EXTRACTING PERSONALITY FEATURES FROM USER GENERATED DATA FOR RECRUITMENT VIA TEXT MINING

[1]PELINVARDARLIER, [2]GOKHANSILAHTAROGLU

[1]Istanbul Medipol University,Department of Human Resources,
[2]Istanbul Medipol University, Department of Management Information Systems,
E-mail: [1]pvardarlier@medipol.edu.tr, [2]gsilahtaroglu@medipol.edu.tr

Abstract- Social media has been used for different purposes under the name of big data to extract useful information. In the field of human resources social media is used manually by managers and operational staff especially for recruitment. On the other hand personality tests are some other instruments to evaluate candidates for hiring process. In this study, we used social media of 30 volunteers to find out their personality traits like dominant, compliance, steadiness, influence. We have interviewed and given them the D.I.S.C test and we have evaluated them with the answers they have given. After this manual assessment process with the permit of those volunteers, we have extracted their tweets from social media i.e Twitter. These text data have been converted into a data warehouse that can be processed with text mining tools. After that, clustering has been applied to the data. Clustering results have been cross validated with the manual assessment results. For each cluster, common key words are discovered to represent characteristics like dominant, compliance, steadiness, influence. Our study shows that, when social media entries or tweets are examined with text mining tools, some words may give a clue about the character of the user. Although this study has used texts produced on social media, any user generated data may be applicable for the test or analysis.

Keywords- Social media; human resources; personality tests; big data; text mining.

I.INTRODUCTION

Data mining is to extract valuable and useful information from raw data,(Han & Kamber, 2011). In other words, it converts raw data into information which can be used for business purposes in many ways. To do this, researchers use a variety of data mining models, techniques, algorithms and tools. In general data mining has three models: Clustering, classification and association rule discovery, (Silahtaroğlu, Veri Madenciliği Kavram ve Algoritmaları, 2013). Clustering employs unsupervised algorithms to learn the hidden facts in large data sets. It is also called segmentation. Clustering techniques and algorithms divide a data set into previously unknown segments. In this way, it discovers information like customer profile, (Chapelle, Schölkopf, & Zien, 2006), customer brand association, (Agrawal & Shafer, 1996), geographic areas and economic parameters, (Roghani, 2015)and so on. Classification uses supervised machine learning algorithms to learn rules to predict some values to appear or come true in the future such as percentage of a price rise(Yuhong & Weihua, 2010),currency ratios (Roghani, 2015), number of visitors to arrive at a hotel or a shop on a future date(Domingues, Soares, & Jorge, 2013) etc. Artificial Neural Networks (ANN) (Graupe, 2013)and Decision Tree algorithms (Perzyk & Soroczynski, 2010)are very useful for this kind of analyses. Association rule discovery algorithms are used to find out the hidden patterns in events. For example, how and why people visit a certain shop on a certain date (Silahtaroğlu, 2015)and what kind of pattern(s) they draw when they shop(Aikaterini, Frangosb, & Frangos, 2013) etc. If a researcher wants to learn the hidden motives some time before an economic crisis, s/he should use association rule discovery tools and algorithms, (Wang P & H., 2002). Furthermore, data mining may be applied to human resources discipline as well. It may be used as a tool during recruitment process. Although a person’s education, experience, skills are very important, his/her personality traits also play a prominent role in professional life. When a person handles a piece of work, both speedand conscientiousness depend not only on the skills, experience and education of the person but also his/her personal characteristics. That’s why many firms conduct a personality assessment procedure when they hire new employees. It is an important phase of recruitment. In this study we present an alternative way to conduct personality assessment procedure with the help of text mining and big data. We show that when online user generated data are analyzed with text mining tools and algorithms we can get valuable information for staff recruitment. In the study we also present the phases to convert UGD into a proper text mining data set.

II. PERSONALITY ASSESSMENT

In the course of the history, competition in business has always been a hot topic. Companies used different tools or instruments to achieve competitive advantage over their rivals. In the classical understanding, capital, innovation or technology are vastly used for competitive advantage however in the modern world and business life human resources have been another point to achieve competitive advantage (Stanujkic, 2015), in addition to this, human resources have become important for strategic management as well (Petkovic M, 2005). So today, companies pay more attention to recruitment than they did in the past. Interviewing, selecting, and orienting new employees are as important as production, marketing or accounting(Arthur, 2012). When it comes to assess a candidate, recruiters in companies consider the requirements of a particular job by analyzing the knowledge, abilities, skills and other requirements needed to perform the job fully and meticulously (Pulakos, 2005), (Cook, 2005). Interviewing and assessment of the candidates are done under the light of all these. Nonetheless, besides the talents and skills of the candidates, their personality, intelligence, psychology etc. are also important assets in professional life (Chaturvedi, 2014). That's why many companies today apply different types of test such as, intelligence tests, psycho tests, personality tests, cognitive tests and so on in order to evaluate candidates (Miller, 2014), (SHRM, 2012). Applications in research and professional experiences prove that there is a close relation between employees' personality, interpersonal style, his / her response to stress situations and requirements and performance of the job (Piotrowski, 2006). For example, one person may like details as another hates so, why not consider this when forming a work group or team in other words recruitment.

In the literature there are a number of personality assessment tests. 16 PF Personality Inventory, Thalento Big 5 Personality, Thomas Personality Inventory, PERI Personality Inventory, NNA Personality Inventory and DISC are some of them. DISC is one of the most commonly used and reliable method to characterize the people's personality (Marston, 2002). DISC encompasses four quadrant behavioral model to examine the behavior of individuals both in their everyday environment and / or in a specific situation (Marston, 2002). There are 28 questions on the test that can be answered in 10 ten minutes or so. After answering survey like questions, the persons profile is extracted through his / he answers. A person may be clustered in one or two personality dimensions (Kim, Jang, & Shin, 2008): D,I,S or C. DISC is an acronym which stands for Dominance, Influence, Steadiness and Conscientiousness. Dominance is related to control, power and assertiveness. People who have this feature in his or her character are probably independent and results driven. They are tendingto take direct action and enjoy challenges. Influence is associated with social situations and communication. Those who have this trait are social and outgoing. They are active team players and enjoy sharing thoughts. Steadiness also refers to being a team player or cooperative. This feature implies that the individual is patient, thoughtful and persistent. They are good listeners and supportive. However they avoid conflicts and keep away from radical changes (Marston, 2002). For example if you want to hire a system analyst to redesign your system or change most of the things in the system, a candidate who bears 'steadiness' will not be appropriate for this post. Conscientiousness is also called as caution and related to structure and organization. These people always have A, B and plans. They like details and always check accuracy. It is quite normal for these people to ask questions like "Why are we doing this?", "How are we going to proceed after that point?" etc.(Marston, 2002).

III. EXTRACTING PERSONALITY CHARACTERISTICS FROM USER GENERATED DATA

a. Data Collection

45 volunteers have participated in this study. They were given a DISC survey to fill out. We haveassessed their personality characteristics from the answers they gave. We shared our assessments with them and they confirmed their personality characteristics both in their everyday environment and in a specific situation like work place, school, and social club and so on. Those who did not confirm our assessment or findings have been taken out from the study so, for data mining analysis we have used 32 persons' test results which were confirmed.

In the second part of data collection, we used KNIME program's Twitter API module to extract tweets of those 32 persons. All of the volunteers have been chosen among those who use Twitter regularly and have hundreds of Tweets. After cleaning data from re-tweets we have decided on some specific words that they may have used in their tweets.We accepted a word as "used" if it is used at least three times by the volunteer.

The words we have chosen to extract and analyze are: But, or, and, yes, no, maybe, namely, never, because of, due to, because, why, who, where, how, what, which, though, although, in addition, for instance, for example.

b. Method and Model

After that we conducted a decision tree analysis to see if specific character types can be inferred or judged from the words they use. For decision tree analysis we used GINI index algorithm(Shafer J.C., 1997)within KNIME program. Although we had only thirty two persons to analyze when the data warehouse has been formed with their tweets and found personality traits we have ended up with a huge data set which consists of thousands of records. However we did not prefer a predicting or scoring session for our model because we had only thirty two persons and it was too few to avoid over learning which decision tree algorithms may suffer from.

c. Analysis and Findings

What we found in this study shows that there is a correlation between the words a person use on Twitter and his / her personal characteristics.

We conducted decision tree learning system in two phases: one to learn persons' prominent characteristics in their natural environment or safe haven like family the other is the mask that they may wear outside their safe haven like school or professional business life. Both analyses have yielded satisfying results. Here we list some of the findings:

Findings for persons' masked characteristics:

The most important word to give a hint about the character of a person is HOW.

If a person does not use HOW, s/he has a chance of having I type character. s/he has no probability to be type of character S or D. There is only a chance of 25% that s/he may bear character type C.

If a person uses HOW but does not use NEVER, s/he may be S type character (66.6 %) but has no chance to be I or C.

If s/he uses HOW, NEVER, but does not say MAYBE then this person is a character type D with a probability of 100 %.

Using HOW, NEVER, MAYBE, NAMELY and YES very often reveals that the person is a I type character.

In the second phase of the study we analyzed and learnt persons' characteristics in their natural environment. Here are the findings:

The most important word to give clue about the character type is NAMELY. Often usage of this word suggests that the person is D type character with a probability of 44%, s/he has got only 8% chance to be character C.

Surprisingly enough dominant characters do not use NO in their natural environment like family but they use WHY and we may infera D type character with a probability of 60 % by checking the usage of NOs and WHY. If one uses WHY and NO, s/he is probably D.

If a person uses BUT, HOWEVER and WHY that's a S character, if WHY is not used it reveals a C character.

CONCLUSION

In this study, we conducted a research in order to find if some certain words used by people on their Twitter accounts may reveal the personal characteristics of the person so that it could be used for recruitment. We used DISC personality test to learn a person's prominent characteristics. DISC is commonly used for recruitment. Our findings have been approved by the volunteers and then we collected their Tweets to build a data warehouse to be used for analysis. Further on we decided some key words. These are;

But, or, and, yes, no, maybe, namely, never, because of, due to, because, why, who, where, how, what, which, though, although, in addition, for instance, for example.

We merged tested, assessed and approved personal characteristics data table with text the database which is taken from Twitter so that we would perform data mining analysis. For analysis we used Gini decision tree algorithm with in KNIME data mining platform. Our findings show that we may predict one's personal character type by checking the usage of some certain words with a probability of %60 to %100, However, for some certain character Types these words are not enough to make a prediction. In addition to this, although we have chosen only 22 words for analysis, only seven of them have given clues about a person's character type. These words are; YES, NO, HOW, WHY, NEVER, MAYBE, NAMELY, BUT, HOWEVER.

This study has been realized with 32 volunteers and 22 words. In this study we have aimed to show that in a detailed study, more accurate results may be held. In fact, study shows that with more volunteers and more words, it is possible to infer ones character type. So it may be used during recruitment process.

REFERENCES

[1]. A.Russell, M. (2013). Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More . O'REILLY MEDIA.

[2]. Agrawal, R., & Shafer, J. C. (1996). Parallel Mining of Association Rules. Proceedings of IEEE Transactions on Knowledge and Data Engineering, 6(8), 962- 969.

[3]. Aikaterini, C. V., Frangosb, C. C., & Frangos, C. C. (2013). Online and mobile customer behaviour: a critical evaluation of Grounded Theory studies. Behaviour & Information Technology, 32(7), 655–667.

[4]. Arthur, D. (2012). Recruiting,interviewing, selecting, orienting new employees (Cilt 5). USA: American Management Association.

[5]. Assuncao, M. D., Calheiros, R. N., Bianchi, S., Netto, M. A., & Buyya, R. (2015). Big Data computing and clouds: Trends and future directions. J. Parallel Distrib. Computing, 79-80, 3-1.

[6]. Chapelle, O., Schölkopf, B., & Zien, A. (2006). Semi-Supervised Learning. Cambridge: MIT Press.

[7]. Chaturvedi, V. (2014). An Investigation into role of competency assesment through assesment center in building employee compentency and organization effectiveness. International Journal of Research in Management & Social Science, 2(2).

[8]. Christou, E. (2015). Branding Social Media in the Travel Industry. Procedia - Social and Behavioral Sciences, 175, 607 – 614.

[9]. Chung, J. (2015). Antismoking campaign videos on YouTube and audience response: Application of social media assessment metrics. Computers in Human Behavior, 51, 114 - 121.

[10]. Chung, N., Lee, S., & Han, H. (2015). Understanding communication types on travel information sharing in social media: A transactive memory systems perspective. Telematics and Informatics, 32, 564–575.

[11]. Cook, M. C. (2005). Psychological Assessment in the Workplace: A Manager's Guide. New York: John Wiley & Sons.

[12]. Domingues, M. A., Soares, C., & Jorge, A. M. (2013). Using Statistics, Visualization And Data Mining For Monitoring The Quality Of Meta-Data In Webportals. Information Systems and e-Business Management, 11(4), 569-595.

[13]. Graham, M. W., Avery, E. J., & Park., S. (2015). The role of social media in local government crisis communications. Public Relations Review, 41, 386 -394.

[14]. Graupe, D. (2013). Principles of Artificial Neural Networks (Cilt 3). USA: World Scientific.

[15]. Han, J., & Kamber, M. (2011). Data Mining Concepts and Techniques (Cilt 3). Morgan Kaufman Publishers Academic Press.

[16]. https://www.siop.org/workplace/employment%20testing/usingoftests.aspx. (tarih yok). 10 6, 2015 tarihinde alındı

[17]. Inmon, W. (2005). Building the Data Warehouse. Wiley.

[18]. Kepes S, M. M. (2011). Big Five Validity and publication bias: Conscientiousness worse than assumed. Annual Conference of the Society for Industrial & Organizational Psychology. Illinois.

[19]. Kim, D., Jang, J., & Shin, S. J. (2008). AC 2008-1263: The Effect Of Personality Type On Team . American Society for Engineering Education, 13(1221), 1-9.

[20]. Kim, S., Koh, Y., Cha, J., & Lee, S. (2015). Effects of social media on firm value for U.S. restaurant companies. International Journal of Hospitality Management, 49, 40-46.

[21]. Li, G., Law, R., Vu, H. Q., Rong, J., & Zhao, X. (. (2015). Identifying emerging hotel preferences using Emerging Pattern Mining technique. Tourism Management, 311-321.

[22]. Link, A. R., Cawkwell, P. B., Shelley, D. R., & Sherman, S. E. (2015). An exploration of online behaviors and social media use among hookah and electronic-cigarette users. Addictive Behaviors Reports, 2, 37-40.

[23]. Luo, N., Zhang, M., & Liu, W. (2015). The effects of value co-creation practices on building harmonious brand community and achieving brand loyalty on social media in China. Computers in Human Behavior, 492-499.

[24]. Marston, W. M. (2002). Emotions of Normal People. Kegan Paul Trench Trubner And Company.

[25]. Miller, V. D. (2014). Meeting the Challenges of Human Resource Management: A Communication Perspective. New York: Routledge.

[26]. Munar, A. M., & Jacobsen, J. S. (2014). Motivations for sharing tourism experiences through social media . Tourism Management, 46-54.

[27]. Nguyen, B., Yu, X., Melewar, T., & Chen, J. (2015). Brand innovation and social media: Knowledge acquisition from social media, market orientation, and the moderating role of social media strategic capability. Industrial Marketing Management.

[28]. Ozer, I., Karpinski, A. C., & Kirschner, P. A. (2014). A cross-cultural qualitative examination of social-networking sites and academic performance. International Conference on Education & Educational Psychology , 112, s. 873–881.

[29]. Perzyk, M., & Soroczynski, A. (2010). Comparative study of decision trees and rough sets theory as knowledge extraction tools for design and control of industrial processes". World Academy of Science, Engineering and Technology, 37.

[30]. Petkovic M, J. N. (2005). Organizacija. Belgrad: Ekonomski Fakultet.

[31]. Piotrowski, C. &. (2006). Current recruitment and selection practices: A national survey of Fortune 1000 firms. North American Journal of Psychology, 8(3), 489-496.

[32]. Premchaiswadi, W., & Romsaiyud, W. (2012). Extracting WebLog of Siam University for Learning User Behavior on MapReduce. Proceedings of 4th International Conference on Intelligent and Advanced Systems (ICIAS) and A Conference of World Engineering.

[33]. Pulakos, E. D. (2005). Selection assessment methods. SHRM Foundation's Effective Practice Guidelines. Virginia: Society for Human Resource Management.

[34]. Ramakrishnan, R., & Gehrke, J. (2003). Database Management Systems. McGraw Hill.

[35]. Roghani, A. (2015). Artificial Neural Networks: Applications in Financial Forecasting. Paperback.

[36]. S.J.H.AlKahtani, Xia, J. (., Veenendaaland, B., Caulfield, C., & Hughes, M. (2015). Building a conceptual framework for determining individual differences of accessibility to tourist attractions. Tourism Management Perspectives, 28–42.

[37]. Shafer J.C., A. R. (1997). SPRINT: A Scalable Parallel Classifier for Data Mining. 22th International Conference on Very Large Databases. Mumbai, India.

[38]. SHRM. (2012). Most employers don't use personality tests. SHRM Poll.

[39]. Silahtaroğlu, G. (2013). Veri Madenciliği Kavram ve Algoritmaları. İstanbul: Papatya Yayıncılık.

[40]. Silahtaroğlu, G. (2015). Predicting Gender of Online Customer Using Artificial Neural Networks. Proceedings of International Conference on Management and Information Technology, (s. 45-50). New York.

[41]. Stanujkic, D. D. (2015). Selection Of Candidates In The Process Of Recruitment And Selection Of Personnel Based On The Swara And Aras Methods. Quaestus(7), 53.

[42]. Wang P, Z. X., & H., O. (2002). Applications of data mining to electronic commerce: A survey, Proceedings Of The Sixth China-Japan. International Conference On Industrial Management, 397-400.

[43]. Wu, H. (2013). Improving user experience with case-based reasoning systems using text mining and Web 2.0. Expert Systems with Applications, 40(2), 500–507.

[44]. Xiang, Z., & Gretzel, U. (2010). Role of social media in online travel information search. Tourism Management, 179–188.

[45]. Yates, D., & Paquette, S. (2011). Emergency knowledge management and social media technologies: A case study of the 2010 Haitian earthquake. International Journal of Information Management, 31, 6-13.

[46]. Yuhong, L., & Weihua, M. (2010). Applications of Artificial Neural Networks in Financial Economics: A Survey. 1, s. 211-214. Hangzhou: Computational Intelligence and Design (ISCID).

[47]. Zeng, B., & Gerritsen, R. (2014). What do we know about social media in tourism? A review. Tourism Management Perspectives, 10, 27-36.

[48]. Zhai, S., Xu, X., Yang, L., Zhou, M., Zhang, L., & Qiu, B. (2015). Mapping the popularity of urban restaurants using social media data. Applied Geography, 63, 113 -120.

★★★