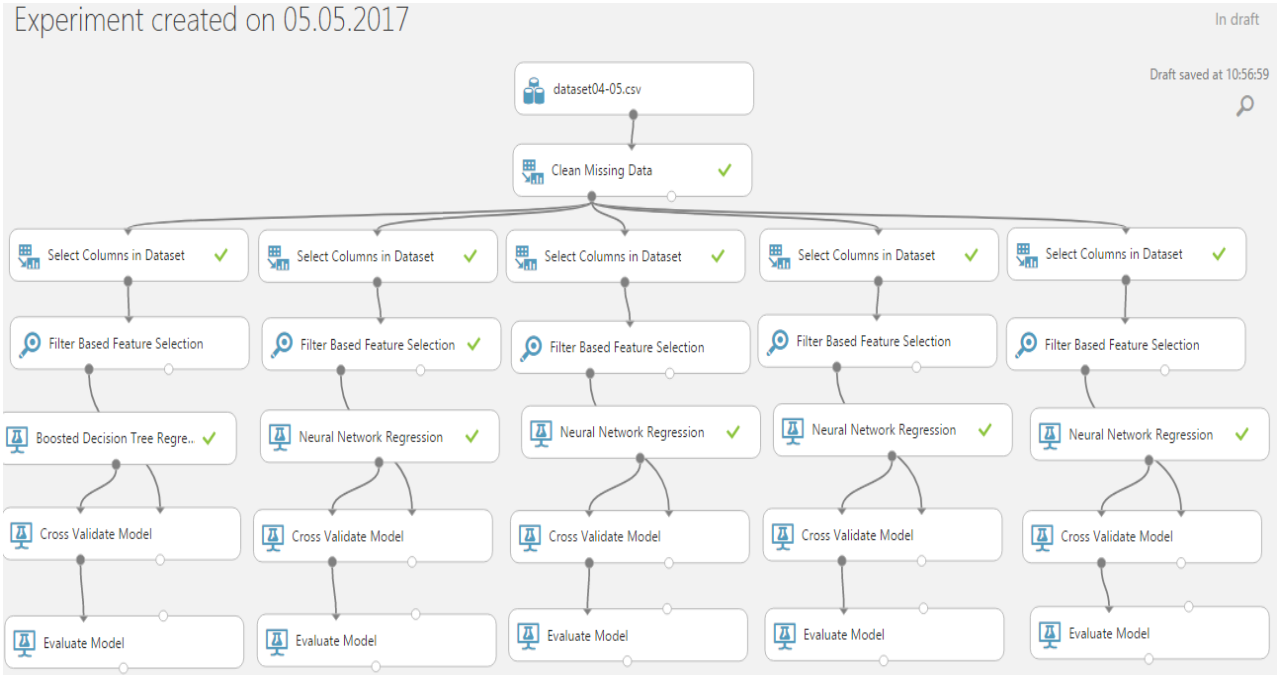

Fig. 2. Experiments using regression algorithms.

4. Experiments

TwitterSharp API version 2.3.1 was used to retrieve the tweets of the users. In the initial study, 50 voluntary subjects joined to this analysis in Istanbul, Turkey. Zemberek natural language processing (NLP) library was used to find the roots of the words. A personality prediction dataset was prepared for Turkish Twitter users and shared in the following link. https://www.dropbox.com/sh/g0jp8m1vvmkauru/AABuUL-ER0_6hhP2_xVwsxRla?dl=0 Experiments are being performed on Azure ML Studio platform as shown in Fig. 2.

A similar dataset is being prepared for Korean users, too. Later, cross-cultural prediction studies will be performed.

5. Conclusion

In this paper, we proposed a new research area called cross-cultural personality prediction based on social media data. In addition, we showed the required technical steps to perform this kind of research. This field will not only affect the personality psychology, but also personality prediction services used in software industry. The software systems have been developed and Twitter data has been collected in Turkey. Currently, different machine learning algorithms are being evaluated for five dimensions of the personality. Stemming and the other necessary steps have been performed as part of natural language processing for Twitter data. After the best models are identified for each culture, cross-cultural prediction experiments will be performed.

References

[1] Burger, J. M. (2014). *Personality*. Wadsworth Publishing.

[2] Benet-Martínez, V. (2008). Cross-cultural personality research. *Handbook of Research Methods in Personality Psychology*, 170-189.

[3] Chiu, C., & Chen, J. (2004). Symbols and interactions: Application of the CCC model to culture, language, and social identity. *Language Matters: Communication, Culture, and Social Identity*. 155-182.

[4] Matsumoto, D. (2000). Culture and psychology: People around the world. USA: Wadsworth.

[5] Twitter Usage Statistics. (2017). Retrieved March 6, from http://www.internetlivestats.com/twitter-statistics

[6] Lima, A. C. E., & De Castro, L. N. (2014). A multi-label, semi-supervised classification approach applied to personality prediction in social media. *Neural Networks*, *58*, 122-130.

[7] Wei, H., Zhang, F., Yuan, N. J., Cao, C., Fu, H., Xie, X., Rui, Y., & Ma, W. Y. (2017). Beyond the words: Predicting user personality from heterogeneous information. *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* (pp. 305-314).

[8] Golbeck, J., Robles, C., Edmondson, M., & Turner, K. (2011). Predicting personality from twitter. *Proceedings of the 2011 IEEE Third International Conference on Social Computing*.

[9] Quercia, D., Kosinski, M., Stillwell, D., & Crowcroft, J. (2011). Our twitter profiles, our selves: Predicting personality with twitter. *Proceedings of the 2011 IEEE Third International Conference on Social Computing*.

[10] Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110(15)*, 5802-5805.

[11] Wald, R., Khoshgoftaar, T., & Sumner, C. (2012). Machine prediction of personality from facebook profiles. *Proceedings of the 2012 IEEE 13th International Conference on Information Reuse and Integration* (pp. 109-115).

[12] Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., & Seligman, M. E.

(2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, *108(6)*, 934.

[13] Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., & Seligman, M. E. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, *108(6)*, 934.

[14] Ortigosa, A., Carro, R. M., & Quiroga, J. I. (2014). Predicting user personality by mining social interactions in Facebook. *Journal of computer and System Sciences*, *80(1)*, 57-71.

[15] Markovikj, D., Gievska, S., Kosinski, M., & Stillwell, D. (2013). Mining facebook data for predictive personality modeling. *Proceedings of the 7th international AAAI conference on Weblogs and Social Media*, Boston, MA, USA.

[16] Chen, T. Y., Chen, T. Y., Tsai, M. C., Tsai, M. C., Chen, Y. M., & Chen, Y. M. (2016). A user's personality prediction approach by mining network interaction behaviors on Facebook. *Online Inf. Review*, *40(7)*, 913-937.

[17] Statistic Brain. (2016). Social Networking Statistics. Retrieved from statistic brain: http://www.statisticbrain.com/social-networking-statistics/

[18] Sumner, C., Byers, A., Boochever, R., & Park, G. J. (2012). Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets. *Proceedings of the* 2012 *11th International Conference on Machine Learning and Applications* (pp. 386-393).

[19] Adali, S., & Golbeck, J. (2012). Predicting personality with social behavior. *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 302-309).

[20] Gou, L., Zhou, M. X., & Yang, H. (2014). Knowme and shareme: Understanding automatically discovered personality traits from social media and user sharing preferences. *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems* (pp. 955-964).

[21] Skowron, M., Tkalčič, M., Ferwerda, B., & Schedl, M. (2016). Fusing social media cues: personality prediction from twitter and instagram. *Proceedings of the 25th International Conference Companion on World Wide Web* (pp. 107-108).

[22] McCrae, R. R., & John, O. P. (1992). An introduction to the five factor model and its applications. *Journal of Personality*, *60(2)*, 175-215.

[23] Personality-BigFiveInventory. Retrieved from: http://fetzer.org/sites/default/files/images/stories/pdf/selfmeasures/Personality-BigFiveInventory.pdf

[24] De, R. B., (2000). *The Big Five Personality Factors: The Psycholexical Approach to Personality*. Hogrefe & Huber Publishers.

[25] Turhan, B., Menzies, T., Bener, A. B., & Di, S. J. (2009). On the relative value of cross-company and within-company data for defect prediction. *Empirical Software Engineering*, *14(5)*, 540-578.

[26] Zimmermann, T., Nagappan, N., Gall, H., Giger, E., & Murphy, B. (2009, August). Cross-project defect prediction: A large scale experiment on data vs. domain vs. process. *Proceedings of the* 7th *Joint Meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering* (pp. 91-100). ACM.

[27] Ma, Y., Luo, G., Zeng, X., & Chen, A. (2012). Transfer learning for cross-company software defect prediction. *Information and Software Technology*, *54(3)*, 248-256.

[28] Peters, F., Menzies, T., & Marcus, A. (2013, May). Better cross company defect prediction. *Proceedings of the 2013 10th IEEE Working Conference on Mining Software Repositories* (pp. 409-418).

[29] Chen, L., Fang, B., Shang, Z., & Tang, Y. (2015). Negative samples reduction in cross-company software

defects prediction, *Information and Software Technology*, *62*, 67-77.

[30] Zhang, F., Zheng, Q., Zou, Y., & Hassan, A. E. (2016, May). Cross-project defect prediction using a connectivity-based unsupervised classifier. *Proceedings of the 38th International Conference on Software Engineering* (pp. 309-320).

[31] Yu, Q., Jiang, S., & Zhang, Y. (2017). A feature matching and transfer approach for cross-company defect prediction. *Journal of Systems and Software*.

[32] Minku, L. L. (2016, September). On the terms within-and cross-company in software effort estimation. *Proceedings of the 12th International Conference on Predictive Models and Data Analytics in Software Engineering*.

[33] Tong, S., He, Q., Chen, Y., Yang, Y., & Shen, B. (2016, December). Heterogeneous cross-company effort estimation through transfer learning. *Proceedings of the 2016 23rd Asia-Pacific Software Engineering Conference* (pp. 169-176).

[34] Minku, L. L., & Yao, X. (2014, May). How to make best use of cross-company data in software effort estimation?. *Proceedings of the 36th International Conference on Software Engineering* (pp. 446-456).

[35] Minku, L., Sarro, F., Mendes, E., & Ferrucci, F. (2015, October). How to make best use of cross-Company data for web effort estimation?.

[36] Turhan, B., & Mendes, E. (2014, August). A comparison of cross-versus single-company effort prediction models for web projects. *Proceedings of the 2014 40th EUROMICRO Conference on In Software Engineering and Advanced Applications* (pp. 285-292).

[37] Mendes, E., Kalinowski, M., Martins, D., Ferrucci, F., & Sarro, F. (2014, May). Cross-vs. within-company cost estimation studies revisited: An extended systematic review. P*roceedings of the 18th International Conference on Evaluation and Assessment in Software Eng.*

[38] Kocaguneli, E., Cukic, B., Menzies, T., & Lu, H. (2013, October). Building a second opinion: learning cross-company data. *Proceedings of the 9th Int'l Conference on Predictive Models in Software Eng*.

[39] Developercloud. Retrieved from https://www.ibm.com/watson/developercloud/doc/personality-insights/index.shtml

[40] De Montjoye, Y. A., Quoidbach, J., Robic, F., & Pentland, A. S. (2013). Predicting personality using novel mobile phone-based metrics. In International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction (pp. 48-55).

[41] Staiano, J., Lepri, B., Aharony, N., Pianesi, F., Sebe, N., & Pentland, A. (2012). Friends don't lie: Inferring personality traits from social network structure. *Proceedings of the 2012 ACM Conference on Ubiquitous Computing* (pp. 321-330).

[42] Cristani, M., Vinciarelli, A., Segalin, C., & Perina, A. (2013). Unveiling the multimedia unconscious: Implicit cognitive processes and multimedia content analysis. *Proceedings of the 21st ACM int'l conference on Multimedia* (pp. 213-222).

[43] Nguyen, T., Phung, D. Q., Adams, B., & Venkatesh, S. (2011). Towards discovery of influence and personality traits through social link prediction.

[44] Aran, O., & Gatica-Perez, D. (2013). Cross-domain personality prediction: from video blogs to small group meetings. *Proceedings of the 15th ACM on International Conference on Multimodal Interaction* (pp. 127-130).

[45] Tim, P., Katrin, S., Sebastian, M., Florian, M., Mohammadi, G., & Vinciarelli, A. (2012). On speaker-independent personality perception and prediction from speech.

[46] Azureml. Retrieved from https://studio.azureml.net/

[47] Ahmetaa. Retrieved from https://github.com/ahmetaa/zemberek-nlp

Cagatay Catal is with the Department of Computer Engineering, Istanbul Kültür University, Turkey

Dr. Cagatay Catal is an associate professor and head of Department at the Department of Computer Engineering in Istanbul Kültür University, Turkey. He received the BS & MSc degrees in computer engineering from Istanbul Technical University and the PhD degree in computer engineering from Yildiz Technical University, Istanbul. He worked 8 years at the Research Council of Turkey (TUBITAK), Information Technologies Institute as senior researcher and involved in the development of several large-scale software-intensive system projects. He's been working in Istanbul Kültür University for 6 years. He has over 60 peer reviewed publications on software engineering and data science in international journals, books, and conferences. His research interests include software engineering, data science, machine learning, big data, and software testing. He is external reviewer for Research Council of Canada, Research Council of Turkey (TUBITAK), and EUROSTARS program funded by European Union.