Alexandria Engineering Journal (2022) 61, 8007–8025

Alexandria University

Alexandria Engineering Journal

www.elsevier.com/locate/aej
www.sciencedirect.com

Predicting personality traits with semantic structures and LSTM-based neural networks

Muhammed Ali Kosan [a,b,*], **Hacer Karacan** [c], **Burcu A. Urgen** [d,e,f]

[a] *Department of Software Engineering, Mus Alparslan University, 49250 Mus, Turkey*
[b] *Department of Computer Science, Gazi University, 06680 Ankara, Turkey*
[c] *Department of Computer Engineering, Faculty of Engineering, Gazi University, 06570 Ankara, Turkey*
[d] *Department of Psychology, Bilkent University, 06800 Bilkent, Ankara, Turkey*
[e] *Interdisciplinary Neuroscience Program, Bilkent University, 06800 Bilkent, Ankara, Turkey*
[f] *Aysel Sabuncu Brain Research Center and National Magnetic Resonance Research Center (UMRAM), Ankara, Turkey*

Received 7 October 2021; revised 15 January 2022; accepted 17 January 2022
Available online 31 January 2022

KEYWORDS

Personality traits;
Prediction;
LSTM;
FastText;
Preprocessing;
Personality dataset

Abstract There is a need to obtain more information about target audiences in many areas such as law enforcement agencies, institutions, human resources, and advertising agencies. In this context, in addition to the information provided by individuals, their personal characteristics are also important. In particular, the predictability of personality traits of individuals is seen as a major parameter in making decisions about individuals. Textual and media data in social media, where people produce the most data, can provide clues about people's personal lives, characteristics, and personalities. Each social media environment may contain different assets and structures. Therefore, it is important to make a structural analysis according to the social media platform. There is also a need for a labelled dataset to develop a model that can predict personality traits from social media data. In this study, first, a personality dataset was created which was retrieved from Twitter and labelled with IBM Personality Insight. Then the unstructured data were transformed into meaningful and processable data, LSTM-based prediction models were created with the structural analysis, and evaluations were made on both our dataset and PAN-2015-EN.

1. Introduction

Intelligence, by definition, is the processing of newly collected or learned data to serve any purpose and the presentation of the knowledge obtained [1–3]. Although it has generally been used for military and political purposes, today it is used in many different areas such as advertising agencies, companies, states, organizations, and even by individuals. For intelligence, there must be a source from which information can be

* Corresponding author at: Department of Computer Science, Gazi University, 06680 Ankara, Turkey.
E-mail addresses: makosan@gazi.edu.tr (M.A. Kosan), hkaracan@gazi.edu.tr (H. Karacan), burcu.urgen@bilkent.edu.tr (B.A. Urgen).
Peer review under responsibility of Faculty of Engineering, Alexandria University.

https://doi.org/10.1016/j.aej.2022.01.050
1110-0168

obtained. The development of technology and the increase in data produced by the spread of the Internet has brought a new dimension to intelligence studies. With these ever-increasing data, it is thought to be less risky to transform them into meaningful information to serve a specific purpose. This process also produces results in a shorter time than other intelligence activities where access to information can be risky and take long periods of time. Given that there is a chance that no information during any intelligence activity, technology-based intelligence activities (especially internet-based intelligence) are less risky. Even when no information is extracted from the data obtained over the Internet, the information is more accessible because the search can be quickly repeated.

The Internet benefits us with size and variety of information. The number of Internet users in the world exceeds 4.9 billion, and the number of websites is approaching 2 billion. Internet traffic generated in the first half of 2021 is around 1.8 Zettabytes, and 100 Terabytes of traffic is generated per second on average. The production rates in social media contents, along with blog contents where personal or corporate information is shared, are presented in Table 1. Since this information includes personal and institutional data, it better reveals the dimensions of accessible information. [4]

In addition, the number of active users of some social media platforms obtained in the first half of 2021 is presented in Table 2.

Based on the amount of content produced and the number of active users, the amount of information generated will continue to increase exponentially. Considering the information dimension created by social media accounts today, it is not possible to examine this information through basic reading and evaluation. In addition, considering the value of the information generated, the use of classical methods is not the right choice. The fact that the information produced on social media platforms is mostly accessible, especially when compared to e-mail, and contains personal information enables us to consider social media platforms more. Additionally, people are using the internet more due to Covid-19 increases the rate of information production on social media platforms.

Although the usage of personal information as intelligence information is possible in all fields, we can say that the most difficult to calculate and the most important distinguishing feature is the personality of the users. At this point, the term user personality seems more appropriate because it is accepted that not every social media user is producing content consistent with his/her actual personality. As a result, shares should be observed as if they are a character produced in a persona manner. Therefore, the prediction of personality from social media platforms can only provide information about user personality. While, this estimation cannot be as precise as the information obtained in a clinical setting, predictability studies can be done on a specific sample. Extraction and prediction of personality traits from social media content can be used in online marketing, recruitment processes, recommendation systems, advertising agencies, crime, and intelligence profiling, etc. [5].

Table 1 General information statistics generated on the Internet [4].

Field	Average for the first half of 2021	Per second (Average)
Email Sending	49.4 Trillion	3 Million
Google Search	1.45 Trillion	92.6 Thousand
Blog Post	1.40 Billion	–
Twitter Post (Tweet)	149 Billion	9.5 Thousand
Watching Youtube Videos	1.41 Trillion	90 Thousand
Instagram Image Post	16.9 Billion	1.07 Thousand
Tumblr Post	29.8 Billion	1.9 Thousand

Table 2 Number of active users of some social media platforms [4].

Social media platform	Number of active users
Facebook	2.84 Billion
Twitter	375 Million
Pinterest	372 Million

In this study, the dataset created from a specific sample taken from social media platforms via Twitter was labeled based on the Big Five Model using IBM Personality Insight. Next, a structural analysis of tweet data was presented, and the important processing steps in the text preprocessing stage were given in a multi-selection manner. Then, Deep Learning training and test analysis were performed using different preprocessing and LSTM-based artificial neural network model combinations. Finally, a comparative analysis was made on the Twitter data with structural analysis, preprocessing model and deep learning model, thus it is aimed to contribute to the literature (for both academia and the public/private sector working for this purpose).

The innovation of this study is presented below:

- Evaluation of the preprocessing steps presented in the structural analysis of textual data
- Comprehensive and categorical analysis of the literature in predicting personality traits
- The importance of structural analysis of Twitter data in predicting personality traits
- Personality traits dataset based on Twitter contents suitable for generalization
- Balanced Bi-LSTM model for personality traits prediction with structural preprocessing steps

2. Literature review

Since the majority of the studies have been conducted in the last 5 years, combined with the desire to create the scope of the study from current studies, the literature review was limited to studies conducted in the last 5 years. The majority of the studies examined were studies on academic and student performance analysis, internet, technology, social media, smartphone addiction, and social science studies based on personality scale survey results, etc. in terms of scope. These studies are not related to prediction personality traits from social media platforms. For this reason, the studies that were examined and determined to be out of scope were excluded from the literature pool. In the remaining studies, topics such as alcohol use, cyberbullying, plant/animal personality, eating habits,

trolling, addiction were covered by questionnaires made using social media. Finally, studies that were not related to the scope of the study were deleted from the literature pool by examining these parameters.

In studies conducted after 2017, social media platforms such as Facebook, Twitter, Instagram, Sina Weibo, Youtube, and Google+ were used in personality prediction. The majority of the work done on the Facebook data uses the myPersonality dataset [6], which was actively distributed a while ago, and currently, only 250 post data are shared publicly. The datasets used in the studies are mostly written in English. The Big Five Inventory (BFI) is the most widely used personality scale due to its widespread acceptance. The data preprocessing methods used in each study vary. However, classification, regression and clustering methods used in the field of artificial intelligence are used in addition to the basic statistical methods for prediction models.

Considering the content and scope, a detailed presentation of the studies obtained in 2017 and later is given in Table 3, in the order of time (the most recent is at the bottom). In the preprocess column in Table 3, the following list is used for easy control of the table:

a. Deleting punctuation marks
b. Deleting numbers
c. Transforming text to lowercase
d. Clearing stop words
e. Deleting excess spaces
f. Root reduction (stemming, lemmatization)
g. Word count
h. LIWC features
i. Content categorization
j. Extracting information from content
k. Feature extraction from image
l. Face detection from picture
m. Filtering non-text objects
n. Feature extraction from text
o. Feature selection
p. Conversion of different languages
q. Deleting special characters
r. Word spelling correction
s. Vectorizing
t. Converting non-text objects to meaningful text
u. PoS tagging (Part-of-speech tagging)
v. Tokenization
w. Word type inference (verb, object, etc.)
x. N-grams
y. Converting emoji and emoticon to meaningful text
z. Positive-negative tagging

The personality tests that were covered in a *meta*-analysis in 2017 [44] are Big-Five Inventory-10, Big Five Inventory, and IPIP. Facebook is mostly used as a social media platform, and the platforms Twitter, Sina Weibo, and Instagram follow Facebook in that order. Language-based content, activities, and images are used as the basis content. All of the studies presented for *meta*-analysis used the myPersonality dataset [6].

The PAN offered by the Webis group [45] was presented as a shared task series on digital text analysis. In 2015, all personality trait prediction studies with the shared task were evaluated and published in a single study [46]. The datasets were presented separately in English, Dutch, Italian, and Spanish with tweets, age, and gender information collected from Twitter. As the training dataset, 152 English, 34 Dutch, 38 Italian, and 100 Spanish user information was given in the study. In the test dataset, 142 English, 32 Dutch, 36 Italian, and 88 Spanish user information was given. The tagged data of the users in these datasets and the Twitter data were presented to the participants in encrypted form. The tweets of each user in the datasets contained only the text of the tweet. In addition, the labeled data of the users are the values belonging to the big-five personality traits (between -0.5 and $+0.5$) obtained by the BFI-10 test, gender, and age. Accuracy for gender and age, RMSE for big-five values were used as performance criteria for these tags. Each group participating in the study applied different methods, and the success rates varied for each output. In the personality trait assessment, the mean best RMSE value was 0.1442 for English datasets. This RMSE value was determined by the two groups, followed by 0.1235 for the Spanish datasets, 0.1044 for the Italian datasets, and finally 0.0563 for the Dutch data. In the studies conducted in 4 languages, the best performance in 3 languages other than Dutch was achieved [47]. The proposed method consists of a combination of Second-Order Attributes (SOA) and Latent Semantic Analysis (LSA). The method proposed by the authors was compared to the Bag-of-words (BOW), SOA, and LSA methods, and the proposed method achieved higher success than these other methods.

As an example of picture-based studies in the literature, Jia Xu et al. [48] presented a personality trait estimation method using the S-NNPP (Soft Threshold-Based Neural Network for Personality Prediction) method over 2D and 2.5D profile pictures. Traditional BP, MobileNetV2, and ResNeSt50 methods were used as comparison methods. They used TPR, FPR, F-measure and ROC Curve methods as the evaluation criteria. In the study, the best performance was obtained with 2.5D profile pictures using the S-NNPP method.

Based on all the studies examined, it was determined that the structural analysis made for each social media environment, and the training and test datasets created with preprocessing were mostly not suitable for other social media environments. At this point, the necessity of most general structural transformations was determined by analyzing the structure of each social media environment within itself and determining all structurally meaningful/meaningless entities. In addition, the effect of diversity in the data preprocessing on differences in model training was not fully emphasized, and in this context, the lack of comparative analysis was identified. The aim of this study was to present a new dataset, a new system, and a comprehensive analysis on user personality detection from Twitter data.

3. Materials and methods

This section explains the proposed preprocessing and prediction model along with the dataset to be used to create the personality traits prediction model from tweet data. The entire workflow of the research is presented in general terms in the graphical abstract presented in Fig. 1. First, the personality traits were labeled with the IBM Personality Insight service, along with the tweet data collected after the scope was determined. Afterwards, a cycle is presented on preprocessing, LSTM-based neural network model, and the evaluation of

Table 3 Literature studies that predict personality traits with social media content.

Study	Dataset Properties	Model	Result
N. Ahmad and J. Siddique [7]	**SM:** Twitter **PI:** DISC **LN:** English **CT:** Tweet **DT:** Collected tweets based on keywords (millions of tweets)	**PP:** a, b, c, d, e, f **MT:** By calculating word frequency and word cloud weight, the personality traits were analyzed to find the most dominant words. **VTm:** Calculated at least 20 times for each scale.	D: get, new I: feelings, social S: help, open C: clear, details
Z. Ahmad et al. [8]	**SM:** Facebook **PI:** Extraversion (E), neuroticism (N), and shyness (S) **LN:** Malay and English **CT:** Post **DT:** 170 users, 12,573 data	**PP:** g, h **MT:** Feature selection was made using Pearson's Correlation and then ZeroR, Hoeffding tree, Naive Bayes, Logistic MultiClass, J48, and OneR methods were applied on WEKA. **VTm:** 10-fold cross-validation, accuracy, f-measure	E: Hoeffding tree, acc 67.81, f-measure 0.66 N: NaiveBayes, acc 58.26 S: J48, acc 70
N. Alsadhan and D.B. Skillicorn [9]	**SM:** Facebook, Twitter, Youtube, and Articles **PI:** NEO-PI-R, Big-Five traits, Myers-Briggs personality types **LN:** English, Spanish, German, Dutch, Italian, Portuguese, French, and Arabic **CT:** Post, Tweet, Vlog, Article **DT:** 9918 Facebook posts, 404 vlogs, 7818 Twitter users	**PP:** g **MT:** A proposed method with SVD (Singular Value Decomposition) **VTm:** Monte Carlo cross-validation, accuracy, f-measure, best choice of ratio	Facebook: f1 0.82, Youtube: f1 0.81, Articles: f1 0.73, Twitter: f1 (English: 0.9, German: 0.76, Italian: 0.9, Spanish: 0.72, French: 0.86, Portuguese: 0.85, Dutch: 0.85, Arabic: 0.76)
S.K. Bhatti et al. [10]	**SM:** Twitter **PI:** Big-Five traits **LN:** English **CT:** Tweet, Profile Picture **DT:** 54,784 Twitter users	**PP:** i, j, k, l **MT:** Using Face++ and EmoVu, information was extracted from the faces in the pictures. In addition, information about the user such as age was estimated from the text-based tweet data and analyzed using Pearson's Correlation with the information obtained from the images. It has been suggested that personality estimation can be made from the profile picture by using the Elastic Net Regularization method with Linear Regression. **VTm:** Pearson's Correlation	r > :135
S.C. Guntuku et al. [11]	**SM:** Twitter **PI:** NEO-PI-R **LN:** - **CT:** Liked and posted pictures **DT:** 1.5 million images from 4000 users	**PP:** k, l **MT:** The effects of features extracted by Pearson's correlation were calculated. It has been suggested that personality estimation can be made from pictures by using Linear Regression and Elastic Net Regularization methods. **VTm:** The features extracted by VGG-Net and Imagga were compared with the proposed feature extraction. RMSE for accuracy and Pearson's correlation for performance. 10-fold cross-validation.	Conscientiousness, agreeableness, and neuroticism presented better results than image features, while openness and extraversion were better predictable from text content.
S. Huang et al. [12]	**SM:** Sina Weibo **PI:** Big Five -Berkeley 44 items **LN:** Chinese **CT:** Text **DT:** 994 users	**PP:** m, n **MT:** Multi-task learning and Robust Multi-task learning methods have been proposed. A comparison of the proposed methods was made with Naive Bayes, Logistic Regression, RepTree, and Random Forest. **VTm:** Precision, recall, f-measure	According to the F-measure criterion, RMTL gave better results than all the tested methods.

Table 3 (*continued*)

Study	Dataset Properties	Model	Result
L. Asadzadeh and S. Rahimi [13]	**SM:** Facebook **PI:** Big-Five **LN:** - **CT:** Like **DT:** 92,255 users from myPersonality database	**PP:** - **MT:** LASSO algorithm **VTm:** MSE, Pearson's Correlation	MSE (O:0.024, C:0.030, E:0.038, A:0.030, N:0.038), Pearson's Correlation (O:0.38, C:0.29, E:0.34, A:0.22, N:0.27)
C. Li et al. [14]	**SM:** Sina Weibo **PI:** Big-Five **LN:** Chinese **CT:** Text **DT:** -	**PP:** n, o **MT:** PCA and Correlation analysis were used for feature selection from the information provided by the user. Then Multiple Regression Model, Gray Prediction Model, and Multi-tasking Model were applied with selected features. **VTm:** MAE	Gray Prediction Model obtained better results with MAE in the range of 0.1–0.2.
J. Lin et al. [15]	**SM:** Sina Weibo **PI:** Big-Five (conscientiousness, extroversion, agreeableness) **LN:** Chinese **CT:** Text **DT:** 968,854 posts	**PP:** - **MT:** In the first experiment, a rule-based personality prediction approach was proposed according to the emotion, meaning, and type of content in the text. The proposed method and Naive Bayes, SVM, Logistic Regression, and Decision Tree methods were compared. In the second experiment, the proposed personality-based emotion classification method was presented. **VTm:** Precision, recall, f-measure	When the results obtained in the first experiment were examined, it was shown that the proposed method was better with precision, and the other methods were better with recall.
T. Tandera et al. [16]	**SM:** Facebook **PI:** Big-Five **LN:** English **CT:** Post **DT:** 10,000 posts were taken from 250 users from the myPersonality dataset. The second dataset was obtained by tagging 150 manually collected user data on applymagicsauce.com.	**PP:** a, b, c, d, e, f, p **MT:** Deep Learning methods (MLP, LSTM, GRU, and CNN) and traditional machine learning methods (Naive Bayes, SVM, Logistic Regression, Gradient Boosting, and Linear Discriminant Analysis) were compared. **VTm:** Accuracy, 10-fold cross-validation	In the myPersonality dataset, the best results were obtained with MLP, and the best results in the second manually collected dataset were obtained with MLP and CNN methods.
M. Vaidhya et al. [17]	**SM:** Facebook **PI:** Big-Five **LN:** English **CT:** Post **DT:** myPersonality, 250 users	**PP:** a, c, d, e, o, q, r, s, t, u, v **MT:** After vectorization with TF-IDF, feature selection was made with PCA. Later, KNN and SVM classification methods were used. **VTm:** k value for KNN, precision, recall, f-measure	The best results in KNN were obtained with different k values for each personality trait. In SVM, on the other hand, the prediction success in personality traits changes according to the use/non-use of PCA.
V. Varshney et al. [18]	**SM:** Facebook, Twitter, Google + and others **PI:** Big-Five **LN:** English **CT:** Posts with text **DT:** myPersonality and others	**PP:** a, b, c, d, e, f, o, s, u, v **MT:** Multinomial Naive Bayes, KNN, and SVM methods were used. **VTm:** -	It was presented that each method would show different success rates in different situations in the estimation processes made with the proposed methods; therefore, it was presented that the most accurate result could be obtained with the majority voting method according to the situation.
R. Akhtar et al. [19]	**SM:** Facebook **PI:** Hogan Development Survey (HDS) **LN:** English **CT:** Post **DT:** 51,712 users	**PP:** h, n, w **MT:** Ordinary Least Squares (OLS) and Lasso algorithms were used. **VTm:** k-fold-cross-validation, min, max, mean, standard deviation, skew, kurtosis, standard error	It was observed that the Excitable, Dutiful, and Bold scales have a strong relationship with language; the Cautious, Colorful, and Leisurely scales have a weaker relationship with language.
B. Ferwerda and M. Tkalcic [20]	**SM:** Instagram **PI:** Big-Five Inventory 44-item **LN:** English **CT:** Content properties, Image **DT:** 54,962 pictures from 193 users	**PP:** k, n **MT:** Besides the ZeroR method, M5′ rules, Random Forest, and Radial Basis Function Network were used for comparison. **VTm:** 10-fold cross-validation, root-mean-square error (RMSE)	Looking at the results obtained, different success rates were observed in different personality traits in both text and picture.
M.	**SM:** Facebook	**PP:** d, s, u, v, w	When all scales were taken as a basis, the

(*continued on next page*)

Table 3 (*continued*)

Study	Dataset Properties	Model	Result
Hassanein et al. [21]	**PI:** Big-Five **LN:** English **CT:** Post **DT:** myPersonality	**MT:** WordNet was used for semantic similarity measurement. Inputs were created as 3 different vectors and each vector was tested with the specified method. Vectors vec1 (Noun, Adjective, Verb, and Pronoun), vec2 (All Nouns), vec3 (Nouns with emotional impact – NRC dictionary of emotion) vectors were compared using JCN and Path Length similarity measurement methods available on WordNet. **VTm:** Accuracy, precision, recall, f-measure	JCN measurement method was found to be more successful in general. In addition, the vector created as vec1 was found to be better than other vectors with accuracy (0.64) and f-measure (0.65) values.
P. Howlader et al. [22]	**SM:** Facebook **PI:** Big-Five **LN:** English **CT:** Post **DT:** myPersonality, 115,872 users	**PP:** d, h, m, n, s **MT:** Different datasets were created with LDA and LIWC features. These datasets were tested using Linear Regression, Support Vector Regression (Linear (L-SVR), Polynomial (P-SVR), and Radial Basis Function (RBF-SVR) methods. **VTm:** MSE	In the experiments, it was determined that P-SVR and RBF-SVR were generally more successful than other methods.
Y. Kim and J. H. Kim [23]	**SM:** Instagram **PI:** Big-Five **LN:** - **CT:** Image **DT:** 25,394 pictures from the accounts of 179 university students	**PP:** k, l **MT:** Pearson's Correlation **VTm:** RMSE	The results showed that the extraversion scale was related to gender, the extraversion, agreeableness, and openness scales were related to the number of faces and the emotional states of the faces on the pictures.
Y. Mao et al. [24]	**SM:** Facebook **PI:** Big-Five **LN:** English **CT:** Post **DT:** myPersonality	**PP:** h, n, s, v **MT:** KNN, Naive Bayes, and Decision Tree methods were used by creating 3 different feature sets (1: 4 general features, 2:1 + psychological features, 3:1 + style features). In addition to the dataset in which all features were discussed, tests were conducted on 3 datasets obtained using Pearson's correlation and Particle Swarm Optimization methods. **VTm:** f-measure, 10-fold cross-validation	In the experiments carried out according to the types of features, different success rates were obtained in each scale with the feature sets containing psychological and stylistic features. In addition, it was determined that the feature set determined by Particle Swarm Optimization was better with great success rates in the analysis based on feature selection.
M. S. Raje and A. Singh [25]	**SM:** Twitter **PI:** 16-Personality Test (MBTI) **LN:** - **CT:** Tweet **DT:** Over 1 million tweets were collected from 450 twitter users	**PP:** h, n **MT:** Pearson's correlation analysis was performed between Twitter data and personality test results. Later, the obtained features and Artificial Neural Network and Logistic Regression methods were used in the created dataset. **VTm:** Accuracy	It was determined that the relationship between Twitter data and many personality test features was high. As a result of testing the classification methods, the accuracy rates varied between 51% and 59% for different scales.
M. M. Tadesse et al. [26]	**SM:** Facebook **PI:** Big-Five **LN:** English **CT:** Post, user information (network size, betweenness, density, brokerage, and transitivity) **DT:** myPersonality	**PP:** h, m, o, s, v **MT:** For feature selection, SNA, LIWC, SPLICE, and their combinations were used. With the obtained 4 different feature sets, XGBoost, Linear Regression, Gradient Boosting, and SVM methods were used. **VTm:** Pearson's Correlation, accuracy	In tests on 4 different feature sets, XGBoost showed a greater success rate than other methods. The highest success rate was seen in the feature set obtained with SNA.
D. Xue et al. [27]	**SM:** Facebook **PI:** Big-Five **LN:** English **CT:** Post **DT:** myPersonality, 115,872 users	**PP**: e, o, r, s, v **MT:** Feature vector obtained using AttRCNN with Statistical Language features and Document vector (Doc2Vec), combinations of feature sets consisting of vectors derived from	As a result of the tests, the best success rates were observed with the features and combinations obtained with AttRCNN. In the estimation methods, the best success rate was obtained with SVR and GBR.

Table 3 (*continued*)

Study	Dataset Properties	Model	Result
		RCNN-CNN and CNN tested using methods such as Support Vector Regression (SVR), Gradient Boosting Regression (GBR), Random Forest, and MLP. **VTm:** MAE, 5-fold cross-validation	
A. Marouf et al. [28]	**SM:** Facebook **PI:** Big-Five **LN:** English **CT:** Post **DT:** myPersonality	**PP:** e, f, h, n **MT:** In feature extraction from the dataset, psycho-language features and feature sets based on language features were created. Naive Bayes, Decision Tree, Random Forest, Linear Regression (Simple Logistic Regression - LogitBoost), and SVM (Sequential Minimal Optimization) methods were used on these feature sets. **VTm:** Precision, recall, f-measure, accuracy, 10-fold cross-validation	Among the estimation methods applied, the Naive Bayes method showed the best success rate with psycho-language features.
H. Zheng and C. Wu [29]	**SM:** Facebook **PI:** Big-Five **LN:** English **CT:** Post **DT:** The myPersonality dataset and the unlabeled dataset of 9856 users	**PP:** h, n, s, v **MT:** Pseudo Multi-view Co-training (PMC) method was proposed because of its advantage in large unlabeled data. In the first experiment, the PMC method was trained with the features obtained with LIWC. In the second experiment, the PMC method was applied with the properties obtained with both LIWC and unigram. **VTm:** Precision, recall, f-measure	The mean f-measure value in personality trait prediction was 0.66 in Experiment 1 and 0.67 in Experiment 2. The proposed method showed better results in all criterion values than the values of the compared study.
N. H. Jeremy et al. [30]	**SM:** Twitter **PI:** Big Five Traits **LN:** Indonesian **CT:** Quantities of Tweets, Followers, Followed, Tweets, Retweets, Tagging, Replies, and Hashtags **DT:** 508 twitter data from unpublished manuscript labeled High: H and Low: L based on 5 personality traits.	**PP:** j, n, x **MT:** kNN, J48, Random Forest, SVM, Naif Bayes **VTm:** Precision, recall, f-measure	The best f-measure value was obtained by the Random Forest method as 0.744.
E. Tutaysalgir et al. [31]	**SM:** Twitter **PI:** Big Five Traits **LN:** Turkish **CT:** Tweet **DT:** Data belonging to 40 volunteers were used as ground data and 2000 random users were used for the clustering test.	**PP:** a, f, m, o, q, s, u **MT:** K-means, Agglomerative Clustering **VTm:** Error rate, silhouette coefficient	K-means clustering was said to have the best silhouette correlation scores with Word2Vec. The most balanced results were obtained with k = 2.
A. A. Marouf et al. [32]	**SM:** Facebook **PI:** Big Five Traits **LN:** English **CT:** Post **DT:** myPersonality	**PP:** a, b, e, f, g, j, n, w **MT:** Naive Bayes, Decision Tree, Random Forest, Simple Logistic Regression (SLR), SVM Feature Selection Methods: Pearson Correlation Coefficient, Correlation-based feature subset, information gain, symmetric uncertainly and chi-squared **VTm:** Accuracy	The best accuracy rate was presented as the average of 5 personality traits, respectively, as SVM 59.6%, SLR 58.28%, and RF 57.38%. By using Pearson Correlation-based feature selection, an accuracy rate between 61.89% and 70.08% was obtained with Naive Bayes and Random Forest methods.
Z. Guan et al. [33]	**SM:** Twitter, Facebook and Youtube **PI:** Big Five Traits **LN:** English **CT:** Text contents **DT:** myPersonality with 500 users, Youtube Personality with 400 records,	**PP:** h, j, n, y, z **MT:** Personality2vec (proposed method), AdaWalk, node2vec, DeepWalk, AttRCNN-CNN, CNN, Mairesse, Doc2Vec, Random Guess **VTm:** MAE	The best MAE value in all datasets was obtained with the proposed method personality2vec. 0.4852–0.5971 for myPersonality, 0.5824–0.7213 for YoutubePersonality, and 0.1007–0.1425 for PAN2015 English.

(*continued on next page*)

Table 3 (*continued*)

Study	Dataset Properties	Model	Result
	and PAN2015 with 294 records in the English dataset		
S. Han et al. [34]	**SM:** Sina Weibo **PI:** NEO-FF 60 questionnaire **LN:** Chinese **CT:** Micro-blogging content **DT:** Dataset with 3.3G Micro-blog repository and Sina Weibo dataset with 400 participants	**PP:** h, j, n, s **MT:** Logistic Regression, SVM, Random Forest Feature Category: CLIWC, BOW, LSA + SOA, Personality Lexicon (PL), PL_Hownet, PL_Tongyici, PL_ITH **VTm:** F-measure	The best f-measure value was obtained as 0.704 in the Random Forest method in the PL_ITH feature category on average.
A. S. Khan et al. [35]	**SM:** Social Media Platforms **PI:** MBTI **LN:** English **CT:** Text-based content **DT:** MBTI dataset with 422,845 records with last 50 posts from 8675 records	**PP:** d, f, s, v **MT:** XGboost (proposed method), NB, KNN, Decision Tree, Random Forest, MLP, SVM, MNB, Logistic Regression, SGD Resampling methods were used in the dataset. **VTm:** Accuracy, Recall, Precision, F-measure	XGboost was found to have very good values in all tests. Especially the results obtained with resampling methods were close to 100% with XGboost.
J. Sun et al. [36]	**SM:** Facebook **PI:** Big Five Traits **LN:** English **CT:** Post, Like **DT:** myPersonality 22,000 users 3,183,816 posts and 301,105 likes	**PP:** s, t **MT:** Logistic Regression, Support Vector Regression, Deep Neural Network methods were used. Inputs were treated as Document Vector (DV) and Rating Vector (LV). **VTm:** MAE	The best MAE values were obtained between 0.5038 and 0.-6354 for 5 personality traits by DV + LV and DNN methods.
X. Sun et al. [37]	**SM:** Youtube, Facebook, Twitter **PI:** Big Five Traits **LN:** English **CT:** Text-based content **DT:** Youtube Personality, MyPersonality, PAN2015, OpenPsychometrics	**PP:** s **MT:** AdaWalk, node2vec, Deep2Walk, 2CLSTMs, doc2vec, Kampman, mairesse, TFIDF, Wei, random guess **VTm:** RMSE	In tests on 4 datasets, the best RMSE values were generally obtained with AdaWalk.
P. Wang et al. [38,39]	**SM:** Sina Weibo **PI:** Four short-answer questions, Proactive Personality Questionnaire **LN:** Chinese **CT:** Micro-blogging **DT:** A total of 4955 posts from 901 participants and the Weibo text dataset	**PP:** d, j, n, o **MT:** SVM, XGboost, KNN, NB, LR **VTm:** Accuracy, F-measure, SEN, SPE, PPV, NPV, AUC	The best results were obtained with the use of both datasets. The SVM and NB methods had the best results.
S. Wang et al. [40]	**SM:** Facebook **PI:** Big Five Traits **LN:** English **CT:** The type of entity the likes and Like **DT:** myPersonality, 19,700 Facebook user data with 8530 likes	**PP:** t **MT:** LR, AttRNN, BiGRU **VTm:** Accuracy	The best results were obtained with the AttRNN method with an accuracy rate of 0.29–0.48.
J. Zhao et al. [41]	**SM:** Facebook **PI:** Big Five Traits **LN:** English **CT:** Post **DT:** 6893 train, 3024 test datasets	**PP:** s **MT:** Bayesian Network, Random Forest, SVM, Attention-based LSTM **VTm:** Precision, Recall, F-measure	The best f-measure rate was 72.2% with the Attention-based LSTM method.
Y. Jiang et al. [42]	**SM:** Sina Weibo **PI:** DISC **LN:** Chinese **CT:** Micro-blogging **DT:** 198 users, only 30 pages of micro-blogging data per user	**PP:** n, o **MT:** NB, Bayes Net, KNN (k = 1,3,5), J48, LibSVM **VTm:** 10-fold cross-validation, Precision, Recall, F-measure, MVV, ROC, PRC, Accuracy	The best f-measure rate was 80.5% with the KNN (k = 1). In other evaluation criteria, the best results were obtained with the KNN method.
S. Basaran	**SM:** Facebook	**PP:** o, s	Two different schemes were created with

Table 3 (*continued*)

Study	Dataset Properties	Model	Result
et al. [43]	**PI:** Big Five Traits **LN:** English **CT:** Likes, Tags, Updates, Events, Groups, Friends, Birthday, Age, Gender, Relationship Status, Interest, Time Zone, and Network Size **DT:** 7438 users, myPersonality	**MT:** Artificial Neural Network **VTm:** 10-fold cross-validation, Accuracy	and without 10-fold cross-validation. The best results were obtained using 10-fold cross-validation.

*Abbreviations: SM (Social media), PI (Personality inventory), LN (Language), CT (Content-type), DT (Detail), PP (Preprocess), MT (Method), VTm (Validation and Test metrics).

analysis results. At this stage, the goal was to obtain the best results of the model created with the analysis of the results by evaluating each process step and model parameter. Considering that the relevant tests were carried out adequately, the PAN-2015 Personality Dataset was compared with the preprocessing and the results of the analysis between models.

3.1. Cover, data aggregation and data labeling

Using 13 words [49] with both positive and negative meanings presented to provide diversity for the users, 1000 tweets were collected for each word between 04:07 and 04:21 (GMT + 3) on the 30th November 2020. By separating the users who wrote 13,000 tweets in total, 12,101 discrete users were obtained. From this user list, the most recent 3200 tweet records of 11,984 users who shared publicly were obtained. The obtained users were then reduced by eliminating those who did not have English content and through the limitations of the IBM Personality Insight service. In addition, users with a maximum of 100 tweets and users with a total tweet data of a maximum of 1200 words were cleaned. Finally, a dataset with a total of 11,769,203 tweet content with 5081 user data was obtained. In this dataset, if a tweet was a reply to another tweet, it was marked with the [REPLY] tag.

The tweet data of 5081 users were converted to the input format of the IBM Personality Insight service, and the results of 35 different personality traits were collected as service output. Obtained personality traits were used as class labels for each user. Relevant personality traits were grouped under 5 main headings. These are openness (O), conscientiousness (C), extraversion (E), agreeableness (A), and neuroticism (N). The other 30 personality traits are the subpersonality traits used to determine the main personality traits, including 6 for each main heading. Each personality trait has decimal values between 0 and 1. This value represents the percentage of the personality trait of the tweet information. It was thought that the use of 5 main personality traits would be sufficient for the experimental study based on the working environment constraint, the size of the data, the compatibility with the PAN-2015 dataset personality traits labels, and the use of the main personality traits.

3.2. Tweet analysis and preprocessing

A tweet can be a combination of many assets and text. Since these assets vary greatly according to the tweet content, it can also make it difficult to find a common denominator for generalization. For example, it may not be easy to find a point of intersection between a mention shared in a tweet and a mention shared in another tweet. In addition, these entities that make it difficult to establish a relationship can bring us to an inextricable position in terms of processing power as it will

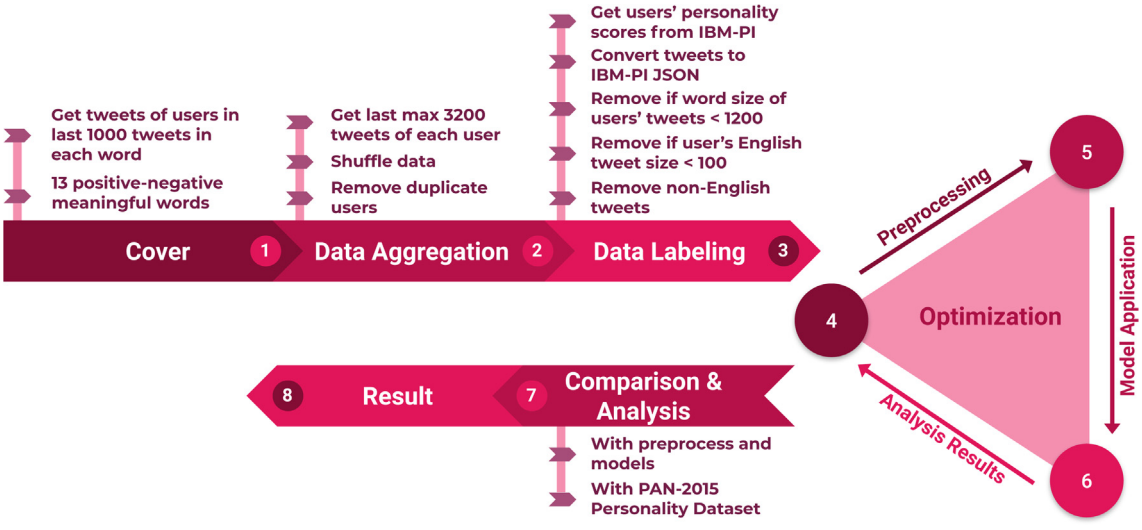

Fig. 1 Research method workflow of our study. (Graphical abstract).

bring the vocabulary word list to gigantic size. At this point, there are several preprocessing steps that transform entities into category tags (mention, hashtag, URL, image, email, retweet, newline), into both meaning and category tags (emoji, emoticon), delete uninterpretable characters (space chars, quotes, numbers, and punctuations), clear labeled data (tags), clear similar words (continuous and discrete duplicate words) and the application of natural language processing steps (post tagging, clear stop words, lemmatization, and stemming) in datasets. The ability to use these preprocessing steps with each other is important on the success rate. Fig. 2 presents which processing steps are applied in the entire dataset, and Fig. 3 presents which preprocessing can be applied to a tweet.

While the areas defined as Checkbox can be selected individually, the areas defined as Radio Button can only be presented as a selection within the group, they are in.

Based on the entire dataset, the number of tweets for each user varies. In this way, the effect of each user on the prediction model can be analyzed by selecting the same maximum number of tweets (select sub tweets in Fig. 2). In other words, if the user has fewer tweets than the maximum selected number of tweets, all tweets of the user are retrieved. In addition, the effects of the tweets on the model can be investigated according to the word count (remove tweets by word count in Fig. 2).

3.3. From text data to computable numeric data

Before applying certain mathematical operations to text data, it is necessary to digitize the text. Although the digitization methods used in deep learning methods are various, two digitization methods were used in this study. The first of these was executed by creating a word pool, which is widely used and evolving into a numerical expression corresponding to each word in the relevant text according to its frequency. The second method was the FastText [50] method, which provides vectorization by subsampling with CBOW and Skip-Gram methods, which have gained popularity in recent years. At this stage, the main point to consider is that only the training dataset should be used during the dictionary creation phase. Although the use of the test dataset increases the success rate of all outputs, it does not generalize. Using our training dataset and PAN-2015 training dataset, a 1-gram (ngram) FastText model with 3 windows with 0.025 alpha with 300-dimensional vectors for the FastText method was trained separately for each dataset. All words were converted to lowercase before digitizing.

3.4. LSTM-based models

Long Short-Term Memory (LSTM) [51–53] is one of the most used variants of Deep learning methods, and Recurrent Neural Network (RNN). The main reason for this is the success rates it has shown in solving many problems.

In the proposed artificial neural network model, firstly, the received tweet data was tokenized, and a dictionary was created. Vectorization was applied for each tweet in the dictionary, and the vectorized tweet information was presented to the model as input. Then the LSTM model was optimized to obtain the best result with its hyper parameters, and the proportional values of personality traits outputs were given. While optimizing LSTM models; batch size, number of epochs, number of hidden layers, different types of hidden layers (LSTM or Dense in + 2nd hidden layer), unit numbers of layers, activation functions (swish, sigmoid, tanh etc.), learning rates, dropouts (recurrent or sequential), and l1/l2 regularization, etc. were used. While optimizing the hyperparameters of the LSTM models, the rest of the parameters were kept constant while finding the ideal value of a parameter. In addition, learning was balanced with dropout, regularization, and controlling the change of the min delta to prevent overfitting. In the model, the Swish activation function was used as the activation function, which achieved better performance than the commonly used activation functions [48]. Additionally, Adam (Learning rate 0.001 and 0.0001) was used as an optimization function, Mean Squared Error (MSE) was used as loss function, and Root Mean Squared Error (RMSE) was used as an evaluation metric. In all models, both normal and recurrent dropouts were taken as 0.3.

There are many different variants of the LSTM model. In this study, Bidirectional LSTM (Bi-LSTM) [54] model was used with the basic LSTM model. The only difference of Bi-LSTM from normal LSTM is that it creates the model by evaluating the input bidirectionally.

The structural representation of our model is presented in Fig. 4, and the model features are presented in Table 4.

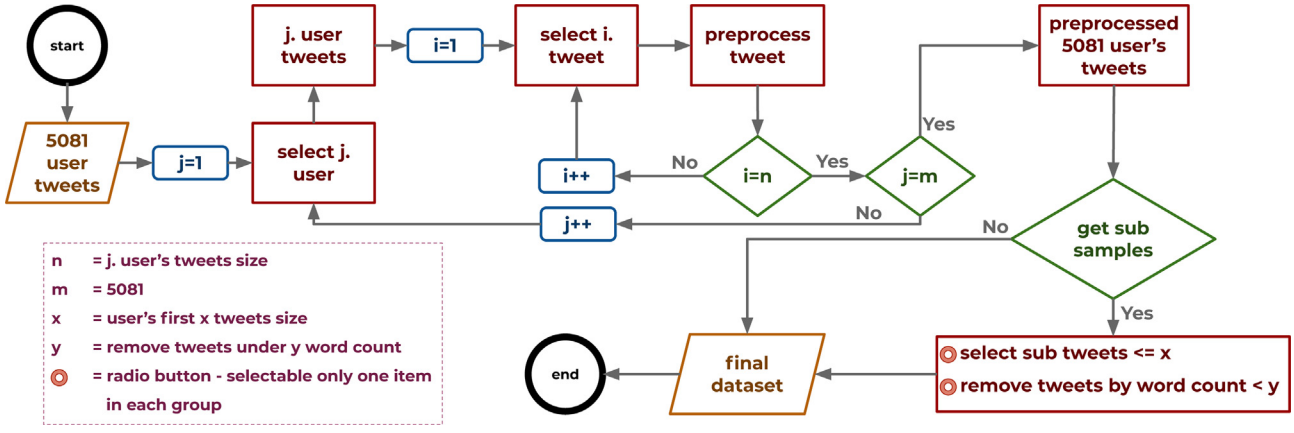

Fig. 2 Preprocessing flowchart for the dataset.

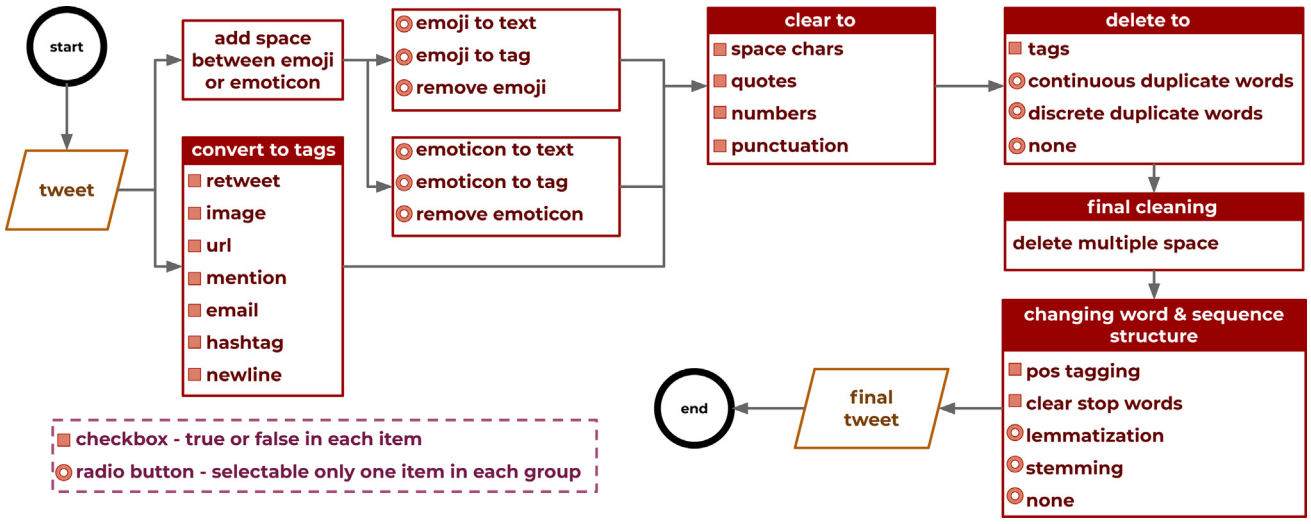

Fig. 3 The preprocessing steps for a tweet (preprocess tweet in Fig. 2).

3.5. Working environment

The working environment is presented in two different aspects as hardware and software.

The working environment with AMD Ryzen 9 3900X 12-Core processor, 64 GB (4*16 GB) 3600 MHz Ram, 500 GB (3400 read/2500 write) SSD, and a Nvidia Chipset RTX 2080Ti 11 GB graphics card were used as hardware. In particular, harmony and balance should be observed in data transmission and processing between hardware.

In terms of software, model tests were carried out with TensorFlow 2.4 on CUDA 10.2 using the Python 3.8.3 programming language on the Windows 10 operating system. In the preprocessing steps, in addition to the basic python codes, emoji and emot libraries were used to make sense of emojis and emoticons. In addition, the NLTK 3.5 library was used for part of speech tagging (POS-Tagging), clear stop words, lemmatization, and stemming methods, and the Gensim 3.8.3 library was used for FastText.

3.6. PAN-2015 personality dataset

The PAN-2015 Personality Dataset is divided into training and test datasets consisting of 4 languages, however only the English datasets were used in this study. There is 152 user information for the English training dataset and 142 user information for the English test dataset. There are 27,344 tweets in total in the two datasets. The PAN-2015-EN Personality Dataset was used as the benchmark dataset on the dataset and models obtained in this study.

3.7. Comparison of our dataset and PAN-2015-EN personality dataset

Our main purpose in creating our dataset was the low amount of data in the available accessible datasets and its scarcity of assets and word diversity. It was also the need for more data by complex models. In addition, there are sub-personality traits that express the main personality traits in the newly created dataset. In order to not complicate the study, evaluation and analysis were not included, as the success rate of the sub-personality traits determines the main value of the main personality traits.

It is known that the number of results to be obtained from the questionnaires and clinical trials to create a new data set will be limited. Due to both the low number of people who will present the relevant data and the fact that the not every person who participates will actively share on their social media accounts.

The numerical information of the two datasets is presented in Table 5, and word clouds according to asset and word density are presented in Fig. 5.

4. Analysis

Our dataset contains 11,769,202 tweet information from 5081 users who tweeted with 13 positive/negative words and were tagged with values between 0 and 1 with the IBM Personality Insights service based on the Big Five Inventory model. The content analysis of this dataset was made with dynamic methods, and certain preprocessing steps were determined according to the assets and structures in its content. Different combinations of these preprocessing steps can produce different results. For example, when converting hashtag phrases to [HASHTAG] tags is disabled and punctuation phrases cleanup is turned on, hashtags are treated as regular words. In addition, there are 638 tweets containing image information, 3,275,304 tweets containing URL information, 9,280,208 tweets containing mention information, 2384 tweets containing email information, 1,200,813 tweets containing hashtag information, and 1,996,337 tweets containing \n (newline) information.

The codes are defined below to more easily express the preprocessing steps in the representations. The codes are Tag Words (**TW**), Emoji and Emoticon (**EE**), Clear to Things (**CL**), Delete to Tags/Words (**DL**), and Word/Sentence (**WS**).

- **TW1:** Retweet to Tag [RETWEET]
- **TW2:** Image to Tag [IMAGE]
- **TW3:** URL to Tag [URL]
- **TW4:** Mention to Tag [USER]

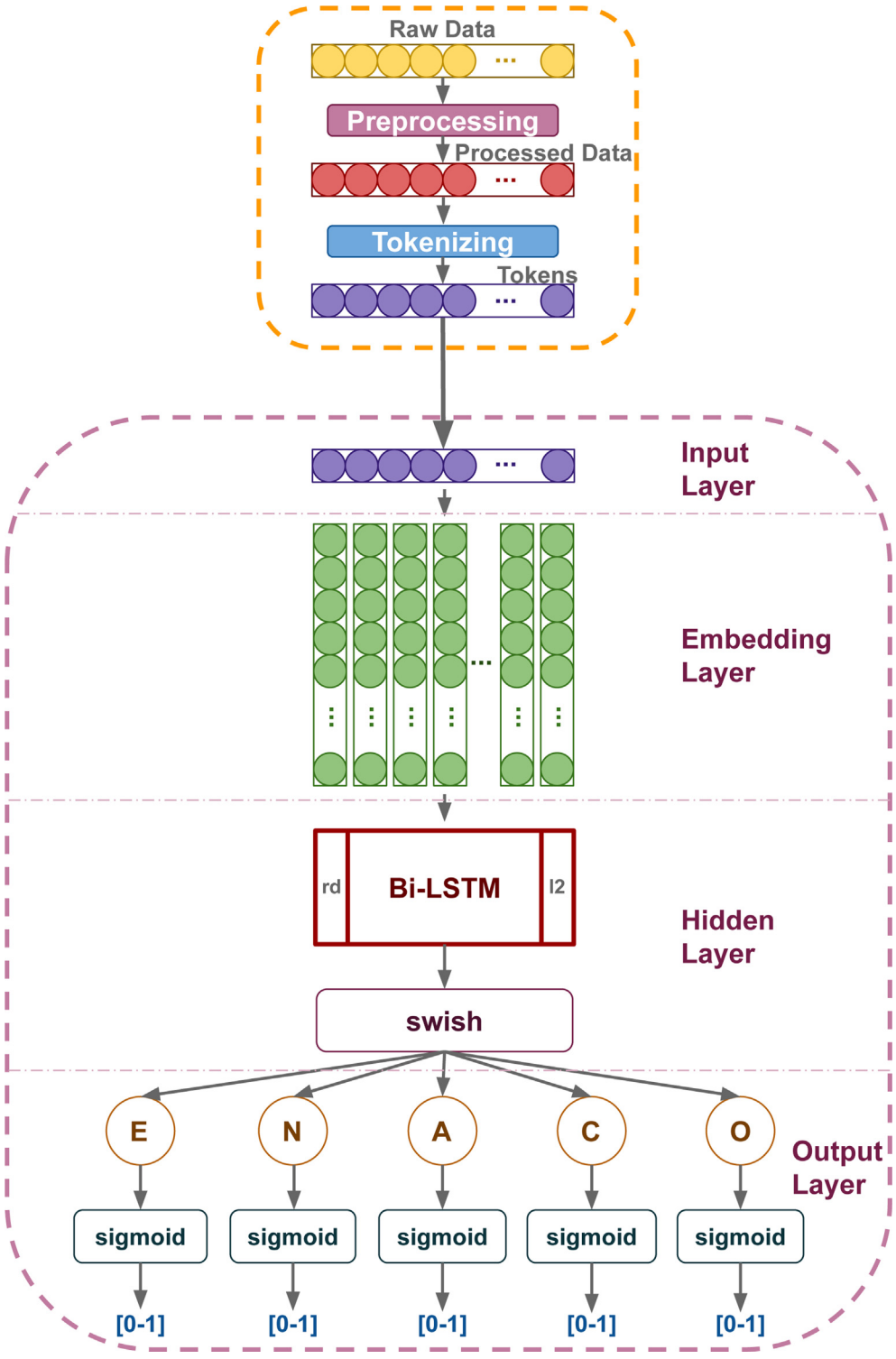

Fig. 4 Structural representation of our model (rd: recurrent dropout, l2: l2 regularizer).

- **TW5:** Email to Tag [EMAIL]
- **TW6:** Hashtag to [HASHTAG]
- **TW7:** Newline (\n) to [NEWLINE]
- **EE1:** Emoji to Text Mean
- **EE2:** Emoji to Tag [EMOJI]
- **EE3:** Remove Emoji
- **EE4:** Emoticon to Text Mean
- **EE5:** Emoticon to Tag [EMOTICON]
- **EE6:** Remove Emoticon
- **CL1:** Clear to Space Chars
- **CL2:** Clear to Quotes
- **CL3:** Clear to Numbers
- **CL4:** Clear to Punctuation Marks
- **DL1:** Delete to Tags ([URL], [USER] etc.)

Table 4 Numerical properties of our model.

Property	Value
Sequence length	48
Embedding dimension	96
Batch size	1024
Epoch	max: 1000 and if the minimum delta is 0.0001 by validation loss, the process is finished
LSTM - units	256
LSTM - recurrent dropout rate	30%
LSTM - l2 activity regularizer value	0.001
LSTM - activation function	swish
Output - activation function	sigmoid
Loss function	MSE
Optimization function	Adam (lr: 0.0001, b1: 0.9, b2: 0.999, e:1e-7)
Evaluation function	RMSE

Table 5 Numerical comparison of our dataset and the PAN-2015-EN personality dataset.

Information	Our dataset	PAN-2015-EN
Number of users	5081	294 (154 train and 142 test)
Number of tweets	11,769,202	27,343
Maximum tweet-length	890	80
Minimum tweet-length	1	1
Most common asset	User	User

- **DL2:** Delete to Continuous Duplicate Words
- **DL3:** Delete to Discrete Duplicate Words
- **WS1:** PoS Tagging
- **WS2:** Clear Stop Words
- **WS3:** Lemmatization
- **WS4:** Stemming

An analysis was made using different combinations according to the model and preprocessing features, and the hyperparameter balancing operations were carried out on the models, in many varieties in order to obtain the best results and to analyze the differences in the results. In our dataset consisting of 11,769,202 tweet records with 5081 users, 60% of the dataset was used for LSTM model training, 20% of the dataset for validation, and 20% for testing. Since the training and test datasets of the PAN-2015 dataset were presented separately, 25% of the training dataset was used as a validation dataset. Root Mean Square Error - RMSE (smallest value best result) was used as the evaluation metric.

It is important to select all preprocesses between TW1-TW7 and CL1-CL4 as meaningless information was found and there is no correlation between the tweets. These 11 preprocesses are called standard preprocessing steps and were used in all subsequent tests. First, the performance of Natural Language Processing methods, which was applied to words and sentences was evaluated on the models. In addition to the standard preprocessing steps, the effects of all preprocessing steps between WS1 and WS4 on LSTM and Bidirectional LSTM (Bi-LSTM) models with the same features are shown in Table 6. Since the RMSE values obtained show different success rates for each personality trait, Arithmetic Mean (A.M.) was used to evaluate the overall success rate of the model. In addition, in vectorizing textual expressions, the standard frequency-based vectorizing process was first applied.

As a result of the analysis made according to the LSTM type and NLP pretreatment presented in Table 6, it is observed that Bi-LSTM gave better results than basis LSTM. In addition, similar results were obtained when no operation was performed (None) and WS2 and WS3 operations were performed in the Bi-LSTM model. Although no operation makes a small difference in success rates with WS2 and WS3 preprocessing, the size of the dictionary can be reduced, and more concentration on distinctive words can be achieved. After this stage of the tests, Bi-LSTM was used, but NLP preprocessing can be used in different options that are WS2, WS3, and None.

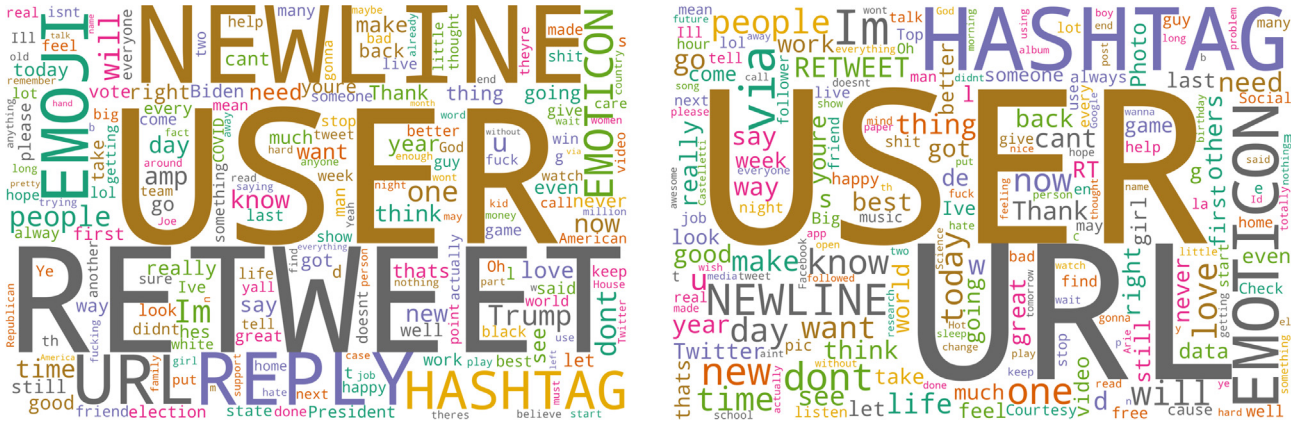

Fig. 5 Comparison of our dataset and PAN-2015-EN personality dataset by word cloud. **(a)** Word cloud of Our dataset, **(b)** Word cloud of PAN-2015-EN personality dataset.

Table 6 LSTM-based model results with Root Mean Square Error (RMSE) according to Natural Language Processing.

Preprocess	Model	Our dataset						PAN-2015					
		E	N	A	C	O	A.M.	E	N	A	C	O	A.M.
WS1	LSTM	0.1602	**0.1817**	0.2133	0.2341	0.1652	0.1909	0.1611	0.2277	0.1521	0.1501	0.1624	0.1707
	Bi-LSTM	0.1590	0.1821	0.2119	0.2329	0.1654	0.1902	**0.1595**	0.2271	0.1504	**0.1471**	0.1612	0.1691
WS2	LSTM	**0.1573**	0.1821	0.2088	0.2307	0.1644	0.1886	0.1607	0.2272	0.1507	0.1482	0.1608	0.1695
	Bi-LSTM	**0.1574**	**0.1811**	0.2086	0.2316	**0.1637**	**0.1884**	**0.1601**	0.2266	0.1496	0.1477	0.1609	**0.1690**
WS3	LSTM	0.1585	0.1857	0.2089	0.2303	0.1654	0.1897	**0.1600**	0.2266	0.1505	0.1477	0.1613	0.1692
	Bi-LSTM	**0.1571**	**0.1817**	**0.2078**	0.2300	0.1655	**0.1884**	0.1602	**0.2264**	**0.1490**	0.1484	0.1601	**0.1688**
WS4	LSTM	0.1575	0.1825	0.2092	**0.2297**	0.1648	0.1887	0.1601	0.2265	0.1507	0.1466	0.1614	0.1691
	Bi-LSTM	0.1590	0.1835	0.2083	0.2314	0.1649	0.1894	**0.1600**	0.2267	0.1492	0.1479	0.1602	**0.1688**
None	LSTM	0.1602	0.1907	0.2091	0.2306	0.1694	0.1920	0.1607	0.2271	0.1506	0.1477	0.1607	0.1693
	Bi-LSTM	**0.1573**	**0.1815**	**0.2077**	**0.2298**	**0.1634**	**0.1879**	**0.1599**	0.2274	0.1495	0.1478	**0.1596**	**0.1688**

*Abbreviations: E (Extraversion), N (Neuroticism), A (Agreeableness), C (Conscientiousness), O (Openness), and A.M. (Arithmetic Mean).

In addition to the standard preprocessing steps, the differences in the success rates of emoji and emoticons on the proposed models according to the preprocessing type are presented in Table 7.

The model, in which emojis and emoticons are evaluated, uses WS2 and WS3 preprocessing steps in addition to the standard preprocessing steps on Bi-LSTM. Although the results of the models applied by transforming them into textual expressions and label expressions in these preprocessing steps differ in each personality trait, arithmetically they have very close results. In the results obtained, it would be more appropriate to convert emojis and emoticons to tags instead of increasing the capacity of dictionary and processing complexity by converting emoji and emoticons into textual expressions.

Model parameters were adapted with the same features and EE2&EE5 were added to standard features. After this test process, these preprocesses were also used in all of the models. Table 8 and Fig. 6 evaluate the effect that tagging expressions in the preprocessing steps had on the success rate in the transformations, depending on whether the tags are deleted or not.

Although the RMSE value obtained in the Extraversion personality trait was better in the case of DL1 = True, it is seen that the other 4 personality traits gave better results in the case of DL1 = False. When all personality traits were evaluated as mean, it was concluded that not performing (DL1 = False) the DL1 preprocessing step would be more appropriate in terms of the model success rate. As a result of this test, the DL1 preprocessing step was added to the standard preprocessing steps as DL1 = False for the subsequent model trials. The reason why the results obtained in Table 8 are lower than the results obtained in Table 7 is due to the fields that are required to be labeled.

Afterwards, the effects of repetitions of words and tags on the models according to the preprocessing types were evaluated. At this stage, the results obtained with the same model and standard preprocessing steps, depending on whether the DL2 and DL3 preprocessing steps were applied or not, are presented in Table 9.

In the absence of preprocessing repeated words, in the DL2 and DL3 preprocessing, results were close to each other in terms of arithmetic mean. However, it is believed that the reduction of sequential repetition and the shortening of the sentence lengths to be processed with DL2 or DL3 will both provide a gain for processing power and help increase the complexity in the models.

5081 users have different numbers of tweets, and each tweet has different lengths. The results of Bi-LSTM with the selection of the tweets with a certain ratio according to the user tweets are presented in Table 10, and the results of the restrictions on the Bi-LSTM according to the word count in the tweets are presented in Table 11.

Although the number of tweets from users was taken from different directions, and we tried to equalize the result distribution, the more data, the higher the performance. Therefore, it is concluded that obtaining all of the user's tweet information is important in terms of the success rate.

Table 7 Model results by emoji and emoticon preprocessing type (via RMSE).

Personality traits	Our dataset			PAN-2015		
	EE1 & EE4	EE2 & EE5	EE3 & EE6	EE1 & EE4	EE2 & EE5	EE3 & EE6
E	**0.1518**	0.1521	0.1550	0.1614	**0.1593**	0.1610
N	**0.1732**	0.1749	0.1757	0.2267	**0.2261**	0.2292
A	0.1979	**0.1975**	0.1991	0.1493	**0.1491**	0.1497
C	0.2163	**0.2155**	0.2174	0.1486	**0.1473**	0.1485
O	0.1574	**0.1571**	0.1588	0.1597	0.1605	**0.1593**
A.M.	**0.1793**	**0.1794**	0.1812	0.1691	**0.1685**	0.1695

Table 8 Model results based on whether or not the tags are deleted (via RMSE).

Personality traits	Our dataset		PAN-2015	
	DL1 = True	DL1 = False	DL1 = True	DL1 = False
E	**0.1534**	0.1546	0.1600	**0.1599**
N	0.1771	**0.1765**	**0.2268**	0.2272
A	0.2046	**0.2006**	**0.1491**	**0.1491**
C	0.2264	**0.2178**	**0.1470**	0.1473
O	0.1603	**0.1589**	0.1604	**0.1598**
A.M.	0.1843	**0.1816**	**0.1687**	**0.1687**

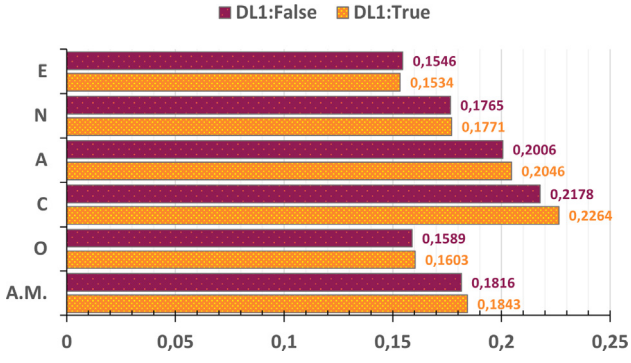

Fig. 6 Model results by emoji and emoticon preprocessing type (Our dataset, via RMSE).

Based on the length of the tweets, before the results in Table 11, it is necessary to know the distribution of the number of tweets according to the length of the tweets. In Fig. 7, shows the distribution of the number of tweets in our dataset, according to the tweet length. When the results in Fig. 7 and Table 11 are considered together, the effect of increasing the success rate with distinctiveness is seen as the tweet length increases. However, based on the distribution in Fig. 7, a large part of the dataset was eliminated by selecting 20 or more tweets. At this point, it was believed that the models' effect on generalization will be low. The selection above 50% can be seen as an ideal distribution for each dataset in order to obtain the best distribution. For our dataset, tweets with a length of 15-word count or more were included in the selection. In addition, in the tweet length distribution of the PAN-2015 dataset presented in Fig. 8, tweets with a length of 10-word count or more are included in the selection because they had 50% over the dataset. Since there were 52 records with 30 or more words in the PAN-2015 dataset, it is not seen in the distribution.

The comparative analysis of the preprocessing and vectorization method and the FastText method on the PAN-2015 dataset and our dataset is presented in Table 12. The Bi-LSTM method, in which we achieve the best performance in each model, is considered as a single hidden layer. According to the results obtained, it is seen that our final model performed better than FastText + BiLSTM model and was more suitable for generalization.

First, the process of convert meaningless entities to meaningful was applied. Later, it was observed that the prediction success rate of Bi-LSTM was high with WS2, WS3 and None types, using NLP methods and LSTM types. After that, the effect of Emoji and Emoticons on the success rate was investigated and it was presented that the best performance was obtained with EE1&EE4 and EE2&EE5 selections, but the best choice would be EE2&EE5 due to the corpus size. Next, it was seen that the cleaning of all tags (DL1) did not increase the success rate. It is presented that DL2 and DL3 methods would be useful in terms of processing performance and complexity in the cleaning of repetitive records (DL2, DL3, and None). Then, it was seen that using all of each users' content was effective in the prediction success rate. Afterwards, the effect of content-based word counts on the success rate was examined, and it was concluded that choosing more than half

Table 9 Model results on deletion of data that is continuous or discrete repetitive (via RMSE).

Personality traits	Our dataset			PAN-2015		
	DL2	DL3	None	DL2	DL3	None
E	**0.1510**	0.1518	0.1515	0.1608	**0.1586**	0.1599
N	0.1733	0.1732	**0.1727**	0.2292	**0.2263**	0.2270
A	**0.1968**	0.1979	0.1974	**0.1499**	0.1502	**0.1499**
C	**0.2147**	0.2163	0.2152	0.1486	**0.1479**	0.1480
O	**0.1563**	0.1574	0.1564	**0.1601**	0.1610	**0.1603**
A.M.	**0.1784**	0.1793	**0.1786**	0.1697	**0.1688**	**0.1690**

Table 10 Model results according to the selection in the number of tweets by users (via RMSE).

Personality traits	Number of Selected Tweets				
	250	500	1000	2000	3200
E	0.1669	0.1626	0.1574	0.1527	**0.1510**
N	0.1936	0.1882	0.1819	0.1760	**0.1733**
A	0.2128	0.2093	0.2044	0.1990	**0.1968**
C	0.2314	0.2254	0.2203	0.2155	**0.2147**
O	0.1741	0.1705	0.1657	0.1607	**0.1563**
A.M.	0.1957	0.1912	0.1859	0.1807	**0.1784**

Table 11 Model results by length of tweets (T.L.) (via RMSE).

T.L.	Our dataset						PAN-2015					
	E	N	A	C	O	A.M.	E	N	A	C	O	A.M.
0	0.1510	0.1733	0.1968	0.2147	0.1563	0.1784	0.1604	0.2264	**0.1492**	0.1481	0.1611	**0.1690**
5	0.1491	0.1707	0.1953	0.2121	0.1542	0.1762	0.1599	**0.2252**	0.1503	0.1478	0.1612	**0.1689**
10	0.1472	0.1654	0.1925	0.2129	0.1449	0.1725	**0.1597**	0.2278	0.1504	0.1472	**0.1597**	**0.1690**
15	0.1434	0.1594	0.1897	0.2109	0.1373	0.1681	0.1627	0.2338	0.1587	**0.1451**	0.1610	0.1722
20	**0.1422**	**0.1577**	0.1889	0.2112	0.1332	**0.1666**	0.1627	0.2422	0.1655	0.1482	0.1578	0.1753
25	0.1465	0.1618	0.1886	0.2097	0.1321	0.1677	0.1640	0.2527	0.1698	0.1616	0.2084	0.1913
30	0.1500	0.1622	**0.1824**	**0.2018**	**0.1217**	**0.1636**	0.1421	0.2518	0.1683	0.1787	**0.1598**	0.1801

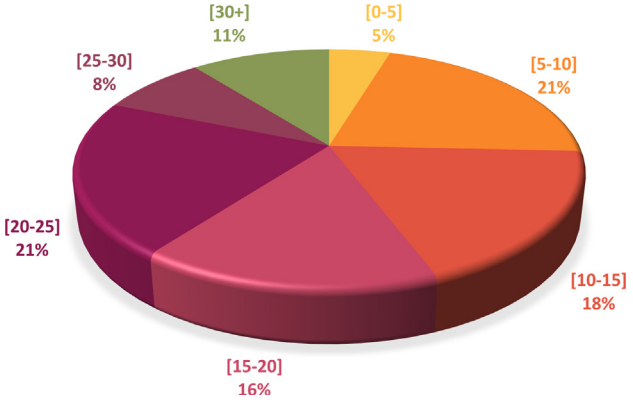

Fig. 7 Distribution of tweet numbers in the dataset by tweet length (our dataset).

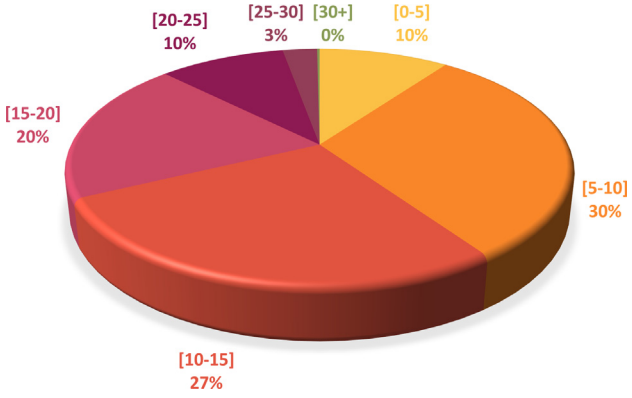

Fig. 8 Distribution of tweet numbers in the dataset by tweet length (PAN-2015 dataset).

of the dataset to obtain balanced diversity had a positive effect on the success rate. Finally, our model and FastText + Bi-LSTM method were compared to both our dataset and the PAN-2015 dataset. It was determined that a better prediction success rate was obtained by using our dataset and our model.

It is seen that sequential trials with structural analysis will affect the success of predicting personality traits for text-based content on any platform. The information types contained in the platform should be correctly extracted and its effect on predicting success rates should be investigated, and the results should be evaluated in terms of the correct data format and success rate.

5. Conclusions

To make a personality estimation based on the data in a sample, it is necessary to perform a structural analysis of the data in the relevant sample. In this study, a structural analysis was made on a Twitter sample, the assets in Tweet information and the structures that would create textual differences were examined. The aim was to keep generalization at the optimum level through the preprocessing steps. Along with the final model created, one of the most important points of this study is the created dataset. It was observed that the success rate of

Table 12 Comparison of Bi-LSTM model on datasets according to vectorization method (via RMSE).

Dataset	Preprocessing + Vectorizing	Personality traits					A.M.
		E	N	A	C	O	
Our dataset	Our Final Model	**0.1434**	**0.1594**	0.1897	0.2109	**0.1373**	**0.1681**
	FastText	0.2388	0.3432	0.2529	0.3677	0.2148	0.2834
PAN-2015	Our Final Model	0.1590	0.2262	**0.1509**	**0.1470**	0.1605	**0.1687**
	FastText	0.1879	0.2566	0.1772	0.1773	0.1670	0.1932

the high-dimensional and ideally distributed dataset had a balanced success rate when compared to the PAN-2015 dataset. In fact, although better results were obtained in some personality traits, according to the arithmetic mean the results were close to each other when all the results were evaluated together.

This study sought to keep the success rate at the most ideal level through hyperparameter tuning applied on LSTM models, which with the preprocessing steps required to optimize the generalization on Twitter data increased the success rate. However, the biggest problem observed as a result of the experiments was the hardware constraints arising from the processed data size and model complexity. At this point, the success rate has been brought to the optimal level by pushing the limits as far as the hardware constraints allowed. Another limitation of the study is the inability to reach other datasets for personality analysis from Twitter data. For this reason, it is thought that our dataset will be a useful source for future studies. Besides its contribution to the literature, we believe that our system can be used and provide benefits in areas such as the intelligence activities of law enforcement agencies, the recruitment processes of companies or institutions, situation analysis of the target audiences of advertising agencies, and the preliminary evaluation in environments serving in the field of psychology, etc.

In the future, we aim to increase the success rate with a hybrid model that could be created from two datasets. Although different structural tests of our dataset (Doc2Vec, TFIDF, etc.) were carried out, ideal results were not obtained. We believe that hybrid trials of different methods could be useful to make progress in this area. In addition, we will continue to create and improve a Turkish personality-traits dataset since no such dataset exists in the literature.

Funding

This research received no external funding.

CRediT authorship contribution statement

Muhammed Ali Kosan: Conceptualization, Methodology, Software, Validation, Investigation, Resources, Data curation, Writing – original draft, Visualization. **Hacer Karacan:** Conceptualization, Validation, Formal analysis, Data curation, Writing – review & editing, Supervision, Project administration. **Burcu A. Urgen:** Validation, Formal analysis, Writing – review & editing, Supervision, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] TDK, İstihbarat Kelime Anlamı, Türk Dil Kurumu. http://www.tdk.gov.tr/index.php?option=com_gts&arama=gts&guid=TDK.GTS.5c272c32beb8e3.27975555.
[2] OD, Intelligence Word Mean, Oxford Dictionary. https://en.oxforddictionaries.com/definition/intelligence.
[3] What is Intelligence?, Office of the Director of National Intelligence. https://www.dni.gov/index.php/what-we-do/what-is-intelligence.
[4] Internet Live Stats, Internet Live Stats. http://www.internetlivestats.com/.
[5] P.S. Dandannavar, S.R. Mangalwede, P.M. Kulkarni, Social Media Text - A Source for Personality Prediction, in: 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), 2018, pp. 62–65.
[6] M. Kosinski, D. Stillwell, T. Graepel, Private traits and attributes are predictable from digital records of human behaviour, Proc. Natl. Acad. Sci. 110 (2013) 5802–5805.
[7] N. Ahmad, J. Siddique, Personality Assessment using Twitter Tweets, Proc. Comput. Sci. 112 (2017) 1964–1973.
[8] Z. Ahmad, S.L. Lutfi, A.L. Kushan, R.T. Yixing, Personality Prediction of Malaysian Facebook Users: Cultural Preferences and Features Variation, Adv. Sci. Lett. 23 (2017) 7900–7903.
[9] N. Alsadhan, D. Skillicorn, Estimating Personality from Social Media Posts, in: 2017 IEEE International Conference on Data Mining Workshops (ICDMW), 2017, pp. 350–356.
[10] S.K. Bhatti, A. Muneer, M.I. Lali, M. Gull, S.M.U. Din, Personality analysis of the USA public using Twitter profile pictures, in: 2017 International Conference on Information and Communication Technologies (ICICT), 2017, pp. 165–172.
[11] S.C. Guntuku, W.S. Lin, J. Carpenter, W.K. Ng, L.H. Ungar, D. Preotiuc-Pietro, Acm, Studying Personality through the Content of Posted and Liked Images on Twitter, Assoc Computing Machinery, New York, 2017.
[12] S.G. Huang, J.H. Zheng, D. Xue, N. Zhao, Predicting Big-Five Personality for Micro-blog Based on Robust Multi-task Learning, in: B. Zou, M. Li, H. Wang, X. Song, W. Xie, Z. Lu (Eds.), Data Science, Pt 1, Springer-Verlag, Berlin, Berlin, 2017, pp. 486–499.
[13] A. Laleh, R. Shahram, Analyzing Facebook Activities for Personality Recognition, in: IEEE International Conference on Machine Learning and Applications (ICMLA), 2017, pp. 960–964.

[14] C. Li, J. Wan, B. Wang, Personality Prediction of Social Network Users, in: 2017 16th International Symposium on Distributed Computing and Applications to Business, Engineering and Science (DCABES), 2017, pp. 84–87.
[15] J. Lin, W. Mao, D.D. Zeng, Personality-based refinement for sentiment classification in microblog, Knowl.-Based Syst. 132 (2017) 204–214.
[16] T. Tandera, D. Hendro, R. Suhartono, Y.L. Wongso, Prasetio, Personality Prediction System from Facebook Users, Proc. Comput. Sci. 116 (2017) 604–611.
[17] M. Vaidhya, B. Shrestha, B. Sainju, K. Khaniya, A. Shakya, Personality Traits Analysis from Facebook Data, in: 2017 21st International Computer Science and Engineering Conference (ICSEC), 2017, pp. 1–5.
[18] V. Varshney, A. Varshney, T. Ahmad, A.M. Khan, Recognising personality traits using social media, in: 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), 2017, pp. 2876–2881.
[19] R. Akhtar, D. Winsborough, U. Ort, A. Johnson, T. Chamorro-Premuzic, Detecting the dark side of personality using social media status updates, Person. Individ. Differ. 132 (2018) 90–97.
[20] B. Ferwerda, M. Tkalcic, Acm, Predicting Users' Personality from Instagram Pictures: Using Visual and/or Content Features?, Assoc Computing Machinery, New York, 2018
[21] M. Hassanein, W. Hussein, S. Rady, T.F. Gharib, Predicting Personality Traits from Social Media using Text Semantics, in: 2018 13th International Conference on Computer Engineering and Systems (ICCES), 2018, pp. 184–189.
[22] P. Howlader, K.K. Pal, A. Cuzzocrea, S.D.M. Kumar, M. Assoc Comp, Predicting Facebook-Users' Personality based on Status and Linguistic Features via Flexible Regression Analysis Techniques, Assoc Computing Machinery, New York, 2018.
[23] Y. Kim, J.H. Kim, Using computer vision techniques on Instagram to link users' personalities and genders to the features of their photos: An exploratory study, Inf. Process. Manage. 54 (2018) 1101–1114.
[24] Y. Mao, D. Zhang, C. Wu, K. Zheng, X. Wang, Feature Analysis and Optimisation for Computational Personality Recognition, in: 2018 IEEE 4th International Conference on Computer and Communications (ICCC), 2018, pp. 2410–2414.
[25] M.S. Raje, A. Singh, Personality Detection by Analysis of Twitter Profiles, in: A. Abraham, A.K. Cherukuri, A.M. Madureira, A.K. Muda (Eds.), Proceedings of the Eighth International Conference on Soft Computing and Pattern Recognition, Springer International Publishing Ag, Cham, 2018, pp. 667–675.
[26] M.M. Tadesse, H. Lin, B. Xu, L. Yang, Personality Predictions Based on User Behavior on the Facebook Social Media Platform, IEEE Access 6 (2018) 61959–61969.
[27] D. Xue, L.F. Wu, Z. Hong, S.Z. Guo, L. Gao, Z.Y. Wu, X.F. Zhong, J.S. Sun, Deep learning-based personality recognition from text posts of online social networks, Appl. Intell. 48 (2018) 4232–4246.
[28] A.A. Marouf, M.K. Hasan, H. Mahmud, Identifying Neuroticism from User Generated Content of Social Media based on Psycholinguistic Cues, in: 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019, pp. 1–5.
[29] H.C. Zheng, C.H. Wu, M. Assoc Comp, Predicting Personality Using Facebook Status Based on Semi-supervised Learning, Assoc Computing Machinery, New York, 2019.
[30] N.H. Jeremy, C. Prasetyo, D. Suhartono, Identifying Personality Traits for Indonesian User from Twitter Dataset, Int. J. Fuzzy Log. Intell. Syst. 19 (2019) 283–289.
[31] E. Tutaysalgir, P. Karagoz, I.H. Toroslu, Clustering based personality prediction on turkish tweets, in: 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2019, pp. 825–828.
[32] A. Al Marouf, M.K. Hasan, H. Mahmud, Comparative Analysis of Feature Selection Algorithms for Computational Personality Prediction From Social Media, IEEE Trans. Comput. Soc. Syst. 7 (2020) 587–599.
[33] Z. Guan, B. Wu, B. Wang, H. Liu, Personality2vec: Network Representation Learning for Personality, in: 2020 IEEE Fifth International Conference on Data Science in Cyberspace (DSC), 2020, pp. 30–37.
[34] S.Q. Han, H.L. Huang, Y.Q. Tang, Knowledge of words: An interpretable approach for personality recognition from social media, Knowl.-Based Syst. 194 (2020) 20.
[35] A.S. Khan, H. Ahmad, M.Z. Asghar, F.K. Saddozai, A. Arir, H.A. Khalid, Personality Classification from Online Text using Machine Learning Approach, Int. J. Adv. Comput. Sci. Appl. 11 (2020) 460–476.
[36] J.S. Sun, Z.Q. Tian, Y.L. Fu, J. Geng, C.L. Liu, Digital twins in human understanding: a deep learning-based method to recognize personality traits, Int. J. Comput. Integr. Manuf. 14 (2020).
[37] X.G. Sun, B. Liu, Q. Meng, J.X. Cao, J.Z. Luo, H.Z. Yin, Group-level personality detection based on text generated networks, World Wide Web 23 (2020) 1887–1906.
[38] P. Wang, Y. Yan, Y.D. Si, G.C. Zhu, X.P. Zhan, J. Wang, R.S. Pan, Classification of Proactive Personality: Text Mining Based on Weibo Text and Short-Answer Questions Text, Ieee Access 8 (2020) 97370–97382.
[39] P. Wang, M. Yan, X. Zhan, M. Tian, Y. Si, Y. Sun, L. Jiao, X. Wu, Predicting Self-Reported Proactive Personality Classification With Weibo Text and Short Answer Text, IEEE Access 9 (2021) 77203–77211.
[40] S. Wang, L. Cui, L. Liu, X. Lu, Q. Li, Personality Traits Prediction Based on Users' Digital Footprints in Social Networks via Attention RNN, in: 2020 IEEE International Conference on Services Computing (SCC), 2020, pp. 54–56.
[41] J.H. Zhao, D.L. Zeng, Y.J. Xiao, L.P. Che, M.J. Wang, User personality prediction based on topic preference and sentiment analysis using LSTM model, Pattern Recognit. Lett. 138 (2020) 397–402.
[42] Y. Jiang, S. Deng, H. Li, Y. Liu, Predicting user personality with social interactions in Weibo, Aslib J. Inform. Manage. (2021).
[43] S. Başaran, O.H. Ejimogu, A Neural Network Approach for Predicting Personality From Facebook Data, SAGE Open 11 (2021), 21582440211032156.
[44] D. Azucar, D. Marengo, M. Settanni, Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis, Person. Individ. Differ. 124 (2018) 150–159.
[45] PAN Shared Tasks, Webis. https://pan.webis.de/.
[46] Francisco Rangel, Fabio Celli, Paolo Rosso, Martin Potthast, Benno Stein, W. Daelemans, Overview of the 3rd Author Profiling Task at PAN 2015, in: Linda Cappellato, Nicola Ferro, Gareth Jones, E.S. Juan (Eds.), CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, CEUR-WS.org, Toulouse, France, 2015.
[47] Miguel A. Alvarez-Carmona, A. Pastor Lopez-Monroy, Manuel Montes-y-Gomez, Luis Villasenor-Pineda, H.J. Escalante, INAOE's participation at PAN'15: Author Profiling task–-Notebook for PAN at CLEF 2015, in: Linda Cappellato, Nicola Ferro, Gareth Jones, E.S. Juan (Eds.), CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, CEUR-WS.org, Toulouse, France, 2015.
[48] J. Xu, W. Tian, G. Lv, S. Liu, Y. Fan, 2.5 D Facial Personality Prediction Based on Deep Learning, J. Adv. Transp. 2021 (2021.).
[49] A. Okrent, 13 Words That Changed From Negative to Positive Meanings (or Vice Versa), Mental Floss, 2019. https://www.mentalfloss.com/article/65987/13-words-changed-negative-positive-or-vice-versa.

[50] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for, Comput. Ling. 5 (2017) 135–146.
[51] F.A. Gers, J. Schmidhuber, F. Cummins, Learning to forget: Continual prediction with LSTM, Neural Comput. 12 (2000) 2451–2471.
[52] Z. Malki, E. Atlam, G. Dagnew, A.R. Alzighaibi, E. Ghada, I. Gad, Bidirectional Residual LSTM-based Human Activity Recognition, Comput. Inform. Sci. 13 (2020) 1–40.
[53] Z. Malki, E.-S. Atlam, A. Ewis, G. Dagnew, O.A. Ghoneim, A. A. Mohamed, M.M. Abdel-Daim, I. Gad, The COVID-19 pandemic: prediction study based on machine learning models, Environ. Sci. Pollut. Res. (2021) 1–11.
[54] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Netw. 18 (2005) 602–610.