Neural Computing and Applications (2023) 35:17147–17165
https://doi.org/10.1007/s00521-023-08603-z

ORIGINAL ARTICLE

Personality traits prediction model from Turkish contents with semantic structures

Muhammed Ali Kosan[1,2] · Hacer Karacan[3] · Burcu A. Urgen[4,5,6]

Received: 1 November 2022 / Accepted: 11 April 2023 / Published online: 27 April 2023

Abstract
Users' personality traits can provide different clues about them in the Internet environment. Some areas where these clues can be used are law enforcement, advertising agencies, recruitment processes, and e-commerce applications. In this study, it is aimed to create a dataset and a prediction model for predicting the personality traits of Internet users who produce Turkish content. The main contribution of the study is the personality traits dataset composed of the Turkish Twitter content. In addition, the preprocessing, vectorization, and deep learning model comparisons made in the proposed prediction system will contribute to both current usages and future studies in the relevant literature. It has been observed that the success of the Bidirectional Encoder Representations from Transformers vectorization method and the Stemming preprocessing step on the Turkish personality traits dataset is high. In the previous studies, the effects of these processes on English datasets were reported to have lower success rates. In addition, the results show that the Bidirectional Long Short-Term Memory deep learning method has a better level of success than other methods both for the Turkish dataset and English datasets.

Keywords Turkish Twitter content · Personality prediction model · Personality dataset · Preprocessing

1 Introduction

Data can be called a mine that has an increasing importance day by day. By receiving and processing the data, you can extract many valuable gems. In today's world, where many personal technologies have entered our lives and personal data have reached gigantic proportions, the importance of human-oriented analysis draws more attention. Especially social media contents contain many contents that can be used in the analysis of people's private lives and personal ideas.

A person's discourse and behavior can contain clues about his/her character profile and personality. At the same time, the content and behavior templates shared on the Internet, especially on social media platforms, are also important in providing the same inferences. At this point, the only difference is that the focus is on the user personality rather than the real personality of the person. People can act on a persona that they have created a persona in their minds to a certain extent in the use of social media platforms. The closest psychological analysis approach to the truth can only be made with a long-term

✉ Muhammed Ali Kosan
makosan@gazi.edu.tr

[1] Department of Software Engineering, Faculty of Engineering and Architecture, Mus Alparslan University, 49250 Mus, Turkey

[2] Department of Computer Science, Graduate School of Informatics, Gazi University, 06640 Ankara, Turkey

[3] Department of Computer Engineering, Faculty of Engineering, Gazi University, 06570 Ankara, Turkey

[4] Department of Psychology, Faculty of Economics, Administrative, and Social Sciences, Bilkent University, 06800 Ankara, Turkey

[5] Interdisciplinary Neuroscience Program, Graduate School of Engineering and Science, Bilkent University, 06800 Bilkent, Ankara, Turkey

[6] Aysel Sabuncu Brain Research Center and National Magnetic Resonance Research Center (UMRAM), Ankara, Turkey

clinical study. Therefore, in studies conducted in the literature, the sample and limitations of the study are stated in this way.

In the analyzes made on personal data, studies are carried out in many areas such as profile extraction, network of relationships, character analysis, and psychological trait extraction. In particular, the determination of personality traits contains clues about the next movement profile of any user, general characteristic structure, and how you can approach people. Therefore, models based on personality trait prediction are used in many areas such as advertising agencies, law enforcement, intelligence activities, competition-based activities, and human resources departments.

In this study, firstly, an overview of the studies examined in the literature is presented in Sect. 2. Next, the general explanation of the research system is presented in Sect. 3. Then, Sect. 4, the process of obtaining the collected personality traits dataset with Turkish content is detailed in the subsections. The prediction model which is applied after the final dataset creation process is completed, is presented in Sect. 5. After explaining the dataset and the estimation model, the experiments are presented in Sect. 6, and the discussion is given on the experiments. Finally, a general evaluation of the study is made in the conclusion part.

The novelties of this study are presented below:

- Comparison of different approaches in preprocessing steps of Turkish content.
- Comparison of popular approaches in vectorization step of Turkish content.
- Comparative and structural analysis in the prediction of personality traits from Turkish content.
- The importance of structural analysis of social media platform data in the prediction of personality traits from Turkish content.
- Personality traits dataset consisting of content written in Turkish.
- The use of the generalized deep learning model in the prediction of personality traits with preprocessing steps in the structural analysis of textual data.

2 Literature review

Studies on the prediction of personality traits in the literature are among the subjects that have been studied for years. In a study, which examines the studies conducted before 2017 for the meta-analysis, it is presented that IPIP, Big-Five, and Big Five Inventory 10 personality traits tests are used [1]. While Facebook social media platform is used intensively, there are Twitter, Sina Weibo, and Instagram among the social media platforms studied. Half of the studies are the ones using the myPersonality dataset. Apart from this, digital footprints of users such as activity on social media, pictures, demographics, language, and likes are used in addition to their posts.

PAN is a series of scientific events organized by WEBIS that incorporates shared tasks on digital text forensic analysis and stylometry. In a shared task conducted in 2015, a personality traits dataset created from twitter data in four different languages was used [2]. This dataset was presented separately as training and test datasets. English, Dutch, Italian, and Spanish languages were preferred for training datasets, and there were 152, 34, 38, and 100 user data, respectively. In addition, for the test datasets, there were 142, 32, 36, and 88 user data for the same languages, respectively. Based on the Big-Five scale, relevant personality traits had values between − 0.5 and 0.5. It was requested to use RMSE as a performance criterion in the estimation of personality traits, and the average RMSE value was taken as the final value. Although each group contributing to the study achieved different success rates, the best success rate was achieved by a group for three languages. The related study consisted of the combination of the second-order attributes (SOA) and Latent Semantic Analysis (LSA) [3]. Bag-of-words (BOW), SOA, and LSA methods were compared to the authors' proposed method. According to the results obtained, it was presented that the proposed method had better values.

The personality traits and user personality-based literature studies were examined in 2017 and later were scanned through Web of Science, Science Direct, and IEEE Xplore, and the same studies were eliminated and analyzed. In addition, subjects such as Internet, smartphone, technology and social media addiction, academic and student performance analysis, social evaluations based on personality traits, alcohol use, cyberbullying, plant/animal personality, eating habits, and trolling, which are not the same as our study subject in terms of subject, were eliminated. When the remaining studies are examined, it is seen that the social media platforms Facebook, Twitter, Instagram, Sina Weibo, YouTube, and Google+ are used. The studies were mostly done on the Facebook platform, and it was observed that this was due to the myPersonality dataset [4] being an accessible dataset for a while. The myPersonality dataset is still available in a reduced version with 250 user data. The list of studies and the number of studies according to social media platforms are presented in Table 1.

In the studies examined, many different preprocessing methods were used, such as punctuation cleaning, root reduction, and conversion of different languages. The statistical representation of these preprocess methods according to the frequency of use is presented in Fig. 1. In addition, the content languages used in the studies were mostly English, but other languages are Spanish, German,

Table 1 Personality traits prediction studies published in 2017 and later according to social media platforms

Social media platform	Studies	Count
Facebook	[5–25]	21
Twitter	[23, 26–36]	12
Instagram	[23, 37, 38]	3
Sina Weibo	[39–46]	8
Multi-platform	[23, 47–52]	7

Dutch, Italian, Chinese, Portuguese, French, Arabic, Malay, and Bahasa. Among the personality traits tests, Big-Five was the most used, but DISC, NEO-PI-R, HDS, MBTI, NEO-FF-60, etc., tests were also used. For prediction methods, classification, regression, and clustering methods were used in different combinations in statistics and artificial intelligence. According to the prediction method used, various test and evaluation methods (Accuracy, f-measure, MAE, RMSE, etc.) were used.

When the literature is examined as a whole, different structural analysis and preprocessing paths for each data environment give different results. In addition, each social media platform has different types of data and entities. At this point, each social media platform should be subjected to a separate structural analysis, and the preprocessing steps should be designed according to the structure of the platform. Especially in artificial intelligence models that have reached a certain level, it is more important to act based on a data-centric approach in order to increase the success rate and approach generalization [53].

In the literature, there is a shortage of studies based on the prediction of personality traits from social media platforms with Turkish content. Studies conducted with Turkish content are not based on personality trait estimation from the content. In this study, it is aimed to create a Turkish personality traits dataset for personality traits prediction from social media platforms with Turkish content and to present a generalized system design based on personality traits prediction from Turkish content.

3 Research methodology

After the literature review of the study, a systematic flow was determined to create the research methodology. This flowchart is presented in Fig. 2. In the research flow, first, it was necessary to draw the boundaries of the scope by considering its contribution to the literature. At this point, there was a lack of Turkish content personality traits

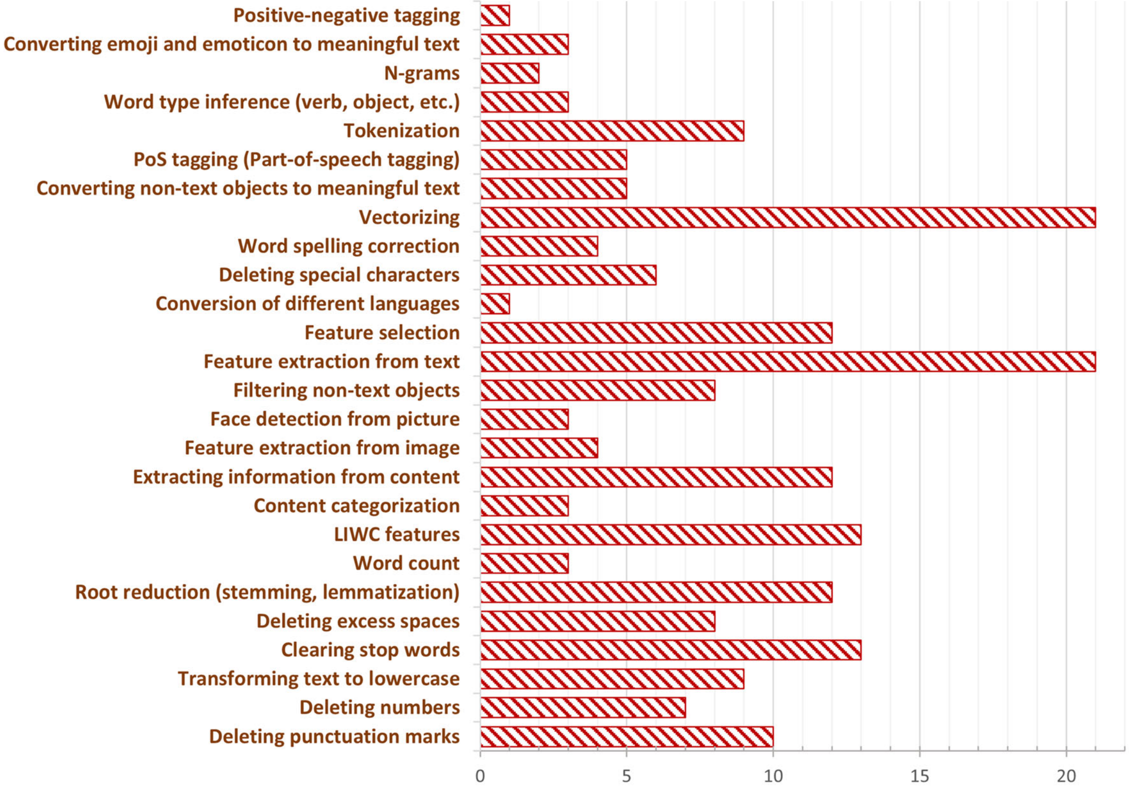
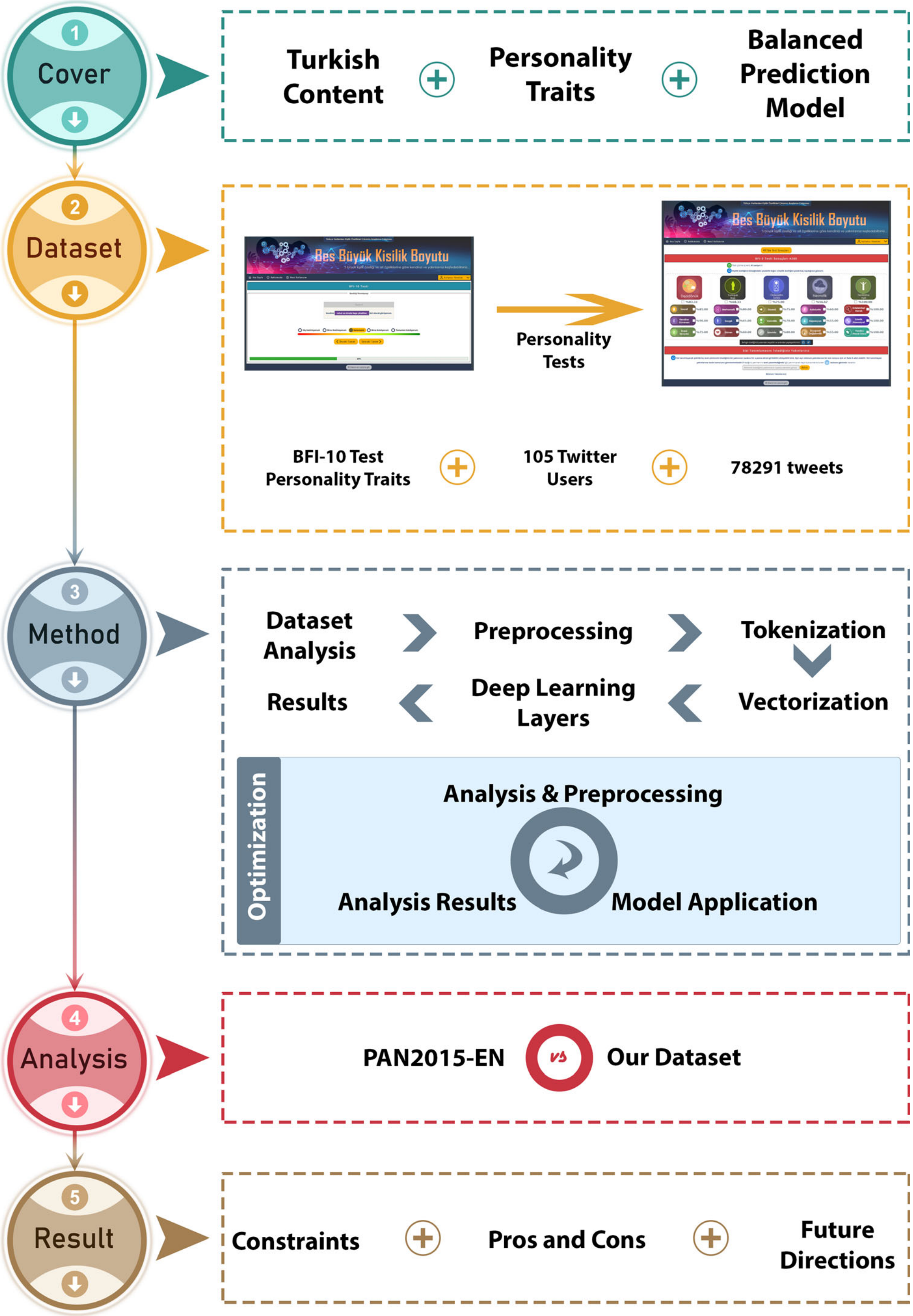


Fig. 1 Usage rates of preprocess methods in the literature

Fig. 2 Research methodology



dataset and Turkish content-based personality traits prediction model. In addition, with the thought that creating a balanced and generalizable methodology for this prediction model would contribute to the real usage areas, it was found appropriate to reveal a data-centric approach by using semantic structure analysis according to the content type. Afterwards, the permissions of the Ethics Commission were collected for the data collection process, and in parallel, a gamification-based web application was developed to fill the Personality Traits Tests by the users and to encourage them. With the release of this web application, many different methods were applied in both official and Internet environments for the promotion of the application. After the data collection process was completed, the analysis of the obtained data and the creation of a usable dataset were carried out. After obtaining the final dataset, experiments were carried out with semantic preprocessing methods as well as different deep learning methods and vectorization methods. Finally, the results were analyzed, and a comparative analysis was presented. Relevant steps are detailed in the next chapters.

4 Turkish personality traits dataset

To create a dataset of personality traits with Turkish content, first of all, personality tests to be applied to users were determined. Next, an Ethics Evaluation was made regarding the tests to be applied, and the permissions of the Ethics Commission were obtained. Then, a web application was developed based on gamification with the determined personality tests, and promotional activities were carried out throughout Turkey. After the data collection process was completed, the data obtained were analyzed, and the final dataset was created. The steps of this process are explained in the sub-headings.

4.1 Determination of personality tests

Five different personality trait tests determined based on the five personality scales are presented in Table 2. More than one personality test were chosen to understand which test the users would be interested in and to clarify the personality traits with the results of more than one test. In addition, users were given the opportunity to be evaluated by their acquaintances.

Personality tests were administered to 10 people for preliminary information purposes. And with this application, average test resolution times were obtained. The test solving times obtained were presented for preliminary information purposes in the web application where personality tests were published. Average test resolution times are 41 s for BFI-10, 53 s for TIPI, 82 s for MINI-IPIP, 112 s for BFI-2S, and 230 s for BFI-2.

4.2 Ethics evaluation

An application was made to the Gazi University Ethics Committee for data collection. The application was approved with the letter dated July 23, 2020, and numbered E.78156. An additional extension was requested as it was thought that enough records could not be reached in data collection due to the pandemic. An additional extension request was approved for 6 months with the meeting held on January 26, 2021.

Obtained personality data and the data collected from social media platforms were anonymized to prevent reverse detection. In addition, it was ensured that the dataset to be created from the relevant data would only be shared with researchers requesting it by official e-mail from research institutions (e.g., university).

Table 2 Personality tests and features

Personality test	References	Number of questions	Scale
BFI-10	[54, 55]	10	5
TIPI	[56, 57]	10	7
MINI-IPIP	[58, 59]	20	5
BFI-2S	[60, 61]	30	5
BFI-2	[60, 61]	60	5

4.3 Data collection system for personality traits

The data collection system was developed as a responsive web application that can be accessed from any device. In addition, the web application was enriched with icon sets where users can always access the tests they apply, request their relatives to evaluate themselves, and share them on social media. The main screen, test solution screen, and result screen of the web application are presented in Figs. 3, 4, and 5, respectively.

4.4 Dissemination and data collection

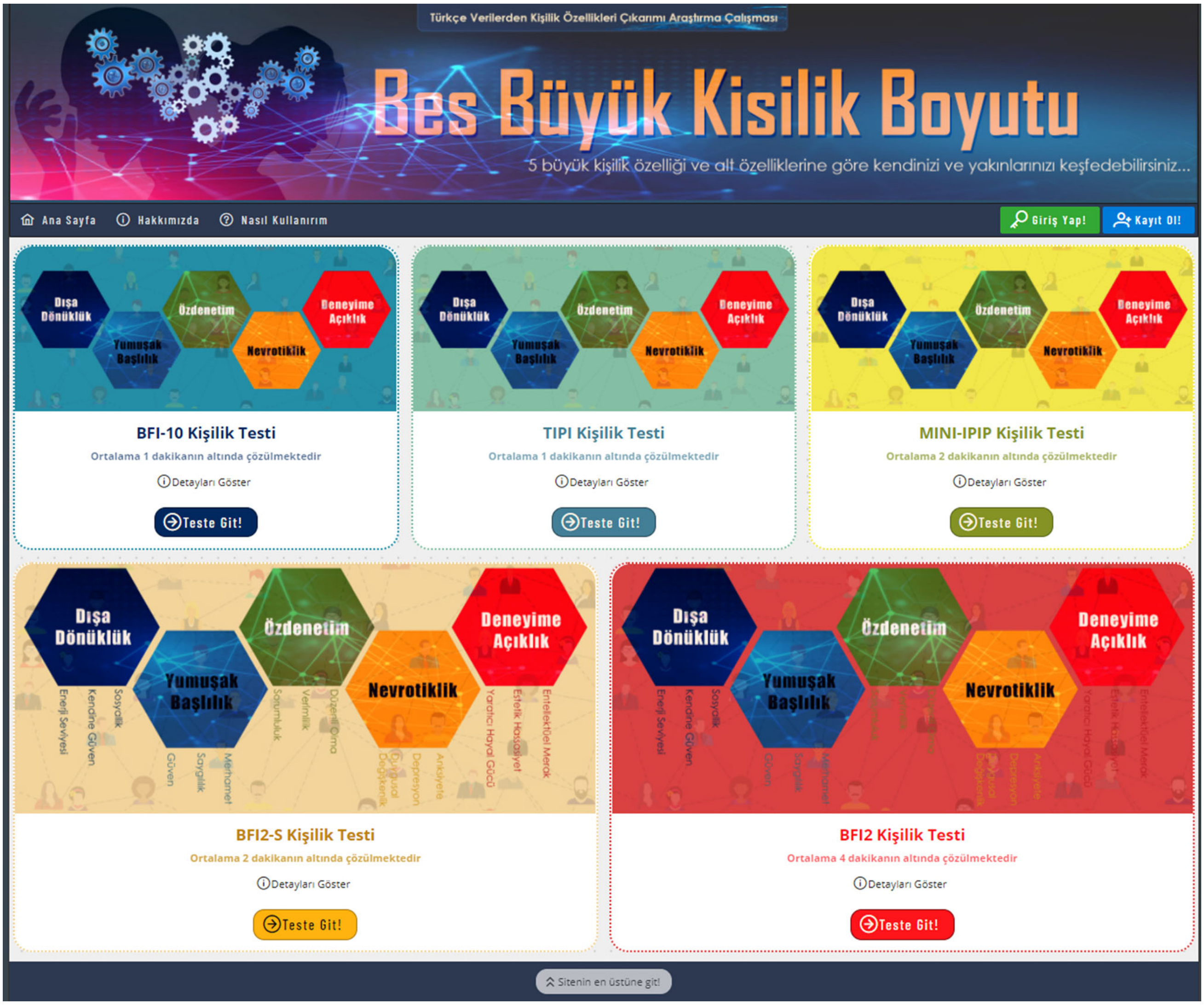
The procedures for the announcement of the web application are listed below:

- An official letter was written to all universities and ministries in Turkey.
- Notifications were made from personal and different social media accounts (Twitter, Linkedin, Facebook, WhatsApp, Instagram, etc.).
- Announcements were made to several student groups at universities.
- Promotions on mailing lists and forums.
- By creating academic and private e-mail lists, mass e-mails were sent so that they would not be included in spam lists.

The data collection process was terminated in August 2021, when the permission of the Ethics Commission expired.

4.5 Analysis of collected data

All user accounts and social media accounts in the collected data are presented in Table 3, the number of tests solved in Table 4, and the social media accounts of users who took the test, and whose data we can access (public) are presented in Table 5. The number of meaningful users who have one of the three social networks (Facebook,



Türkçe Verilerden Kişilik Özellikleri Çıkarımı Araştırma Çalışması

Beş Büyük Kişilik Boyutu

5 büyük kişilik özelliği ve alt özelliklerine göre kendinizi ve yakınlarınızı keşfedebilirsiniz...

BFI-10 Kişilik Testi

Ortalama 1 dakikanın altında çözülmektedir

TIPI Kişilik Testi

Ortalama 1 dakikanın altında çözülmektedir

MINI-IPIP Kişilik Testi

Ortalama 2 dakikanın altında çözülmektedir

BFI2-S Kişilik Testi

Ortalama 2 dakikanın altında çözülmektedir

BFI2 Kişilik Testi

Ortalama 4 dakikanın altında çözülmektedir

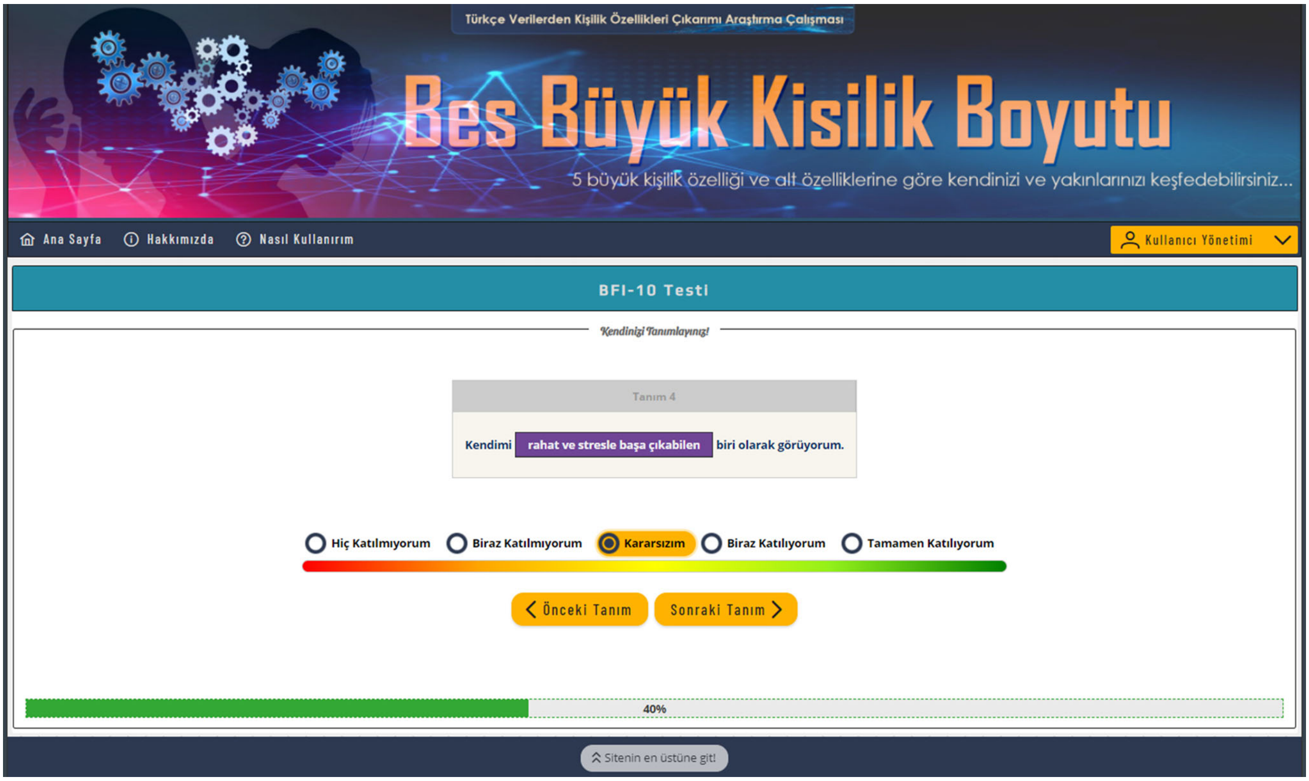
Fig. 3 Web application developed to collect data—home screen

Twitter, and Instagram) and solve any personality test is 259.

When the accounts in the data presented in Table 5 were examined, it was seen that the shares on Instagram were both at a low level, and the information such as the explanation for the photos was concentrated on the tags. Posts on Facebook were usually about sharing the posts of other groups and updating photos and information. In the analysis made on the textual data, the tweet information of 113 users from 123 users was collected by evaluating the test solutions made for Twitter. A total of 105 of these 113 users solved the BFI-10 test, but the remaining eight completed one of the other tests. Since a balanced distribution could not be observed at this point, it was thought that it would be more appropriate to evaluate all solutions as the product of the same test. Thus, a dataset consisting of 105 users who solved the BFI-10 test and had a meaningful twitter account was created.

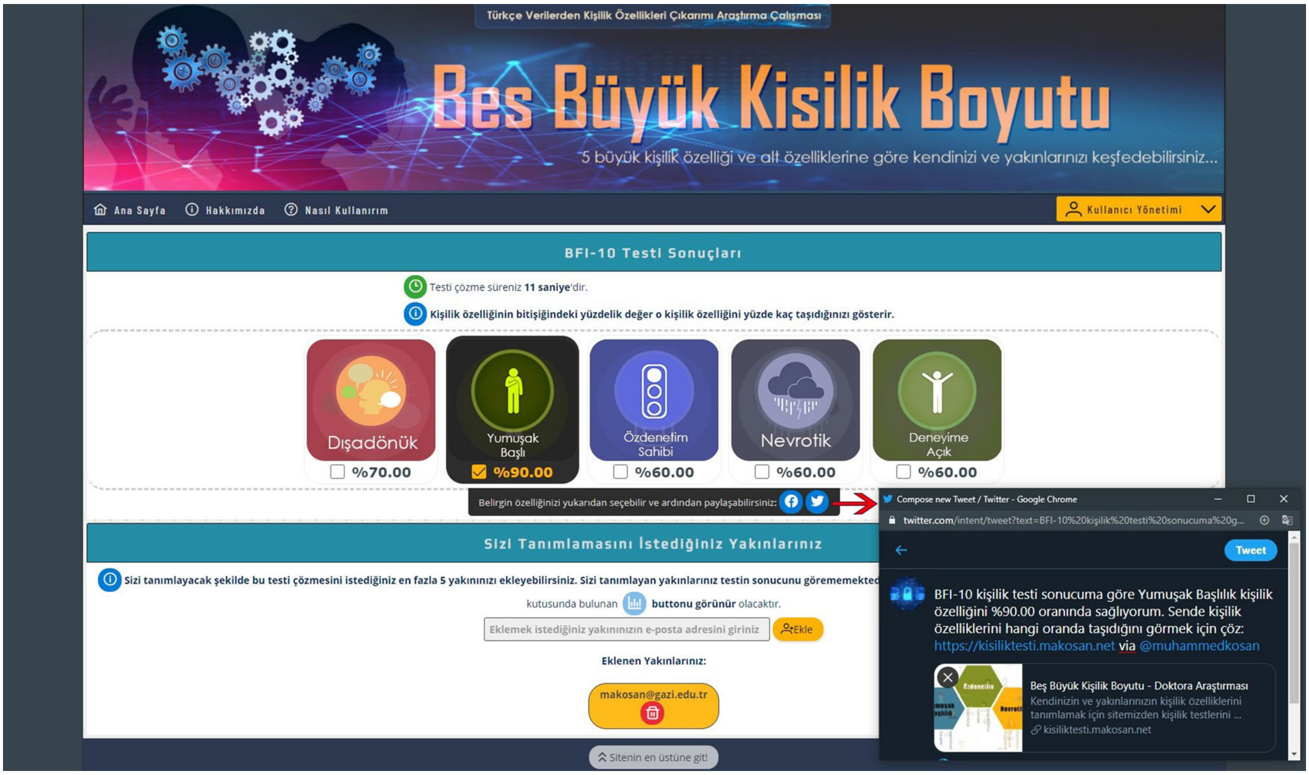
4.6 Creating the final dataset

In the examinations made on the collected user information, it was observed that Twitter users shared posts in different languages. In addition, it was seen that a tweet consisted of a mixture of more than one language or only shared with hashtag, mention, URL, etc. Such posts were marked as Undefined (Und) content language by Twitter. Dataset statistics according to tweet information marked as Turkish (Tr) and/or Und are presented in Table 6 and Fig. 6. Considering the information obtained, the data distribution was close to each other according to the different tweet numbers.



Kendimi rahat ve stresle başa çıkabilen biri olarak görüyorum.

Fig. 4 Web application developed to collect data—personality test screen


Türkçe Verilerden Kişilik Özellikleri Çıkarımı Araştırma Çalışması
Bes Büyük Kisilik Boyutu
5 büyük kişilik özelliği ve alt özelliklerine göre kendinizi ve yakınlarınızı keşfedebilirsiniz...
Ana Sayfa
Hakkımızda
Nasıl Kullanırım
Kullanıcı Yönetimi
BFI-10 Testi Sonuçları
Dışadönük
□ %70.00
Yumuşak Başlı
☑ %90.00
Özdenetim Sahibi
□ %60.00
Nevrotik
□ %60.00
Deneyime Açık
□ %60.00
Sizi Tanımlamasını İstediğiniz Yakınlarınız
Eklemek istediğiniz yakınınızın e-posta adresini giriniz
Ekle
Eklenen Yakınlarınız:
makosan@gazi.edu.tr
Sitenin en üstüne git!
Compose new Tweet / Twitter - Google Chrome
twitter.com/intent/tweet?text=BFI-10%20kişilik%20testi%20sonucuma%20g...
Tweet
Özdenetim
Yumuşak Başlılık
Nevrotik

Testi çözme süreniz **11 saniye**'dir.

Kişilik özelliğinin bitişiğindeki yüzdelik değer o kişilik özelliğini yüzde kaç taşıdığınızı gösterir.

Belirgin özelliğinizi yukarıdan seçebilir ve ardından paylaşabilirsiniz:

Sizi tanımlayacak şekilde bu testi çözmesini istediğiniz en fazla 5 yakınınızı ekleyebilirsiniz. Sizi tanımlayan yakınlarınız testin sonucunu görememekted
kutusunda bulunan **buttonu görünür** olacaktır.

BFI-10 kişilik testi sonucuma göre Yumuşak Başlılık kişilik özelliğini %90.00 oranında sağlıyorum. Sende kişilik özelliklerini hangi oranda taşıdığını görmek için çöz: https://kisiliktesti.makosan.net via @muhammedkosan

Beş Büyük Kişilik Boyutu - Doktora Araştırması
Kendinizin ve yakınlarınızın kişilik özelliklerini tanımlamak için sitemizden kişilik testlerini ...
kisiliktesti.makosan.net

Fig. 5 Web application developed to collect data—personality test result screen

Table 3 Data collected—account information

Parameter	Value
Number of total users	899
Number of users who solved any test	813
Number of users who did not solve any tests	86
Number of Twitter accounts	253
Number of Facebook accounts	255
Number of Instagram accounts	449
Number of YouTube accounts	4
Number of Blog accounts	4

Table 4 Data collected—numbers of tests solved according to personality tests

BFI-10	TIPI	MINI-IPI	BFI-2	BFI-2S
762	407	370	276	274

Table 5 Data collected—numbers of social media accounts of users who solved personality tests according to social media platform

Twitter	Facebook	Instagram
123	77	59

When the tweets collected from 105 Twitter users were analyzed; there were 78,291 tweets in total. However, it contained data in the number of 1,104,699 words/entities (not unique). When Turkish and Undefined tweets were considered, the information presented in Table 7 was obtained. All non-Tr and non-Und tweet data were cleared from the dataset.

The final dataset created was used as Tr and Und. Each personality trait value (PTV) was between 0 and 1. The tweet distributions of each personality trait in our dataset according to PTV ≤ 0.5 (less than or equal to) and $0.5 < \text{PTV}$ (greater than) conditions are presented in Table 8. A general balance was observed in the distribution. But only the agreeableness (A) personality trait showed a tweet density of 0.5 and above. This was considered as something to be ignored. Abbreviations for personality traits and arithmetic mean used in the tables are Extraversion (E), Neuroticism (N), Agreeableness (A), Conscientiousness (C), Openness (O), and Arithmetic Mean (A.M.), respectively.

To see the word and entity density of the dataset, different distributions were observed by applying the cleaning process according to Zemberek and NLTK Stop Words along with the standard preprocessing steps on the dataset. These distributions are presented in Fig. 7 for NLTK and Fig. 8 for Zemberek. When examined together in Figs. 7 and 8, different word densities occurred according to the Stop Words lists. In addition, it was observed that presence/word densities were distinctive in some cases according to personality traits.

5 Prediction model

After the dataset was obtained, a template was presented to create the most appropriate deep learning model with combinations of different approaches over a standard systematic scheme for the deep learning prediction model. This template is presented in Fig. 9 in general terms. The details of the template were explained in the sub-headings.

5.1 Preprocessing

Based on the preprocessing experiments on English datasets, the following preprocessing steps were applied in the estimation model as standard. These preprocessing steps [32] are as follows:

- Reply to tag (RPT).
- Retweet to tag (RTT).
- Image to tag (IT).
- URL to tag (UT).
- Mention to tag (MT).
- Email to tag (ET).

Table 6 Number of users based on count of tweets according to Turkish (Tr) and Undefined (Und) tweets

Tweets count range	Count of users (Tr)	Count of users (Tr and Und)
0–100	24	21
101–200	10	10
201–500	28	29
501–1000	15	15
1001–2000	11	12
2001–3200	17	18
Total	105	105

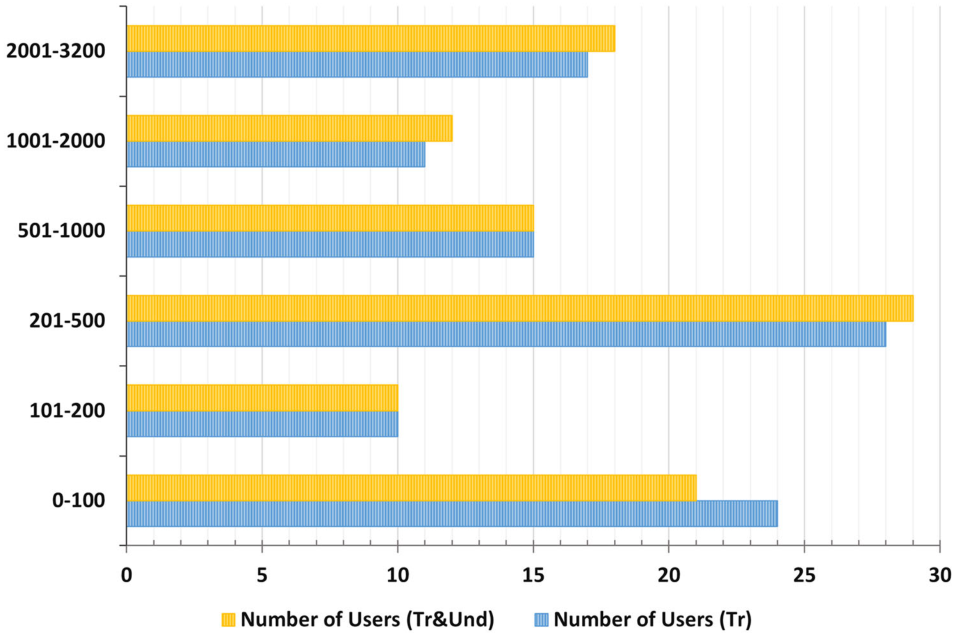


Fig. 6 Number of users based on count of tweets according to Turkish (Tr) and Undefined (Und) tweets

Table 7 Count of tweets and words/entities by Turkish and Undefined tweets

	Turkish tweets (Tr)	Turkish and/or Undefined tweets (Tr and Und)
Count of tweets	73,896	78,291
Average count of tweets by users	798	840
Number of words/entity of tweets written in the selected language	1,091,745	1,104,699
User-based average of the number of words/entity of tweets written in the selected language	13,027	13,167

Table 8 Distribution of the count of tweets according to PTV ≤ 0.5 and $0.5 <$ PTV conditions of personality traits

Condition	E	N	A	C	O
$0.5 <$ PTV	43,747	39,766	81,298	64,357	51,255
PTV ≤ 0.5	47,458	51,439	9907	26,848	39,950

- Hashtag to tag (HT).
- Newline to tag (NT).
- Add space between emoji and emoticon (BEE).
- Emoji to tag (EJT).
- Emoticon to tag (ETT).
- Clear to space chars (CSC).
- Clear to quotes (CQ).
- Clear to numbers (CN).
- Clear to punctuations (CP).
- Delete multiple space (DMS).

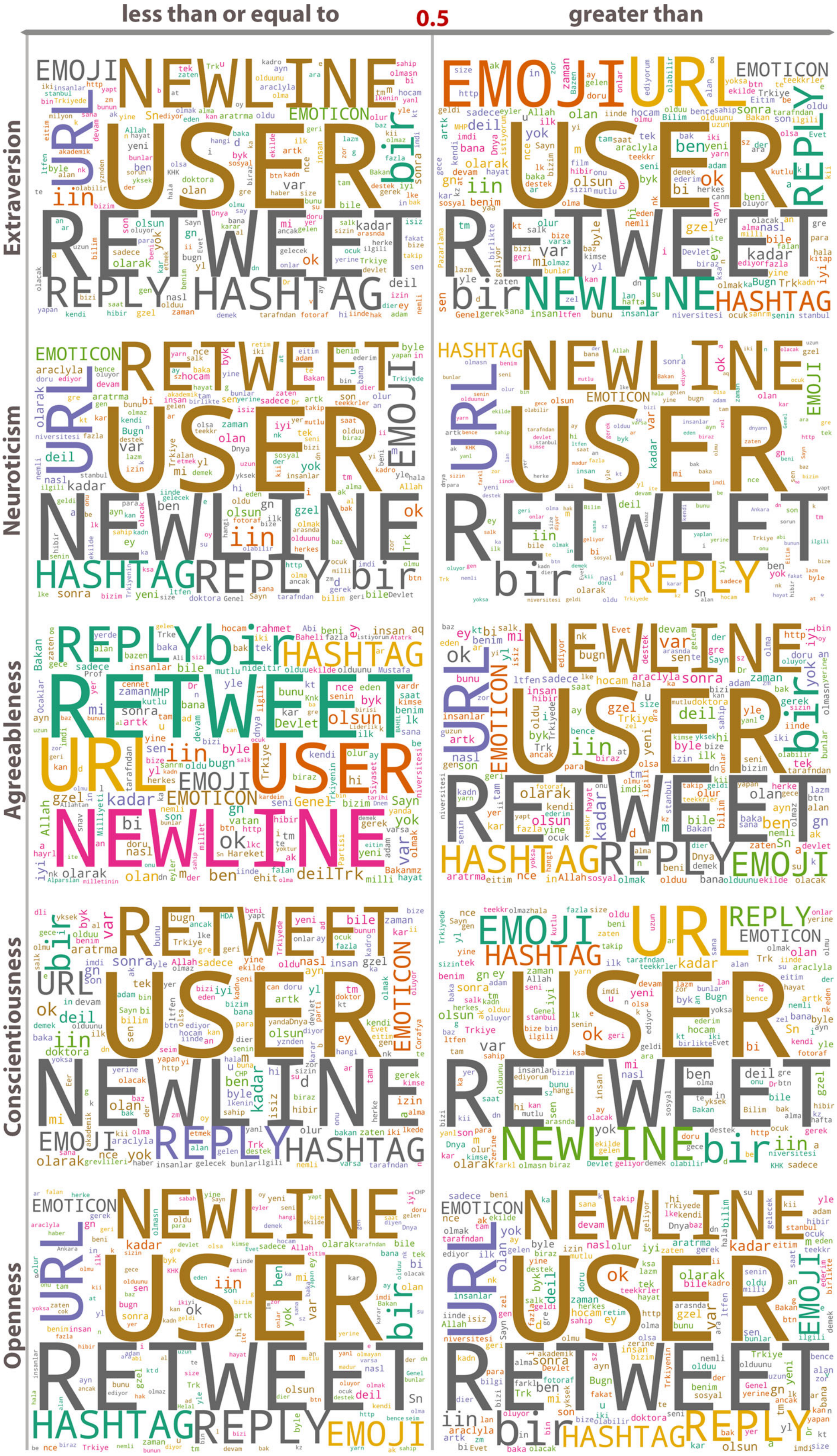
Apart from this, the preprocessing steps, which were tried with language-specific methods due to language differences, are listed below:

- Normalization (NRM).
- PoS tagging (POS).
- NLTK—Clear Stop Words (NCS).
- Zemberek—Clear Stop Words (ZCS).
- Lemmatization (LEM).
- Stemming (STM).

The cleaning of unimportant words was subjected to different trials according to both NLTK and Zemberek Stop Word lists. The description list of the preprocessing methods is presented in Table 9.

The presented preprocessing steps can create some problems when used together. For example, since images are included in the content as links, operations related to images should be applied before URL filtering processes. The flow of preprocessing steps created with this logic, and which can be used optionally, is presented in Fig. 10.

Fig. 7 Word/entity density maps created using NLTK Stop Words based on personality traits



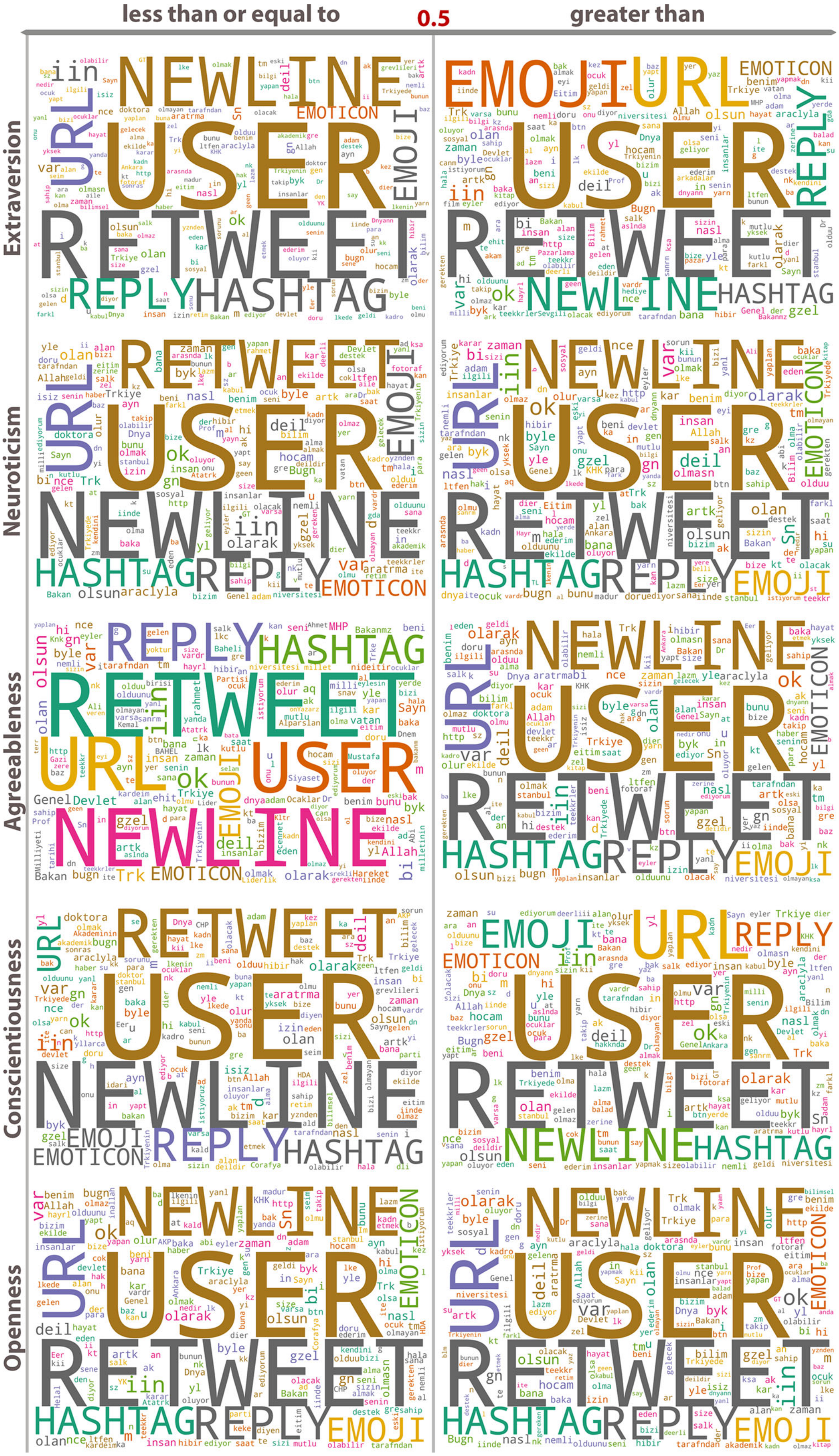


Fig. 8 Word/entity density maps created using Zemberek Stop Words based on personality traits

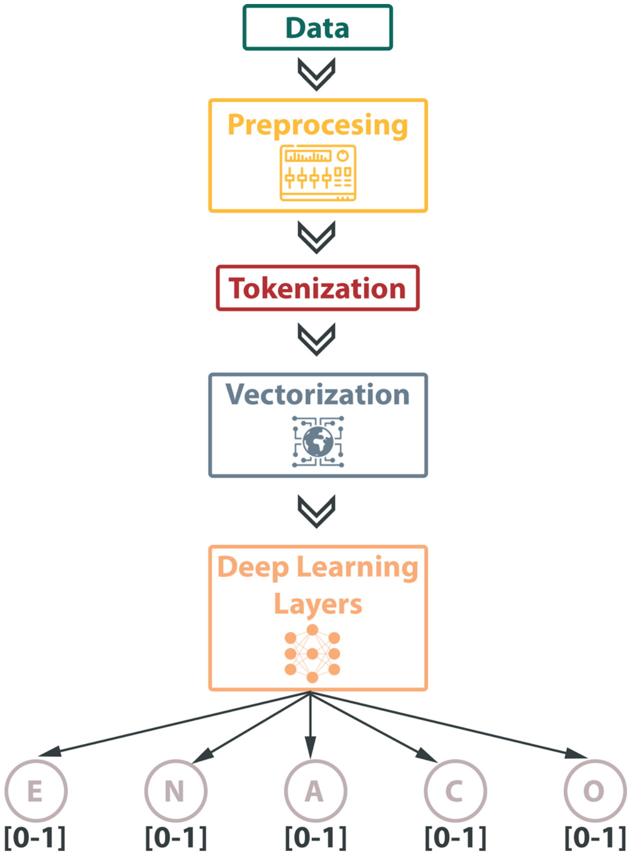


Fig. 9 Deep learning model

5.2 Tokenization and vectorization

Every textual expression must be converted to a numeric expression to be processed mathematically. In this study, the textual expression was first divided into distinctive expressions by applying the white space-based tokenization method (word-based tokenization). Then, with certain digitization methods, parts of the data were converted into numerical expressions to be distinguished from each other. In this study, BERTurk (Turkish BERT) [62], Electra [62, 63], FastText [64, 65], Doc2Vec [66, 67], and Entity Frequency vectorization methods were applied together with the basic vectorization method using the dictionary created from the training dataset.

The Entity Frequency vectorization method was used as a vector method created according to the size of common entities that can be found in a tweet. Our aim in trying this method was only to understand whether the influence levels of entities on tweets contribute to the success rate of the personality prediction model. For the Entity Frequency vectorization method, the entities of Reply, Retweet, User, Hashtag, Url, Image, Email, Newline, Emoji, and Emoticon formed a vector, respectively. The vector creation process was completed according to which frequency level an entity was on the content.

5.3 Deep learning methods

Bidirectional models, which train the vectors taken as input by bidirectional calculation with better results, showed better success rate than unidirectional models. For this reason, in this study, bidirectional uses of RNN [68], GRU [69], LSTM [69–71], and 1D Convolutional LSTM [69, 72] architectures were studied. The representation of the variations according to the deep learning architecture and deep learning layers used in this study is presented in Fig. 11, and the parameter values of the methods used are presented in Table 10.

In the studies of creating a deep learning model, the Bidirectional Long Short-Term Memory (Bi-LSTM) model [32], which was first applied on the English dataset, was applied. Hyperparameters were optimized to increase the success rate. While the models were optimized, the number of epochs, the number of units of the layers, the batch size, the number of different hidden layers, the activation functions (sigmoid, tanh, swish, etc.), the learning rate, the dropout (normal or recurrent), l1/l2 regularization, etc., parameters were used. Dropout, regularization, and min-delta change control processes were carried out to make learning in a balanced way (to avoid overfitting and underfitting). Swish activation function [73] is used in the hidden layer. Trials and previous studies had an impact on this selection. In addition, root-mean-squared error (RMSE) was used as evaluation metric, mean-squared error (MSE) was used as loss function, and Adam was used as optimization function.

5.4 Working environment

To express the working environment as hardware infrastructure;

- AMD Ryzen 9 3900X 12-Core Processor.
- 64 GB (4 × 16 GB) 3600-MHz RAM.
- 500-GB SSD (3400 read/2500 write).
- Nvidia Chipset RTX 2080Ti 11-GB GPU.

For an efficient communication between hardware, the balance of reading, writing, and transmission between hardware should be observed.

In the software-based working environment, Tensorflow 2.4, Cuda 10.2, and Python 3.8.3 are used on the Windows 10 operating system. In addition to basic Python operations, NLTK 3.5, Gensim 3.8.3, and Zemberek-grpc 0.16.1 libraries were used for preprocessing steps and vectorization approaches.

Table 9 Preprocessing methods (PPM)—description list

PPM	Description
RPP	If a post is a reply to another post, the [REPLY] tag is added
RTT	If a post is retweeted from another post, the [RETWEET] tag is added
IT	If a post contains an image, the [IMAGE] tag is added where the image entity is in the post
UT	If a post contains a URL, the [URL] tag is added where the URL entity is in the post
MT	If a post refers to another account (e.g., @user and mention), the [USER] tag is added where the mention entity is in the post
ET	If a post contains an e-mail address, the [EMAIL] tag is added where the e-mail address entity is in the post
HT	If a post contains a hashtag (e.g., #hashtag), the [HASHTAG] tag is added where the hashtag is in the post
NT	If a post contains a new line (\n) information, the [NEWLINE] tag is added where the new line information is in the post
BEE	By creating spaces between emoji and emoticons, operations on emoji and emoticons are facilitated. It is also used as a facilitation step for EJT and ETT
EJT	If a post contains an emoji, the [EMOJI] tag is added where the emoji is in the post
ETT	If a post contains an emoticon, the [EMOTICON] tag is added where the emoticon is in the post
CSC	If a post contains a special space definition (\t \n \r \f \v) rather than the basic space operation, the basic space character is added where the special space definition is in the post
CQ	If a post contains quotation marks, the basic space character is added where the quotation marks in the post. Since quotation marks can be adjacent to other characters, the base space character is added first. In this way, the filtering process is carried out by eliminating possible problems. Afterwards, excess spaces are cleaned with DMS
CN	If a post contains digits, the digits are cleared
CP	If a post contains punctuation marks, the punctuation marks are cleared
DMS	For the tokenization process to be done correctly, the excess space characters both in the normal content and created by different preprocessing steps are cleaned
NRM	If there are misspelled words in a post, the correct spellings of the words are added
POS	A new sentence structure is created by adding the sentence element of each word to the word in the post
NCS	Using the Turkish Stop Words list in the NLTK library, the unimportant words in the sentence are cleaned
ZCS	Using the Turkish Stop Words list in the Zemberek library, the unimportant words in the sentence are cleaned
LEM	By finding the roots of the words, a new sentence is formed with the roots
STM	A new sentence is formed with the new words obtained by clearing the prefixes and suffixes of the words

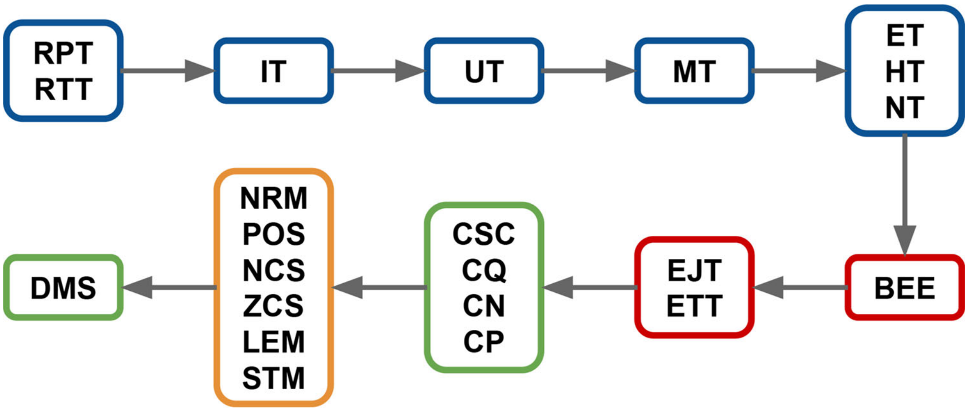


Fig. 10 Preprocessing steps—display of the sequence of operations to be followed in combinations

5.5 Other datasets

Since no personality traits dataset with Turkish content could be found to compare the dataset created in this study, the PAN-2015-EN [3] and 5081-EN [32] datasets used in the literature were used for comparison purposes. Considering the language structure, the prediction success rates will not be close due to the preprocessing steps applied and the different vectors that will emerge in the vectorization. For this reason, a comparison of language-based operations will not be made.

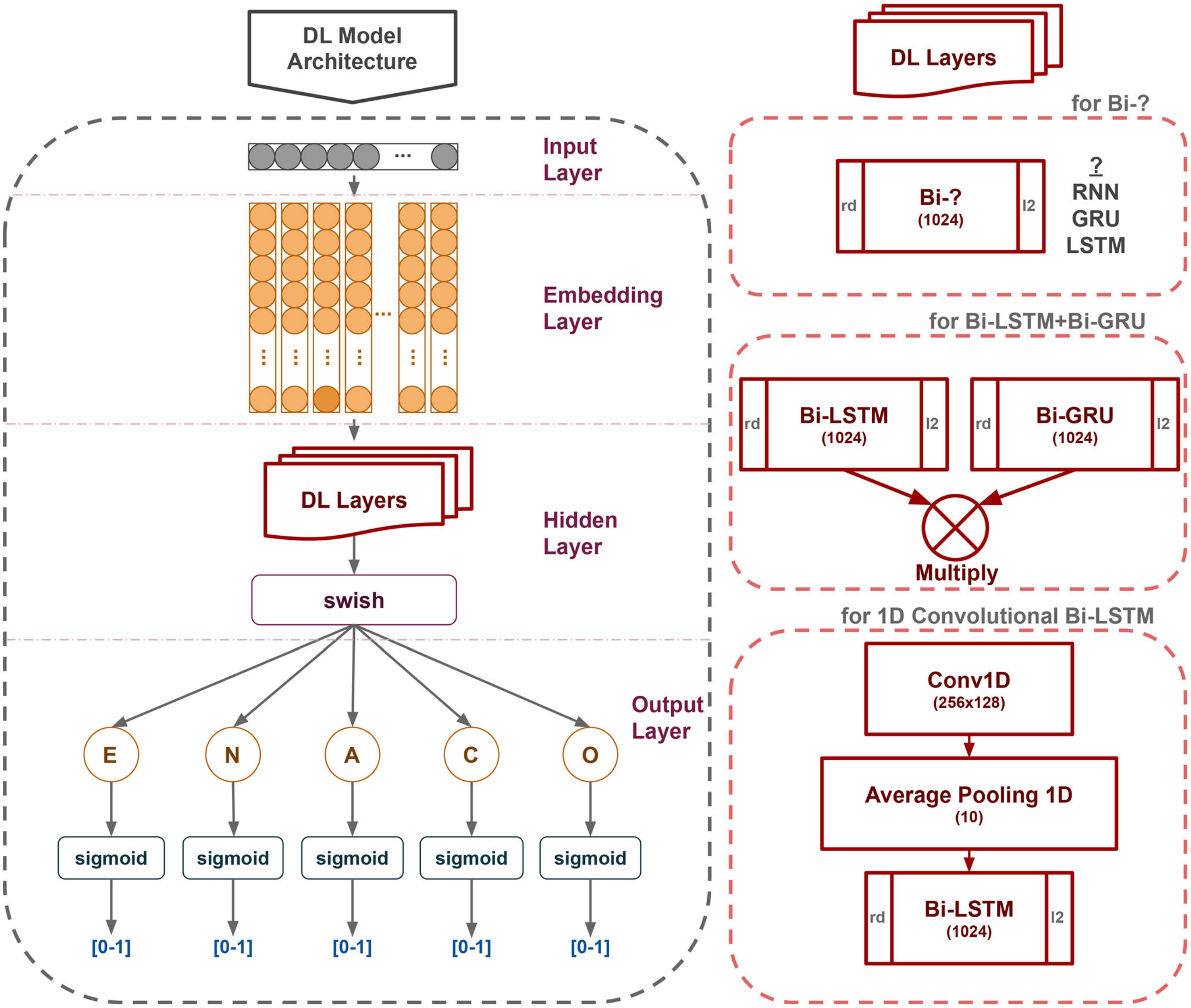


Fig. 11 Proposed deep learning architecture and deep learning hidden layers

6 Experiment and discussion

The deep learning model comparing PAN-2015-EN and 5081-EN datasets was first applied on our dataset (105-TR) with Turkish content, and the results are presented in Table 11. When the RMSE values were examined, it was seen that the optimized model for English language did not have the same effect on Turkish content, based on language differences.

Based on the difference in language structure and usage diversity, a hyperparameter estimation parameter was selected on the model, and other parameters were kept constant. Optimization was done by changing the selected parameter. Afterwards, all parameters were selected in order, and the same operations were applied. In addition, experiments on model complexity and hybrid solutions were also conducted. The results obtained with the Model_TR_OP obtained by the hyperparameter tuning made with the Model_EN_to_TR are presented in Table 12. The Model_EN_to_TR was obtained by applying our Turkish dataset with the deep learning model architecture created using the English dataset. The Model_TR_OP, on the other hand, represents the new results obtained on the Turkish dataset with hyperparameter tuning and with the adjustments made in the model structure.

According to the language structure used in the created dataset and the vector structures created, there can be differences in the success rates of the models. For this reason, firstly, different model trials according to the deep learning architecture were applied on the 105-TR dataset, and the most suitable deep learning model was tried to be found. Comparative results according to the deep learning model are presented in Table 13. The values marked in bold in

Table 10 Model parameters

Parameter	Value
Sequence length	32
Embedding dimension	256
Batch size	128
Epoch	Max: 1000 and if the minimum delta is 0.0001 by validation loss, the process is finished
Conv1D size	256 × 128 (filter size × kernel size)
Pool size	10
Hidden layer—activation function	Swish
Hidden layer—units	1024
l2 regularizer	0.001
Recurrent dropout rate	%30 (1D Convolutional Bi-LSTM), %90 (anothers)
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 11 Comparison of datasets according to the created model (using RMSE)

Dataset	E	N	A	C	O	A.M.
5081-EN	0.1434	0.1594	0.1897	0.2109	0.1373	0.1681
PAN-2015	0.1590	0.2262	0.1509	0.1470	0.1605	0.1687
105-TR	0.3153	0.2660	0.1693	0.2132	0.2333	0.2394

the tables represent the best values. The bold areas presented only in the arithmetic mean column (A.M.) indicate that at least 1 of the personality traits is best predicted.

According to the results obtained, the average success rate was the same in Bi-GRU and Bi-LSTM 1024-unit models, but the best success rates were observed for three different personality traits in the results obtained with Bi-LSTM. At this point, considering the overall success rate and discrete success rate in model selection, the use of the Bi-LSTM model seemed more appropriate.

In addition to the experiments such as deep learning model, hybrid approaches, and hyperparameter tuning, some additions were made to the preprocessing steps in the deep learning model template. In addition to the normalization process on the relevant tweets for the effect of misspellings on the success rate of the model, Part-of-Speech Tagging, Stemming, Lemmatization, and Clear Stop Words operations were applied to examine the effect on model performance. The results obtained with NLP operations on word/sentence structure and Bi-LSTM model are presented in Table 14, provided that the basic preprocessing remains constant.

According to Table 14, the success rate when no word/sentence preprocessing was close to the success rate when it was applied with Clear Stop Words and Stemming

Table 12 Comparison of first model and last model (using RMSE)

Model	E	N	A	C	O	A.M.
Model_EN_to_TR	0.3153	0.2660	0.1693	0.2132	0.2333	0.2394
Model_TR_OP	0.2972	0.2495	0.1590	0.1995	0.2135	0.2237

Table 13 Comparison of results obtained with different deep learning models in fixed preprocessing and vectorization (using RMSE)

Model	E	N	A	C	O	A.M.
Bi-RNN	0.3358	0.2684	0.3316	0.2773	0.2829	0.2992
Bi-GRU	**0.2933**	**0.2490**	0.1607	0.2000	0.2154	**0.2237**
Bi-LSTM	0.2972	0.2495	**0.1590**	**0.1995**	**0.2135**	**0.2237**
Bi-LSTM + Bi-GRU	0.3019	0.2601	0.1615	0.2028	0.2202	0.2293
Convolutional Bi-LSTM	0.3168	0.2628	0.1641	0.2085	0.2209	0.2346

Table 14 Comparison of results obtained with word/sentence operations applied according to the current model (using RMSE)

Word/sentence NLP method	E	N	A	C	O	A.M
None	0.2972	0.2495	**0.1590**	0.1995	**0.2135**	**0.2237**
Normalization	0.2959	0.2559	0.1610	0.2032	0.2181	0.2266
PoS tagging	0.2957	0.2521	0.1604	0.1993	0.2178	0.2251
NLTK—Clear Stop Words	0.2982	**0.2478**	0.1606	0.1992	0.2167	**0.2245**
Zemberek—Clear Stop Words	**0.2922**	0.2509	0.1606	0.1995	0.2157	**0.2239**
Lemmatization	0.2965	0.2540	0.1619	0.2023	0.2200	0.2269
Stemming	0.2953	0.2482	**0.1591**	**0.1981**	0.2140	**0.2229**

operations. In cases where we wanted to minimize word pools, dictionary reduction could be done by applying Clear Stop Words and Stemming operations.

Finally, different approaches were used in vectorization processes, provided that the Bi-LSTM model remained constant. Experiments with these approaches are presented in Table 15. The vectorization step was applied according to the training dataset we created, including the models that were ready.

Table 15 shows that the success rate of the vectorization process was high from the word frequencies created with the dictionary created from the training dataset, which we call the standard vectorization method. In addition, it was observed that the success rates were high in different personality traits in the experiments with the models that had 32 K and 128 K dictionaries belonging to BERTurk. Based on generalization, it was thought that better results would be obtained in the models created with BERTurk as a higher capacity dictionary was operated. In addition, based on model performance, it was seen that standard vectorization could operate with less processing capacity due to fewer dictionaries and less scale at dictionary frequencies.

7 Conclusion

To predict the personality traits of users from content written in any language, it is necessary to analyze the content and create a model architecture by considering the structure of the language. Numerical identifiers that will occur due to structural differences between languages will also have different characteristics. For this reason, it may be necessary to consider a separate model architecture for each language.

In this study, for the estimation of personality traits from Turkish content, first, common social media platforms were considered, but Twitter contents and personality traits were used with the final dataset. By applying the successful model performed on English datasets on the dataset, we created from Turkish content, we observed the effect of structural differences between languages on the estimation model. In addition, we observed what kind of results we could achieve with different preprocessing steps. Afterwards, the success rate effects of different deep learning architectures on the dataset were examined, and the effect of Bi-LSTM on performance was revealed once again. Finally, the different methods were applied in the vectorization step, and it was observed that the standard and BERTurk methods showed a good success rate.

Each architecture and intermediate operations may differ according to different datasets and platform-based data. The dataset created and used in this study was based on the tweet structure created on Twitter. Tweets had a certain character limit (280 characters), and the entities in their content were varied. Likewise, it is known that there may be different entities on different platforms. For this reason, it is important to restructure each model to be created according to the platform on which it will be based.

In our experiments, as the size of the data collected is increased in a meaningful and balanced way, it is seen that the success rate increases. One of the main limitations of this study is the cost of creating a huge personality trait

Table 15 Comparison of results obtained with different vectorization processes used fixed preprocessing and Bi-LSTM model (using RMSE)

Vectorization method	E	N	A	C	O	A.M.
Standard	0.2972	0.2495	0.1590	**0.1995**	**0.2135**	**0.2237**
BERTurk—32 K uncased	**0.2922**	0.2498	0.1603	0.2028	0.2149	**0.2240**
BERTurk—128 K uncased	0.2946	**0.2485**	**0.1578**	0.2009	0.2143	**0.2232**
Doc2Vec	0.3202	0.2677	0.1690	0.2182	0.2348	0.2420
Electra	0.2958	0.2529	0.1607	0.2008	0.2171	0.2254
Entity Frequency	0.3047	0.2629	0.1653	0.2069	0.2322	0.2344
FastText—wiki.tr	0.3284	0.2681	0.1702	0.2249	0.2398	0.2463
FastText—cc.tr.300	0.3659	0.2758	0.1709	0.2403	0.2415	0.2589

dataset from meaningful content texts that can be used with personality tests. Although a sufficient level of dataset is created, it is an undeniable fact that creating a balanced dataset with higher dimensions has an impact on performance and generalization. Another limitation is the working environment. Large-scale cloud services were not used because all processes were managed with specific configurations in the hardware operating environment, and sensitive data were worked on. Another limitation can be presented as the inability to test on much more complex models due to hardware constraints in our personal working environment.

With this study, it was aimed to contribute to all areas where a preliminary assessment was needed about people/ users in public and private institutions, in addition to academic studies in this field. In the future, different model applications will be conducted on hybrid approaches of different prediction models that have been developed or are being developed. In addition, different data processing models templates will be created on data templates in different social media platforms using a data-centric approach. Furthermore, the effect of different tokenization methods, which vary according to language, will be examined on personality trait prediction models. Finally, the entities (image, web page, etc.) on social media platforms will be examined in depth, and the effects on success rate will be observed with the changes made in the content.

Authors' contributions MAK and HK helped in conceptualization; MAK worked in methodology; MAK worked in software; MAK, HK, and BAU helped in validation; MAK worked in investigation; MAK worked in resources; MAK and HK helped in data curation; MAK helped in writing—original draft; HK and BAU helped in formal analysis; MAKosan helped in visualization; HK and BAU helped in writing—review and editing; HK and BAU worked in supervision; and HK and BAU worked in project administration.

Funding This research received no external funding.

Data availibility statement The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declaration

Conflict of interest All authors declare that there is no conflict of interest.

References

1. Azucar D, Marengo D, Settanni M (2018) Predicting the Big 5 personality traits from digital footprints on social media: a meta-analysis. Personal Individ Differ 124:150–159
2. Anonymous PAN Shared Tasks. In: Webis. https://pan.webis.de/
3. Rangel F, Celli F, Rosso P et al (2015) Overview of the 3rd author profiling task at PAN 2015. In: Cappellato L, Ferro N, Jones G, Juan ES (eds) CLEF 2015 evaluation labs and workshop—working notes papers. CEUR-WS.org, Toulouse, France
4. Kosinski M, Stillwell D, Graepel T (2013) Private traits and attributes are predictable from digital records of human behavior. Proc Natl Acad Sci 110:5802–5805
5. Ahmad Z, Lutfi SL, Kushan AL et al (2017) Personality prediction of Malaysian Facebook users: cultural preferences and features variation. Adv Sci Lett 23:7900–7903
6. Laleh A, Shahram R (2017) Analyzing facebook activities for personality recognition. In: 2017 16th IEEE international conference on machine learning and applications (ICMLA), pp 960–964
7. Tandera T, Hendro SD et al (2017) Personality prediction system from Facebook users. Procedia Comput Sci 116:604–611
8. Vaidhya M, Shrestha B, Sainju B et al (2017) Personality traits analysis from Facebook data. In: 2017 21st international computer science and engineering conference (ICSEC), pp 1–5
9. Akhtar R, Winsborough D, Ort U et al (2018) Detecting the dark side of personality using social media status updates. Pers Individ Differ 132:90–97
10. Hassanein M, Hussein W, Rady S et al (2018) Predicting Personality traits from social media using text semantics. In: 2018 13th international conference on computer engineering and systems (ICCES), pp 184–189
11. Howlader P, Pal KK, Cuzzocrea A et al (2018) Predicting Facebook-users' personality based on status and linguistic features via flexible regression analysis techniques. Assoc Computing Machinery, New York
12. Mao Y, Zhang D, Wu C et al (2018) Feature analysis and optimisation for computational personality recognition. In: 2018 IEEE 4th international conference on computer and communications (ICCC), pp 2410–2414
13. Tadesse MM, Lin H, Xu B et al (2018) Personality predictions based on user behavior on the Facebook social media platform. IEEE Access 6:61959–61969
14. Xue D, Wu LF, Hong Z et al (2018) Deep learning-based personality recognition from text posts of online social networks. Appl Intell 48:4232–4246
15. Marouf AA, Hasan MK, Mahmud H (2019) Identifying neuroticism from user generated content of social media based on psycholinguistic cues. In: 2019 international conference on electrical, computer and communication engineering (ECCE), pp 1–5
16. Zheng HC, Wu CH, Assoc Comp M (2019) Predicting personality using Facebook status based on semi-supervised learning. Assoc Computing Machinery, New York
17. Al Marouf A, Hasan MK, Mahmud H (2020) Comparative analysis of feature selection algorithms for computational personality prediction from social media. IEEE Trans Comput Soc Syst 7:587–599
18. Sun JS, Tian ZQ, Fu YL et al (2020) Digital twins in human understanding: a deep learning-based method to recognize personality traits. Int J Comput Integr Manuf 34:14
19. Wang S, Cui L, Liu L et al (2020) Personality traits prediction based on users' digital footprints in social networks via attention RNN. In: 2020 IEEE international conference on services computing (SCC). IEEE, pp 54–56
20. Zhao JH, Zeng DL, Xiao YJ et al (2020) User personality prediction based on topic preference and sentiment analysis using LSTM model. Pattern Recognit Lett 138:397–402
21. Başaran S, Ejimogu OH (2021) A neural network approach for predicting personality from Facebook data. SAGE Open 11:21582440211032156

22. Bakry MR, Nasr MM, Alsheref FK (2022) Personality classification model of social network profiles based on their activities and contents. Int J Adv Comput Sci Appl 13:16–21
23. Kamalesh MD, Bharathi B (2022) Personality prediction model for social media using machine learning Technique. Comput Electr Eng 100:12
24. Yang B (2022) Analysis model of personality and psychological characteristics of network users under high-pressure working environment. Secur Commun Netw 2022:10
25. Zhou LX, Zhang ZY, Zhao LJ et al (2022) Attention-based BiLSTM models for personality recognition from user-generated content. Inf Sci 596:460–471
26. Ahmad N, Siddique J (2017) Personality assessment using Twitter tweets. In: ZanniMerk C, Frydman C, Toro C, Hicks Y, Howlett RJ, Jain LC (eds) Knowledge-based and intelligent information and engineering systems. Elsevier Science Bv, Amsterdam, pp 1964–1973
27. Bhatti SK, Muneer A, Lali MI et al (2017) Personality analysis of the USA public using Twitter profile pictures. IEEE, New York
28. Guntuku SC, Lin WS, Carpenter J et al (2017) Studying personality through the content of posted and liked images on Twitter. Assoc Computing Machinery, New York
29. Raje MS, Singh A (2018) Personality detection by analysis of Twitter profiles. In: Abraham A, Cherukuri AK, Madureira AM, Muda AK (eds) Proceedings of the eighth international conference on soft computing and pattern recognition. Springer International Publishing Ag, Cham, pp 667–675
30. Jeremy NH, Prasetyo C, Suhartono D (2019) Identifying personality traits for Indonesian user from Twitter dataset. Int J Fuzzy Log Intell Syst 19:283–289
31. Tutaysalgir E, Karagoz P, Toroslu IH (2019) Clustering based personality prediction on Turkish tweets. In: 2019 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM). IEEE, pp 825–828
32. Kosan MA, Karacan H, Urgen BA (2022) Predicting personality traits with semantic structures and LSTM-based neural networks. Alex Eng J 61:8007–8025
33. Karanatsiou D, Sermpezis P, Gruda D et al (2022) My tweets bring all the traits to the yard: predicting personality and relational traits in online social networks. ACM Trans Web 16:26
34. Matsumoto K, Kishima R, Tsuchiya S et al (2022) Relationship between personality patterns and harmfulness: analysis and prediction based on sentence embedding. Int J Inf Technol Web Eng 17:24
35. Rathi S, Verma JP, Jain R et al (2022) Psychometric profiling of individuals using Twitter profiles: a psychological natural language processing based approach. Concurr Comput Pract Exp 34:19
36. Elbaghazaoui BE, Amnai M, Fakhri Y (2023) Predicting the next word using the Markov chain model according to profiling personality. J Supercomput 16
37. Ferwerda B, Tkalcic M, Acm, (2018) Predicting users' personality from Instagram pictures: using visual and/or content features? Assoc Computing Machinery, New York
38. Kim Y, Kim JH (2018) Using computer vision techniques on Instagram to link users' personalities and genders to the features of their photos: an exploratory study. Inf Process Manag 54:1101–1114
39. Huang SG, Zheng JH, Xue D et al (2017) Predicting big-five personality for micro-blog based on robust multi-task learning. In: Zou B, Li M, Wang H, Song X, Xie W, Lu Z (eds) Data science, Pt 1. Springer, Berlin, pp 486–499
40. Li C, Wan J, Wang B (2017) Personality prediction of social network users. In: 2017 16th international symposium on distributed computing and applications to business, engineering and science (DCABES), pp 84–87
41. Lin J, Mao W, Zeng DD (2017) Personality-based refinement for sentiment classification in microblog. Knowl-Based Syst 132:204–214
42. Han SQ, Huang HL, Tang YQ (2020) Knowledge of words: an interpretable approach for personality recognition from social media. Knowl-Based Syst 194:20
43. Wang P, Yan Y, Si YD et al (2020) Classification of proactive personality: text mining based on Weibo text and short-answer questions text. IEEE Access 8:97370–97382
44. Wang P, Yan M, Zhan X et al (2021) Predicting self-reported proactive personality classification with Weibo text and short answer text. IEEE Access 9:77203–77211
45. Jiang Y, Deng S, Li H et al (2021) Predicting user personality with social interactions in Weibo. Aslib J Inf Manag 73(6):839–864
46. Yang K, Yuan H, Lau RYK (2022) PsyCredit: an interpretable deep learning-based credit assessment approach facilitated by psychometric natural language processing. Expert Syst Appl 198:13
47. Alsadhan N, Skillicorn D (2017) Estimating personality from social media posts. In: 2017 IEEE international conference on data mining workshops (ICDMW), pp 350–356
48. Varshney V, Varshney A, Ahmad T et al (2017) Recognising personality traits using social media. In: 2017 IEEE international conference on power, control, signals and instrumentation engineering (ICPCSI), pp 2876–2881
49. Guan Z, Wu B, Wang B et al (2020) Personality2vec: network representation learning for personality. In: 2020 IEEE fifth international conference on data science in cyberspace (DSC). IEEE, pp 30–37
50. Khan AS, Ahmad H, Asghar MZ et al (2020) Personality classification from online text using machine learning approach. Int J Adv Comput Sci Appl 11:460–476
51. Sun XG, Liu B, Meng Q et al (2020) Group-level personality detection based on text generated networks. World Wide Web 23:1887–1906
52. Lopez-Santillan R, Gonzalez LC, Montes-Y-Gomez M et al (2023) When attention is not enough to unveil a text's author profile: enhancing a transformer with a wide branch. Neural Comput Appl 34:20
53. Strickland E (2022) Andrew NG: Unbiggen AI. In: IEEE spectrum. https://spectrum.ieee.org/andrew-ng-data-centric-ai
54. Rammstedt B, John OP (2007) Measuring personality in one minute or less: a 10-item short version of the Big Five Inventory in English and German. J Res Pers 41:203–212
55. Horzum MB, Tuncay A, Padir MA (2017) Adaptation of big five personality traits scale to Turkish culture. Sakarya Univ J Educ 7:398–408
56. Gosling SD, Rentfrow PJ, Swann WB Jr (2003) A very brief measure of the Big-Five personality domains. J Res Pers 37:504–528
57. Atak H (2013) On-Maddeli Kişilik Ölçeği'nin Türk Kültürü'neUyarlanması
58. Donnellan MB, Oswald FL, Baird BM et al (2006) The mini-IPIP scales: tiny-yet-effective measures of the Big Five factors of personality. Psychol Assess 18:192
59. Korkmaz M, Somer O, Güngör D (2013) Ergen örneklemde beş faktör kişilik envanteri'nin cinsiyetlere göre ortalama ve kovaryans yapılarıyla ölçme eşdeğerliği. Eğitim ve Bilim 38
60. Soto CJ, John OP (2017) The next Big Five Inventory (BFI-2): developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. J Pers Soc Psychol 113:117
61. Soto C The Big Five Inventory–2 (BFI-2). In: Colby College—Personality Lab. https://www.colby.edu/psych/personality-lab/#4

62. Schweter S (2020) BERTurk—BERT models for Turkish. In: Zenodo. https://doi.org/10.5281/zenodo.3770924
63. Clark K, Luong M-T, Le QV et al (2020) Electra: pre-training text encoders as discriminators rather than generators. arXiv:2003.10555
64. Bojanowski P, Grave E, Joulin A et al (2017) Enriching word vectors with subword information. Trans Assoc Comput Linguist 5:135–146
65. Grave E, Bojanowski P, Gupta P et al (2018) Learning word vectors for 157 languages. arXiv:1802.06893
66. Mikolov T, Chen K, Corrado G et al (2013) Efficient estimation of word representations in vector space. arXiv:1301.3781
67. Lau JH, Baldwin T (2016) An empirical evaluation of doc2vec with practical insights into document embedding generation. arXiv:1607.05368
68. Schuster M, Paliwal KK (1997) Bidirectional recurrent neural networks. IEEE Trans Signal Process 45:2673–2681
69. Kumar JA, Abirami S (2021) Ensemble application of bidirectional LSTM and GRU for aspect category detection with imbalanced data. Neural Comput Appl 33:14603–14621
70. Gers FA, Schmidhuber J, Cummins F (2000) Learning to forget: continual prediction with LSTM. Neural Comput 12:2451–2471
71. Graves A, Schmidhuber J (2005) Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw 18:602–610
72. Kiranyaz S, Avci O, Abdeljaber O et al (2021) 1D convolutional neural networks and applications: a survey. Mech Syst Signal Process 151:107398
73. Ramachandran P, Zoph B, Le QV (2017) Searching for activation functions. arXiv:1710.05941

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.