Title: Gender Prediction From Social Media Comments With Artificial Intelligence

Authors: Özer Çelik, Ahmet Faruk Aslan

Recieved: 2019-04-30 17:42:24

Accepted: 2019-09-03 23:11:19

Article Type: Research Article
Volume: 23
Issue: 6
Month: December
Year: 2019
Pages: 1256-1264

Sakarya University Journal of Science 23(6), 1256-1264, 2019

Gender Prediction from Social Media Comments with Machine Learning

Özer ÇELİK*[1], Ahmet Faruk ASLAN[2]

Abstract

In the 21st century, which can be termed as age of artificial intelligence, machine learning (ML) techniques that can become widespread and improve themselves can be given more quality services to humanity in many fields. As a result of these ML developments, nowadays many companies use predictive models to estimate customer behavior. Also, with increasing use of social media, the companies have started to deliver their products and services to their customers via social media accounts. But every customer is not interested in all product or service. Each customer's area of interest is different. Gender is one of the main reasons for this difference. If the gender of a social media user is determined correctly, the amount of sales may be increased by offering the appropriate products or services. The main aim of our study is an estimation of genders of the commenters thanks to machine learning techniques by analyzing the comments of companies posting on Facebook. In context of the study, the genders of the commenters labelled based on commenters' name. The data set is divided into training and test data as 70-30%. As a result of the study, it was seen that machine learning methods predicted with similar accuracy rates, while the highest accuracy rate (74.13%) was obtained by logistic regression method.

Keywords: gender prediction, artificial ıntelligence, machine learning, natural language processing, sentiment analysis

1. INTRODUCTION

The use of social media has made it easier to investigate psychological and social problems [1]. Thus, it is allowed more data-based study beside hypothesis-testing of social science process [2]. With the help of social media, it could be followed up psychological well-being [3, 4, 5], and a host of other behavioral, psychological, medical phenomena [6] and disease rates [7]. Unlike classic hypothesis based on social science, such wide-scale social media researches rarely pay attention to or have access to age and gender information, which can have a significant impact

* Corresponding Author: ozer@ogu.edu.tr

[1] Eskisehir Osmangazi University, Department of Mathematics and Computer Science, Eskisehir, Turkey. ORCID: 0000-0002-4409-3101

[2] Eskisehir Osmangazi University, Department of Mathematics and Computer Science, Eskisehir, Turkey. ORCID: 0000-0003-1583-6508

over many problems. For example, males live nearly five years shorter than females [8]. Men and women differ usually significantly in their interests and work choice [9]. Moreover, social media language change by age [10, 11] and gender [12]. A male could have bias on twitter [13], while social media generally skew towards being young and female.

Today, many companies aim to deliver their services and products to their customers by social media accounts. But a customer is not interested in all product or service. Each customer's product type of interest is different. Gender is one of the main reasons for this difference. If the gender of a social media user is determined correctly, the amount of sales may be increased by offering the appropriate products or services.

Companies measure the satisfaction of their customers upon the physical or online product/service sales from social media. It is quite difficult to analyze the comments for the sharings during a whole day. It can easily be studied context emotion analysis (positive, negative, neutral) in natural language processing which is a subfield of artificial intelligence. In order to consider their services, the companies need to analysis according to the different information of their customers in addition to these analyses. One of the main factors in providing products according to the customers' information is gender.

Online behavior is representative of many aspects of a user's demographics [14, 15]. Many studies have used linguistic cues (such as ngrams) to determine if someone belongs to a certain age group, be it on Twitter or another social media platform [16, 17, 18, 19]. Gender prediction has been studied across blogs [20, 21], Yahoo! search queries [22], and Twitter [15, 18, 20, 23]. Because Twitter does not make gender or age available, such work infers gender and age by leveraging profile information, such as gender-discriminating names or crawling for links to publicly available data (e.g. [20]).

2. MACHINE LEARNING

Machine learning is the algorithm and statistical models used to perform a specific task based on patterns and inferences, instead of using an open instruction of computer systems. It is a sub-topic of artificial intelligence. Machine learning algorithms establish a statistical model of sample data, named as training data, owing to make predictions or decisions without being distinctly programmed [24]. Machine learning is utilized in many applications, such as email filtering and computer vision, where it is not possible to develop appropriate instructions. Machine learning is science of computational statistics, which based on making predictions by using computers. Machine learning focuses on estimations from the learned data based on known features.

Data mining is a field of study within machine learning and focuses on exploratory data analysis through unsupervised learning [25]. Data mining focuses on discovering unknown (historical) features in the data. This is a step in the analysis of information discovery in databases.

In this study, Artificial Neural Networks (ANN), Decision Tree (DT), Support Vector Machine (SVM), Naive Bayes (NB), Logistic Regression (LR), k-Nearest Neighbor (KNN) and Extreme Gradient Boosting (XGBoost) by using machine learning techniques.

2.1. Artificial Neural Networks

The full manuscript must not exceed 20 pages. It must include an abstract of up to 300 words. The 20 pages should include all tables, figures, and references.

2.2. Decision Tree

DT is the decision structure that performs learning from known data classes by inductive method. Decision tree is a learning algorithm that allocates large amounts of data into small data groups utilizing simple decision-making steps. As a result of each accomplished separation, the members in the result group are more like each other. Decision tree with descriptive and predictive features is one of the most preferred classification algorithms due to its reliable, easy to interpret and integratable into databases [27].

2.3. Support Vector Machine

The support vector machine, also known as support vector network, is one of the supervised classification techniques laid down by Cortes and Rapnik (1995) [28]. SVM is the machine learning algorithm which performs estimate and generalization about new data by performing learning on data that unknown the distribution. The fundamental principle of the SVM is based on the asset of a hyperplane that best distinguishes the data of two classes. SVM is divided into two according to the linear separation and nonlinear separation of the data set [29].

2.4. Naive Bayes

The Naive Bayes classification is a classification by utilizing statistical methods for labeling data. Because it is easy to use, it is often preferred in classification problems. It is generally aimed to calculate the probability values of the effects of each criterion in the Bayesian classification. Naive Bayes calculates the conditional probability of the class to which the data belongs, in order to predict the probability of a class with a data. Bayes theorem is used in this process.

2.5. Logistic Regression

Logistic regression is a method of classifying the relationship between multiple independent variables and dependent variables. Although it has usually been used in medical field in the past, it is an advanced regression method which has gained popularity in social sciences today. Logistic regression is a technique used as an alternative to this method due to the inadequacy of Least Squares Method (LSM) in a multivariate model with dependent and independent variable discrimination. In logistic regression analysis, the probability of the dependent variable with two values is predicted. In addition, the variables in the model are continuous. Because of this feature, it is a technique frequently used for classifying observations.

2.6. k-Nearest Neighbor

The k-nearest neighbor algorithm, which submitted by Fix and Hodges in 1951, are based on the logic that the data closest to each other belong to the same class. The main purpose is to classify the new incoming data by using the data previously classified. The data, which is unknown to which class it belongs to, are called test samples, the previously classified data are called learning samples. In the KNN algorithm, the distance of the test sample from the learning samples is calculated, and then the k-learning sample closest to the test sample is selected. If the selected k samples have mostly belonged to which class; the class of the test sample is also determined as this class [30].

2.7. Extreme Gradient Boosting

XGBoost is a scalable, portable and computationally compiled package of gradient tree strengthening algorithm. While the gradient tree algorithm tries to solve the optimization problem in two basic steps (first determines the direction of the step, then determines the step size), XGBoost finds the step size and direction at one time. Additionally, the XGBoost name refers to the engineering target that pushes the boundary of computing resources for increased tree algorithms. Many researchers use XGBoost due to the reason. The algorithm was designed for performance of compute time and memory resources in the implementation. The design goal is to make the best use of current resources owing to train the model [31].

3. DATA ANALYSIS

3.1. Data Set

Ready package implementations and strong programs in data science are used in machine learning process. The one used in Waikato Environment for Knowledge Analysis (WEKA) machine learning is the most popular open source coded program.The data sets are in Attribute-Relation File Format (arff) or Comma Separated Values (csv) format for WEKA . So, there is no need to deal with any programming language. The results are gained at training phase by selecting the ready machine learning methods. [32].

Besides, Matlab, Python and R programming languages can be used at machine learning. In our study Python programming language and Scikit Learn library were used. The CountVectorizer method in Scikit Learn library were used in pre-data processing phase. In order to operate Jupyter Notebook and Python and R languages provided by Microsoft for free on cloud Azure Notebook platform were used.

In Python, Scikit Learn library is used at classification in machine learning, clustering and estimation [43]. A total of 8770 comments collected by us have been investigated (1533 female and 7237 male). 1533 female, 1533 male, balanced data set was prepared by stratified sampling. Male and female were coded as 0 and 1, respectively. In addition, the name variable taken part in the data set is also included in the model.

Table 1. Sample Comments in The Data Set

Name	Comment	Gender	Emotion
Furkan Y.	laptop sogutucusu var	0	Neutral
Ziya Y.	ben burda musteriye sunulan avantaji anlayamadim bankadan faizini vererek krediyi cekiyorsun zaten canli para kredi karti pesin fiyati taksitli fiyati olsun	0	Uncertain
Evren G.	hastasiyim telefonun yarin alacagim nasipse	0	Positive
Ferhat U.	ne alacam dandik alana gerizekali derim israf verilen para	0	Negative
Zuhal O.	karlar ulkesi elbisenin fiyati nedir	1	Neutral
Hicran K.	benimkide remington	1	Uncertain
Zeynep K.	benim evimde her sey vestel memnunum	1	Positive
Esin Y.	kesinlikle tavsiye etmiyorum ttelekom olduktan sonra nede telefon hicbisey cekmiyor yer bursa apzima bile almak istemiyorum sozlesmem bitsin bayilerinin onunden dahi gecmem	1	Negative

3.2. Data Processing

In this study, Facebook were selected as a social media platform. The comments of some brands were specified on Facebook. The sharings of the pages, comments, the message text, message transmission date, comments and commenters' information on Facebook can be reached by using several Facebook Application Programming Interfaces (APIs). In order to archive the datas gained via API, SQLite database could be used [17]. By using Ruby programming language, we took the datas via Facebook API and archived on SQLite database.

Then, the arrangement and labelling of the data set before training were done. The comments in data set were cleared from meaningless words and punctuations marks via several pre operations. In order to realize the machine learning, the labelling was done according to the names of users that commented and the gender marks used in Turkish.

A model was created with the data set in our study, Python programming language of Scikit Learn library, KNN and SVM classification algorithms. With the test data reserved in %50 rate, the accuracy rates were calculated on the created model.

The Accuracy Rate (ACC), a commonly used success evaluation method, was used in our study. The accuracy method is the rate of the sample number the system classifies as trues (True Positive (TP) and True Negative (TN)) to all sample number. And the error rate is the rate of the sample number calculated false (False Positive (FP) and False Negative (FN)) to all sample number. It is expected to have the accuracy rate is higher than the false rate at the end of the study.

Success scores are calculated with the help of the confusion matrix (Table 2).

Table 2. Confusion Matrix

		Actual		**Total (%)**
		0	1	
Predicted	0	TP	FP	Precision Score
	1	FN	TN	Negative Predictive Value (NPV)
Total		Recall Score, Sensitivity	Specificity	ACC

The success measures and formulas used in our study, which were calculated with the help of Confusion Matrix;

$$ACC = (TP + TN)/(TP + TN + FP + FN)$$

$$Precision = TP/(TP + FP)$$

$$NPV = TN/(TP + FP)$$

$$Recall = TP/(TP + FN)$$

$$Specificity = TN/(FP + TN) \qquad (1)$$

There are several more accuracy scores calculated with the help of confusion matrix. In addition to, power of the study, type II error, type I error are calculated respectively via TP value, FN value and FP value.

All analysis and processing A computer with Windows 10 64-bit operating system, quad-core Intel Skylake Core i5-6500 CPU with 3.2 GHz 6MB Cache and 8GB 2400MHz DDR4 Ram memory was used.

4. CONCLUSION

The data set is divided into training and test data as 70-30%. It was observed accuracy rates in Table 3 through the results of the study. According to Table 3, the highest accuracy rate was achieved by logistic regression algorithm (74.13%). Approximately 70% accuracy rate was obtained with other algorithms. The comments of the data set were used directly in the training without morphological analysis. It is assumed that if the morphological analysis of these comments is done and then used in the training, the accuracy rates will be higher.

One of the biggest constraints of the analysis developed in order to make predictions on the text is that grammar rules are not frequently observed. This leads to incorrect estimates and therefore low accuracy rates. In similar gender prediction studies in different languages, success rates of approximately 70-80% have been achieved. It is also possible to achieve higher accuracy rates with more data sets.

Table 3. The Accuracy Rates of The Algorithms

Algorithm	Accuracy Rate (%)	Rate of 0 - 1 (%)
ANN	70.65	77.0-65.0
DT	69.93	79.0-57.0
SVM	72.83	79.0-67.0
NB	70.76	83.0-58.0
LR	74.13	78.8-69.3
KNN	63.04	49.0-77.0
XGBoost	67.39	86.0-49.0

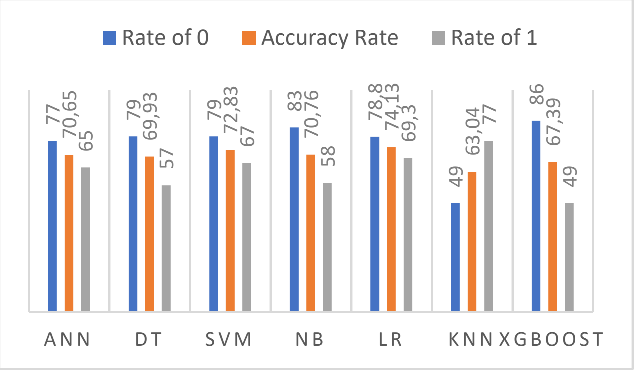

Figure 1. The Accuracy and The Category Rates of The Algorithms

As a result of the research, Receiver Operator Characteristics (ROC) curve graph calculated by confusion matrix data is given in Figure 1. Owing to ROC curve graph, it is possible to see graphically the performances of the models created. It is thought that the reason why the logistic regression algorithm predicts it with higher accuracy rate may be that gender is in binary format.

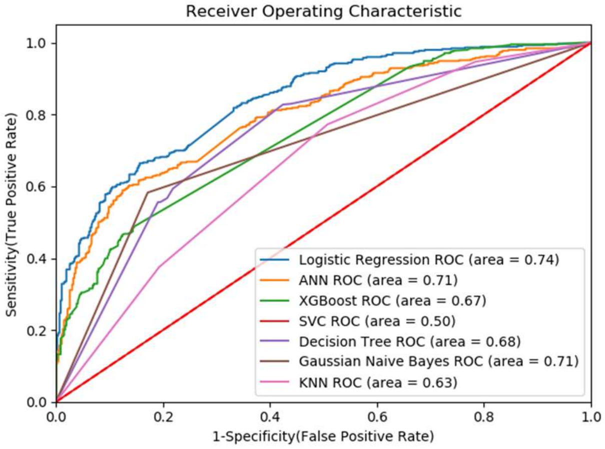

Figure 2. ROC Curve Graph

The results of the logistic regression which gives the highest success rate are given in Table 4.

Table 4. Confusion Matrix of Logistic Regression

		Actual		Total (%)
		Male	Female	
Predicted	Male	368	99	78.80
	Female	139	314	69.32

Total	72.58	76.03	74.13

5. DISCUSSION

For a long time, researchers have investigated for a better understanding of human psychology by examining words people use [34, 35, 36]. According to Tauszczik & Pennebaker say it: Language is the most collective and reliable way for human to convert their inner thoughts and emotions into a form that others can understand. So, language and words are the very items of psychology and communication [37].

There are several studies on the emotion analysis of the sharings on social media and the comments to these sharings. Pang, Lee and Vaithyanatham realized emotional analysis study using SVM and NB machine learning algorithms on English sentences. In this study SVM (%82,9) algorithms gets the highest rate [38]. Similar studies were made for Turkish texts as well. Cetin and Amasyali realized emotion analysis study for Turkish Twitter datas in 2013 using NB, Random Forest, Sequential Minimum Optimizasyon (SMO), KNN and Instance-Based (IB1) learning algorithms [39]. Apart from the comments on social media sites it is also made the emotion analysis of the comments to the cinema, meal order, hotel etc. sites. Sevindi used the machine algorithms of DT, KNN, NB and SVM in 2013 in his Turkish cinema comments emotion analysis studies and had successful results [40]. Nizam and Sakın acquired the best social media emotion analysis performance with %72,33 accuracy rate from SMO using NB, RF, SMO, DT (J48) and IB1- the controlled machine learning methods [41].

Sap et. al. said that demographic lexica have used for widespread in social science, economic, and business applications. A lexica (words and weights) was predicted from words with the demographic tags in Twitter, Facebook, and blog data for age and gender by utilizing classification and regression methods. It was determined the lexica publicly available, was effective technique in language-based age and gender prediction over Facebook and Twitter. Then the lexica were evaluated for generalization across social media genres as well as in limited message situations [42].

As a result of the literature review, it was seen that at least 72% accuracy was achieved with different ML algorithms in the studies of gender estimation from the comments made on different social media platforms. In our study, the correct prediction rate of 74.13% was achieved with Logistic Regression algorithm. In order to achieve a higher success rate in such studies, it is necessary to have fewer errors such as spelling and grammatical errors in interpretations as suggested in other studies.

6. REFERENCES

[1] D. Lazer, D. Brewer, N. Christakis, J. Fowler, and G. King, "Life in the network: the coming age of computational social science." Science (New York, NY), 323(5915), 721, 2009.

[2] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, R. E. Lucas, M. Agrawal, and L. H. Ungar. "Characterizing Geographic Variation in Well-Being Using Tweets." In ICWSM (pp. 583-591), 2013.

[3] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth. "Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter." PloS one, 6(12), e26752, 2011.

[4] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz. "Predicting depression via social media." ICWSM, 13, 1-10, 2013.

[5] H. A. Schwartz, , J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, and L. H. Ungar. "Personality, gender, and age in the language of social media: The open-vocabulary approach." PloS one, 8(9), e73791, 2013.

[6] M. Kosinski, D. Stillwell, and T. Graepel. "Private traits and attributes are predictable from digital records of human behavior." Proceedings of the National Academy of Sciences, 201218772, 2013.

[7] M. J. Paul, and M. Dredze. "You are what you Tweet: Analyzing Twitter for public health." Icwsm, 20, 265-272, 2011.

[8] A. Marengoni, S. Angleman, R. Melis, F. Mangialasche, A. Karp, A. Garmen, and L. Fratiglioni. "Aging with multimorbidity: a systematic review of the literature." Ageing research reviews, 10(4), 430-439, 2011.

[9] R. R. McCrae, and P. T. Costa Jr. "A five-factor theory of personality." Handbook of personality: Theory and research, 2(1999), 139-153. 1999.

[10] M. L. Kern, J. C. Eichstaedt, H. A. Schwartz, G. Park, L. H. Ungar, D. J. Stillwell, and M. E. Seligman. "From "Sooo excited!!!" to "So proud": Using language to study development." Developmental psychology, 50(1), 178, 2014.

[11] J. W. Pennebaker, and L. D. Stone. "Words of wisdom: Language use over the life span." Journal of personality and social psychology, 85(2), 291, 2003.

[12] D. A. Huffaker, and S. L. Calvert. "Gender, identity, and language use in teenage blogs." Journal of computer-mediated communication, 10(2), JCMC10211, 2005.

[13] A. Mislove, S. Lehmann, Y. Y. Ahn, J. P. Onnela, and J. N. Rosenquist. "Understanding the Demographics of Twitter Users." ICWSM, 11(5th), 25, 2011.

[14] M. Pennacchiotti, and A. M. Popescu. "A Machine Learning Approach to Twitter User Classification." Icwsm, 11(1), 281-288, 2011.

[15] Rao, D., Yarowsky, D., Shreevats, A., and Gupta, M. (2010, October). Classifying latent user attributes in twitter. In Proceedings of the 2nd international workshop on Search and mining user-generated contents (pp. 37-44). ACM.

[16] F. Al Zamal, W. Liu, and D. Ruths. "Homophily and Latent Attribute Inference:

Inferring Latent Attributes of Twitter Users from Neighbors." ICWSM, 270, 2012.

[17] A. Shlomo K. Moshe, W. P. James, and S. Jonathan. "Automatically profiling the author of an anonymous text." Communications of the ACM, 52(2):119–123, 2009.

[18] D. Nguyen, R. Gravel, D. Trieschnigg, and T. Meder. "" How Old Do You Think I Am?" A Study of Language and Age in Twitter." In ICWSM, 2013.

[19] F. Rangel, and P. Rosso. "Use of language and author profiling: Identification of gender and age." Natural Language Processing and Cognitive Science, 177, 2013.

[20] J. D. Burger, and J. C. Henderson. "An Exploration of Observable Features Related to Blogger Age." In AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs (pp. 15-20), 2006.

[21] S. Goswami, S. Sarkar, and M. Rustagi. "Stylometric analysis of bloggers' age and gender." In Third International AAAI Conference on Weblogs and Social Media, 2009.

[22] R. Jones, R. Kumar, B. Pang, and A. Tomkins. "I know what you did last summer: query logs and user privacy." In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (pp. 909-914). ACM, 2007.

[23] W. Liu, and D. Ruths. "What's in a Name? Using First Names as Features for Gender Inference in Twitter." In AAAI spring symposium: Analyzing microtext (Vol. 13, No. 1, pp. 10-16), 2013.

[24] M. A. Keane. "Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming." Artificial Intelligence in Design '96. Springer, Dordrecht. pp. 151–170, 1996.

[25] J. H. Friedman. "Data Mining and Statistics: What's the connection?" Computing Science and Statistics. 29 (1): 3–9, 1998.

[26] M. Gerven, and S. Bohte. "Artificial neural networks as models of neural information processing." Frontiers Media SA, 2018.

[27] A. S. Albayrak, and O. G. S. K. Yilmaz. "Veri madenciliği: Karar ağacı algoritmaları ve İMKB verileri üzerine bir uygulama." Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 14(1), 2009.

[28] O. Celik, and S. S. Altunaydin. "A Research on Machine Learning Methods and Its Applications." Online Learning, 1(3), 2018.

[29] H. Guneren. "Destek vektör makineleri kullanarak gömülü sistem üzerinde yüz tanıma uygulaması", 2015.

[30] H. Ozkan. "K-Means Kümeleme ve K-NN Sınıflandırma Algoritmalarının Öğrenci Notları ve Hastalık Verilerine Uygulanması Bitirme Tezi", İstanbul Teknik Üniversitesi, İstanbul, 2013.

[31] J. Brownlee. "A Gentle Introduction to XGBoost for Applied Machine Learning. Machine Learning Mastery." Available online: http://machinelearningmastery.com/gentle-introduction-xgboost-appliedmachine-learning/ (accessed on 2 March 2018), 2016.

[32] https://www.cs.waikato.ac.nz/ml/weka/, (Access Date: 01.02.2018).

[33] http://scikit-learn.org/, (Access Date: 01.02.2018).

[34] P. Stone, D. Dunphy, M. Smith. "The General Inquirer: A Computer Approach to Content Analysis." MIT press, 1966.

[35] M. Coltheart. "The mrc psycholinguistic database." The Quarterly Journal of Experimental Psychology 33: 497–505, 1981.

[36] J. W. Pennebaker, M. R. Mehl, K. G. Niederhoffer. "Psychological aspects of natural language use: our words, our selves." Annual Review of Psychology 54: 547–77, 2003.

[37] Y. Tausczik, J. Pennebaker. "The psychological meaning of words: Liwc and computerized text analysis methods." Journal of Language and Social Psychology 29: 24–54, 2010.

[38] B. Pang, L. Lee, and S. Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics, 2002.

[39] M. Cetin, and M. F. Amasyali. "Supervised and traditional term weighting methods for sentiment analysis." In Signal Processing and Communications Applications Conference (SIU), 2013 21st (pp. 1-4). IEEE, 2013.

[40] B. I. Sevindi. "Comparison of supervised and dictionary based sentiment analysis approaches on Turkish text" (Doctoral dissertation, Master thesis, Gazi University, Turkey), 2013.

[41] H. Nizam, and S. S. Akin. "Machine Learning in Social Media and the Comparison of the Balanced and Non-balanced Data Sets in Emotion Analysis." XIX. Internet Conference in Turkey, 2014.

[42] M. Sap, G. Park, J. Eichstaedt, M. Kern, D. Stillwell, M. Kosinski, and H. A. Schwartz. "Developing age and gender predictive lexica over social media." In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1146-1151), 2014.