Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 9, No. 4, 2021, 451-463.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

DENOVA: Predicting Five-Factor Model using Deep Learning based on ANOVA

Motahare Nasiri and Hossein Rahmani*

School of Computer engineering, Iran University of Science and Technology, Tehran, Iran.

Article Info

Article History:
Received 19 January 2021
Revised 07 May 2021
Accepted 11 June 2021

DOI:10.22044/JADM.2021.10471.2186

Keywords:
Personality Dimensions, Five-Factor Model (FFM), ANOVA, Deep Learning, Word Embedding, Text Mining.

**Corresponding author: h_rahmani@iust.ac.ir (H. Rahmani).*

Abstract

Determining the personality dimensions of the individuals is very important in the psychological research works. The most well-known example of personality dimensions is the five-factor model (FFM). There are two approaches, manual and automatic, for determining the personality dimensions. In a manual approach, the Psychologists discover these dimensions through the personality questionnaires. As an automatic way, varied personal input types (textual/image/video) of people are gathered and analyzed for this purpose. In this work, we propose a method called DENOVA (DEep learning based on the aNOVA), which predicts FFM using deep learning based on an Analysis of variance (ANOVA) of words. For this purpose, DENOVA first applies ANOVA in order to select the most informative terms. Then DENOVA employs Word2Vec in order to extract document embeddings. Finally, DENOVA uses support vector machine (SVM), logistic regression, XGBoost, and multi-layer perceptron (MLP), as classifiers in order to predict FFM. The experimental results obtained show that DENOVA outperforms on average, 6.91%, the state-of-the-art methods in predicting FFM with respect to accuracy.

1. Introduction

According to psychological research, personality dimensions can reflect human interests and preferences [1-4]. Today, recognizing the interests and preferences of the human beings is consequential and practical in various fields [1, 5, 6]. Varied practical applications can use knowledge hidden in the people's personality dimensions for their purposes. Among them, the recommender systems recommend the best suggestion for music, movies, books, etc. considering the personality dimensions [7-11], the human resource department hires people according to their personality dimensions [12, 13], and the fraud management systems make a more accurate prediction of the offender or fraudster among several people by considering personality dimensions. For example, most people who commit crimes are neurotic people [14, 15].

There are several data analysis approaches to determine the personality dimensions of people. The input features of these methods are varied from standard questionnaires to complex image/audio features. Questionnaires are the oldest method of predicting personality dimensions. Today, the researchers use well-known psychological questionnaires such as NEO [16], BFI [17], and Goldberg [18]. In all the three questionnaires, there are the five options "strongly disagree", "disagree", "no opinion", "agree", and "strongly agree" as an answer. An alternative way to understand the personality dimensions of people is to analyze their writing style (textual features). How a person writes is fixed over time, and can be used as a source of information in order to examine a person's personality dimensions [3, 6, 19, 20]. Personality dimension can be predicted by analyzing the facial expressions of people in the photos or the types of reactions and movements of people in the video [21, 22]. According to the research works, "smile"

is one of the most important features for the extraversion and agreeableness people. There are a few methods [23-25] that only use audio as an input feature in order to predict the personality dimensions. In these methods, the audio features such as speech activity, word n-gram, and sound frequency are applied for the analysis [26].

Among the mentioned input features, the textual data is the most available and even trustable input feature [19, 21, 27]. People are eager to freely show their emotions and feeling in their weblogs/social media etc. [2, 19, 28].

The most notable example of personality dimensions is the Five-Factor Model (FFM) [8], which models personality based on five dimensions: Openness to Experience (Opn), Conscientiousness (Con), Extraversion (Ext), Neuroticism (Neu), and Agreeableness (Agr). The most significant features of each personality dimension are shown in the following [19, 28-30]:

1. Neuroticism
   People having this personality character experience a lot of stress, and are always very worried about various issues. They become upset quickly, and experience sudden and drastic changes in their emotions.

2. Conscientiousness
   People having this personality character prepare themselves for the events and projects in advance. They enjoy having a pre-arranged schedule, so they prioritize their tasks, and get important tasks done first. They pay a special attention to details.

3. Extraversion
   People having this personality character enjoy being the center of attention, also like to start conversations, and usually speak before thinking. They enjoy meeting new people, so they have a lot of friends and acquaintances.

4. Openness to Experience
   People having this personality character are generally very creative. They love new experiences and new challenges. They enjoy thinking about abstract things (like philosophy).

5. Agreeableness
   People having this personality character are very interested in other people and humans. They are compassionate, and care for others, and empathize with them. They like to do things for the happiness of the others, and help others, if necessary.

Accordingly, in this work, we focused on the textual input data, and proposed a method called DENOVA (DEep learning based on the ANOVA). DENOVA first applies Analysis of Variance (ANOVA) in order to select the most discriminant terms in the context of FFM prediction, and then DENOVA uses deep learning to involve the context of informative terms in the prediction task. At the end, DENOVA utilizes SVM, Logistic Regression, XGBoost, and MLP for the final prediction. The main contribution of this work is intelligent ways of combining deep learning methods with a statistical ANOVA method in order to discover the discriminative and informative features for the task of predicting FFM.

The structure of this paper is as what follows. Section 2 overviews the available methods in predicting FFM from the text. DENOVA is described in details in Section 3. The numerical evaluation is presented in Section 4. Section 5 includes a discussion of the results and proposes promising directions for the future research works.

2. Background

Nowadays, considering personality dimensions has an important role in varied applications [1, 10, 27]. In this section, we review the available text-based approaches for predicting FFM. Majumder et al. [19] have proposed a method that predicts FFM using the Essay dataset [31], which includes anonymous-written texts of 2,468 students and their personality scores. They extracted the text features with the Word2Vec, CNN, NRC [32], and Mairesse tools [33], and finally, they applied the SVM and MLP algorithms for classification. Tighe et al. [3] have used the Essay dataset and LIWC [34] tools in order to analyze the text. They applied ZeroR, LibSVM, SMO, and SimpleLogistic for classification. The result of their research work shows that conscientiousness, agreeableness, and neuroticism use negative emotions. conscientiousness, openness to experience, and agreeableness use swear words. Poria et al. [29] have utilized LIWC and MRC [35] in order to extract features from the Essay dataset. They combined two concepts, ConcepNet [36] and EmoSenticNet [37], in order to predict emotions in text, and applied SVM as a classifier. Tandera et al. [38] have used the MyPersonality

dataset [39], which includes the status updates of 251 Facebook users. They applied the LIWC, SPLICE tools [40] and the Glove algorithm in order to extract the features. Then they executed MLP, LSTM, SVM, and Naïve Bayes for classification. Tadesse et al. [41] have used LIWC and SPLICE in order to analyze the text of the MyPersonality dataset. Finally, they utilized the linear regression, Gradient Boosting, SVM, and XGBoost classification methods. They realized that the extraverted people usually use past-tense verbs and additionally, and prefer to write short-length messages. The neurotic users update their status with negative emotions such as anger and anxiety. Nowson et al. [42] have examined the text of 71 bloggers. They applied LIWC, MRC, and n-gram in order to extract the text features, and then applied SVM for classification. The results of their research work showed that women used more pronouns and words in writing that indicate their emotions and physical states. In contrast, men often talk about foreign events and use more articles. Drexel [43] has analyzed Indonesian WhatsApp users' messages. He executed Word2Vec and FastText [44-46] for extracting the features. He applied AdaBoost and Gaussian Naïve Bayes for classification. Philip et al. [47] have used two Facebook and Twitter datasets. They employed WordNet [48] in order to extract the features, and applied the SVM and Naïve Bayes algorithms for classification. Zheng et al. [49] have used the MyPersonality dataset and utilized n-gram, LIWC, and Word2Vec algorithm for feature extraction. Finally, they applied the semi-supervised learning algorithm for classification. Their research work showed that neuroticism people use dirty and curse words more often. extraversion people use words related to their life, like "weekend", "holiday", "dressing", etc. Salem et al. [1] have collected the last 3,200 Arabic tweets of 92 Egyptian Twitter users who responded to the NEO questionnaire [16]. They used TF * IDF and n-gram in order to extract features from text and multinomial Naïve Bayes, SVM, decision tree, and KNN for classification.

3. Proposed Method

The main steps of our proposed DENOVA method are shown in Figure 1, and are described in details in the following sub-sections.

3.1. Dataset

DENOVA uses textual data (human-written texts) in order to predict FFM. In this work, DENOVA uses the two following datasets:

1. Essay dataset
   In order to collect the Essay [31] dataset, the researchers asked some psychology students to write down everything that came to their minds in 20 minutes. Additionally, the students were asked to answer the BFI personality questionnaire. The result of the BFI questionnaire was stored as participants' personality dimensions (labels for supervised learning). As a result, the Essay dataset contains 2468 anonymous written texts with corresponding personality labels.

2. MyPersonality dataset
   MyPersonality [39] was a Facebook application that allowed its users to participate in a psychological research work by answering the BFI personality questionnaire. The participants' status updates are considered as the input textual data, and the result of the BFI questionnaire is stored as the personality labels. The MyPersonality dataset contains the information of 251 active Facebook users.

3.2. Pre-processing

DENOVA uses the NLTK library for pre-processing the texts [50, 51]. At first, DENOVA removes the characters like "@", "#", "$", and "%" from the text. Then it removes all the links in the texts, and converts all the letters to the lowercase. Next, it tokenizes all the words based on the space between them, and uses the lemmatizer library [52, 53]. Finally, it removes all the stopwords.

3.3. ANOVA

Analysis of variances (ANOVA) [54] is a collection of statistical models that can be used for discovering the discriminant features among the groups. The method was invented by R.A. Fisher, a famous biologist and statistician. We applied this statistical approach in our research work to find the most discriminant terms in the context of FFM prediction. Our personality detection task is a multi-label classification task in which our classifier would recommend the five personality labels neuroticism, conscientiousness, extraversion, openness to experience, and agreeableness for each person. In order to address this multi-label classification task, we proposed the two approaches DENOVA_Rest and DENOVA_5Way. DENOVA_5Way (discussed in details in Section 3.3.1) considered all the five

labels directly in one classification task, while DENOVA_Rest (which is discussed in details in Section 3.3.2) maps the multi-label classification task into 5 binary classification tasks.

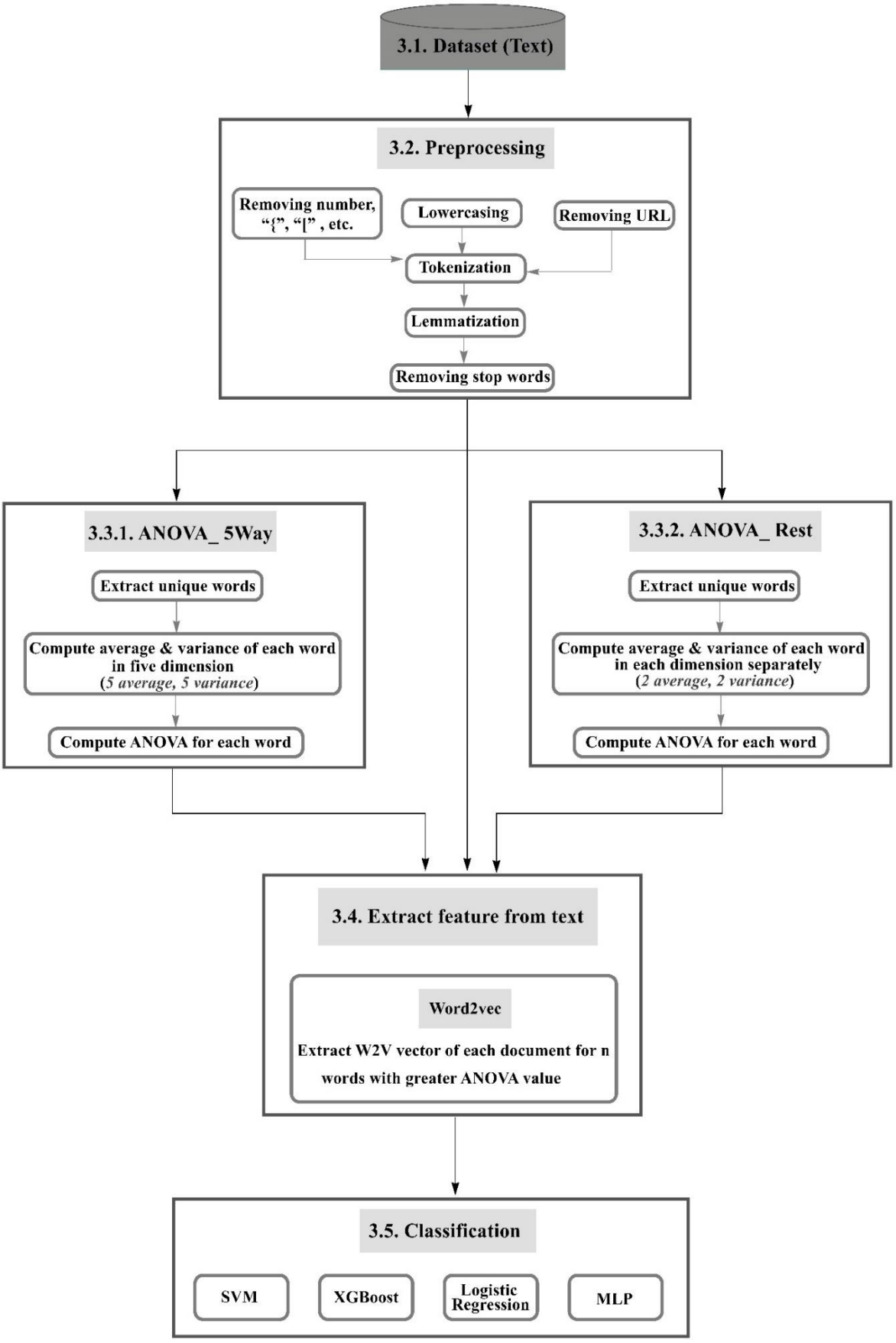


Figure 1. An overview of DENOVA method.

3.3.1. DENOVA_5Way

DENOVA_5Way addresses the task of five personality dimensions directly, and uses ANOVA in order to find the discriminative words for each personality dimension in the Essay and MyPersonality datasets. For this purpose, DENOVA_5Way follows the following steps:

1. DENOVA_5Way collects unique words (about 35,000 words) from the two datasets Essay and MyPersonality. In order to do this step, we saved two datasets (Essay and MyPersonality) in a CSV file, and after pre-processing, stored their text in an array. Finally, we used the array functions in Python and extracted non-duplicate elements of the array as unique words.

2. DENOVA_5Way considers FFM as five baskets of neuroticism, conscientiousness, extraversion, openness to experience, and agreeableness. Then DENOVA_5Way pre-processes the existing texts for each person. Each person has a "yes" or "no" value for each personality dimension.

3. DENOVA_5Way obtains the mean and variance for each unique word in five different baskets. At this step, we have five averages and five variances per word.

4. DENOVA_5Way calculates the ANOVA value for word w_i according to "(1),":

$$ANOVA(w_i) = \frac{\sigma^2(\mu_{i1},\ \mu_{i2}, \mu_{i3}, \mu_{i4}, \mu_{i5})}{\mu(\sigma_{i1}^2, \sigma_{i2}^2, \sigma_{i3}^2, \sigma_{i4}^2, \sigma_{i5}^2)} \qquad (1)$$

where μ_{ik} is the average frequency of word w_i in personality dimension k, σ_{ik}^2 is the variance word w_i in personality dimension k, and k includes the five personality dimensions neuroticism, conscientiousness, extraversion, openness to experience, and agreeableness. Finally, DENOVA_5way indicates the informativeness of each unique word in discriminating the five personality dimensions.

3.3.2. DENOVA_Rest

DENOVA_Rest maps the task of predicting the five personality dimensions into 5 binary classification tasks. DENOVA_Rest uses ANOVA in order to find the discriminative words in the five personality dimensions in the Essay and MyPersonality datasets. For this purpose, DENOVA_Rest follows the following steps:

1. DENOVA_Rest extracts the unique words (about 35,000 words) from the two datasets Essay and MyPersonality. In order to do this step, we saved the two datasets (Essay and MyPersonality) in a CSV file, and after pre-processing, stored their text in an array. Finally, we used the array functions in Python and extracted the non-duplicate elements of the array as unique words.

2. DENOVA_Rest transforms the multi-label approach (DENOVA_5Way) into a binary approach (DENOVA_Rest). DENOVA_Rest, each time, takes only one specific personality dimension and calculates the ANOVA value for the words of that personality dimension. For this purpose, DENOVA_Rest first reads and pre-processes the text written by each person. It divides the text into two baskets: one text basket with the desired personality dimension, and the other one contains the text corresponding to the remaining personality dimensions.

3. DENOVA_Rest obtains the unique words used in the whole dataset, and then computes the mean and variances of each unique word in each one of the two baskets. For each word w_i, DENOVA_Rest has two mean values and two variance values, one value for the text-basket related to a specific-chosen personality dimension (such as neuroticism) and one value for the text-basket related to the other four personality dimensions (such as conscientiousness, extraversion, openness to experience, and agreeableness).

4. DENOVA_Rest measures the ANOVA value for word w_i according to "(2),":

$$ANOVA(w_i) = \frac{\sigma^2(\mu_{i1},\ \mu_{i2})}{\mu(\sigma_{i1}^2, \sigma_{i2}^2)} \qquad (2)$$

where μ_{i1} is the average frequency of word w_i in the desired personality dimension, μ_{i2} is the average frequency of word w_i in the other four personality dimensions, σ_{i1}^2 is the variance of w_i in the desired personality dimension, and σ_{i2}^2 is the variance of word w_i in the other four personality dimensions. Finally, DENOVA_Rest indicates the informativeness of each unique word in discriminating a specific personality dimension

comparing to the other four personality dimensions. As a result, DENOVA_Rest calculates 5 discriminate scores ("(2),") for each unique word w_i.

3.4. Feature Extraction

After calculating the discriminativeness scores (with respect to ANOVA_Rest or ANOVA_5Way), in this section, first, we select the unique words whose scores are higher than λ value, and then we use the Google's pre-trained Word2Vec model [55, 56] in order to build a semantic vector representation of those selected words. The vector length is 300 features. For this purpose, the following steps are carried out:

- DENOVA pre-processes each person's text of the Essay dataset.

- For each word in the person's text, we calculate the word's ANOVA value according to "(1)," and "(2),". If we consider the ANOVA value as a word's importance, we follow the process shown in Figure 2 to first select the most discriminated words with the highest ANOVA values, and secondly, aggregate the Word2Vec vectors of those selected words to represent the person's text in an aggregated and still informative way, and finally, apply the classifiers to predict FFM.

```
For   F = start:1  to  end:20,000   step:1000
{
   Renge = F word with the highest value of  ANOVA
   For  Person  in EssaysDataset:
   {
         Person.PText = Preprocess(Person.Text)
         PersonW2VLists = []
         For word in Person.PText
         {
                  If word in Renge:
                  {
                  PersonW2VLists.add(W2V(word))
                  }
         }
         Person.W2VFeatures = Means of W2V
         Vectors in PersonW2VLists
   }
Model = Classification(
                  EssaysDataset.PersonsW2VFeatures,
                  EssaysDataset.PersonsDimensions,
                  "10 Fold Cross Validation")

print(Model.Accuracy)
}
```

Figure 2. Pseudo-code of DENOVA approch.

3.5. Classification

After selecting the informative features, DENOVA applies SVM [57] (as classifier C_1), logistic regression [58] (as classifier C_2), XGBoost (as classifier C_3), and MLP (as classifier C_4), and uses a 10-fold cross-validation for evaluating the classifiers' accuracies.

The words with the highest ANOVA values are considered as the informative input features F, and accordingly, feed into the classifier C_i (i ranges from 1 to 4). As shown in Figure 2, we evaluate DENOVA with a varied number of input features. The number of features varies from 1 ($|F| = 1$) to 20,000 ($|F| = 20000$) with $step = 1000$. Applying each classifier C_i on each feature count $|F|$ would result in an evaluation e_{ik}, where k varied from 1 to 20000. In order to aggregate the evaluation values, we apply the "Mean" function. Mean (e_{ik}), $k = 1$ to 20,000, indicates the average accuracy of classifier C_i with respect to the varied number of input features (From $|F| = 1$ to $|F| = 20000$).

4. Numerical Evaluation

In this section, we overview the results of applying DENOVA_Rest and DENOVA_5Way on the Essay and MyPersonality datasets.

We applied DENOVA_5Way on the two datasets Essay and MyPersonality, and Table 1 shows 10 words with the highest ANOVA values in the DENOVA_5Way approach. These words are considered as the most discriminant features in predicting the five personality dimensions. As discussed earlier, these 10 words are discriminative but it is not clear how much they are informative concerning one specific personality dimension. Surprisingly, there are some names such as "gibson" and "messi" among the 10-top most informative words.

Table 1. 10 words with the highest ANOVA value in the DENOVA_5Way method.

Word	ANOVA value
gibson	0.123987
miscellaneous	0.123321
resurrection	0.122054
messi	0.121502
disjointed	0.119587
provisional	0.113085
locate	0.11024
airborne	0.11024
intrude	0.108586
reptilian	0.107769

We applied DENOVA_Rest on the two datasets Essay and MyPersonality, and Table 2 shows 10

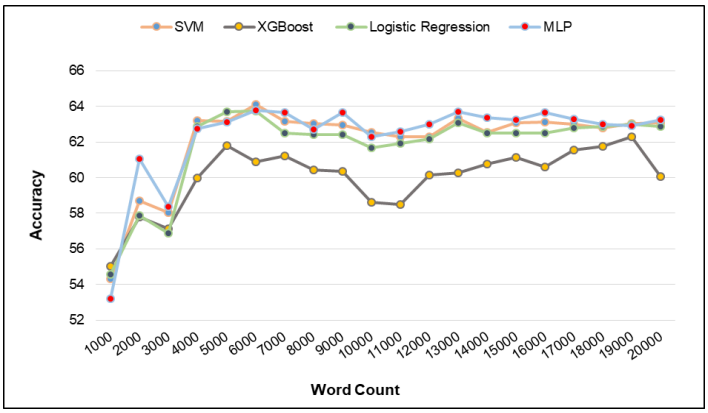
words with the highest ANOVA values in the DENOVA_Rest approach for each personality dimension. These words are considered as the most discriminant features in predicting one specific personality dimension. For example, in the personality dimension of "extraversion", there are words like "fun", "perhaps", etc. that are consistent with the basic definitions of this personality dimension. In the "neuroticism" personality dimension, there are words such as a "beat", "stress", etc. that reflect the negative and rough feelings in these people.

In the DENOVA_5Way method, the words separate the five personality dimensions from each other. These words do not represent a specific personality dimension but they are unique words that are less frequently repeated in the texts, and are considered as the most discriminant features in predicting the five personality dimensions. However, the DENOVA_Rest method transforms the multi-label approach into a binary approach, and specifically examines the words related to each personality dimension. As a result, DENOVA_Rest selects the words that represent a specific personality dimension much better than DENOVA_5Way, and accordingly, the selected words are more interpretable in the context of a specific personality dimension.

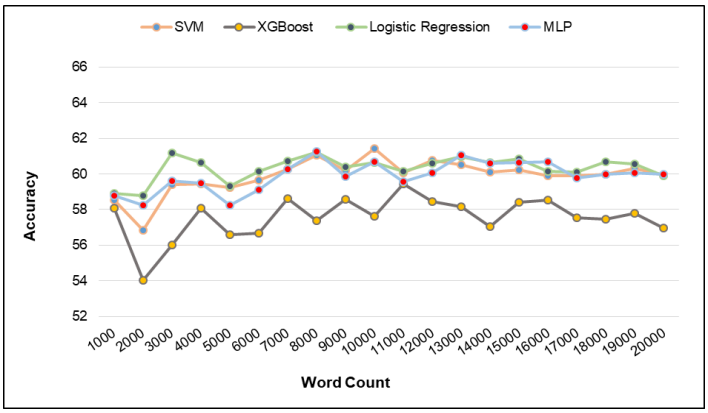
Table 2. 10 words (per personality dimension) with the highest ANOVA values in DENOVA_Rest.

Index	Agreeableness	Conscientiousness	Extraversion	Neuroticism	Openness to experience
1	wan	offensive	sorority	beat	class
2	family	decision	economy	stressed	school
3	bay	fuzzy	perhaps	scared	homework
4	dizzy	student	dreaded	feel	home
5	apology	standardized	boyfriend	hurt	college
6	translation	sleepless	shyness	carey	accurate
7	arrangement	joseph	programming	froze	world
8	awsome	able	fun	hate	go
9	harbor	vocabulary	generally	inviting	spin
10	retail	conciousness	report	acquired	going

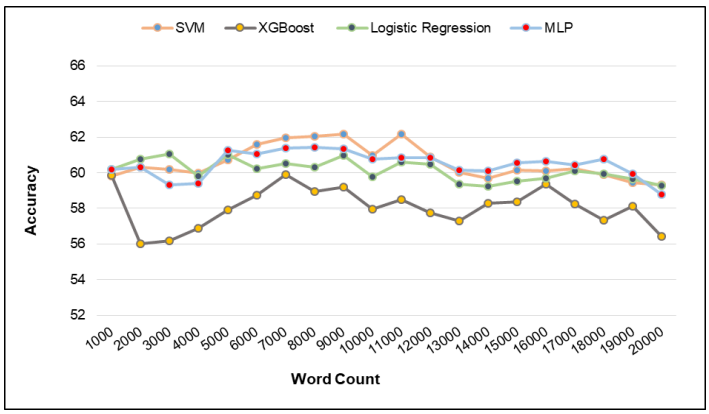
Using the two approaches ANOVA_Rest and ANOVA_5Way, we calculate the ANOVA value for each word w_i. In the next step, we select the K (from 1 to 20,000 step 1000) most discriminative words with the highest ANOVA values, and then we build Word2Vec for the selected words. Finally, these vectors are considered as the input vectors to four classifiers. Figures 3 and 4 show the results of applying 4 classifiers to the varied number of input features selected with respect to ANOVA_5Way and ANOVA_Rest for the Essay dataset, respectively.



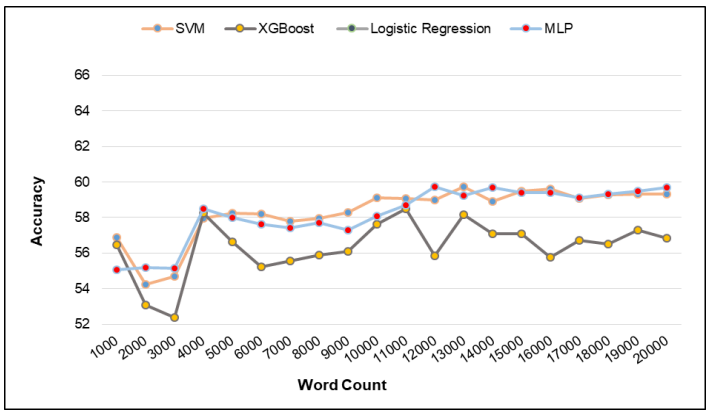
(a) Openness to experience dimension



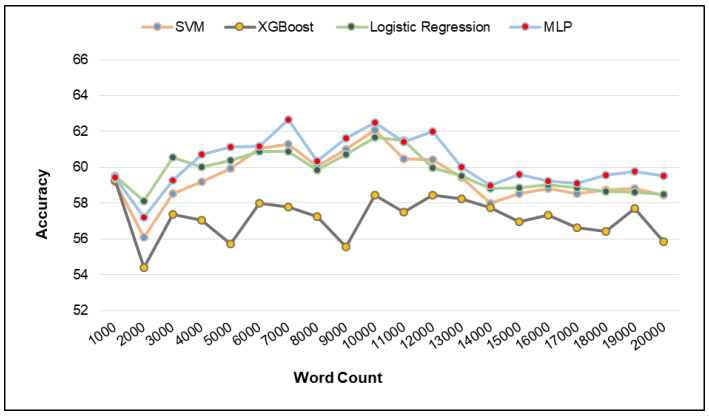
(b) Conscientiousness dimension



(c) Agreeableness dimension

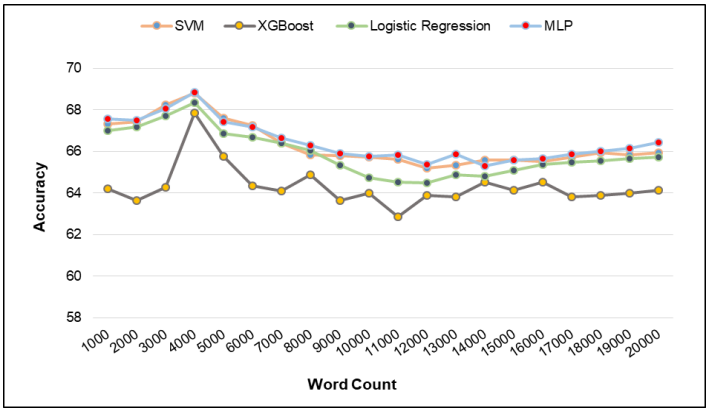


(d) Neuroticism dimension

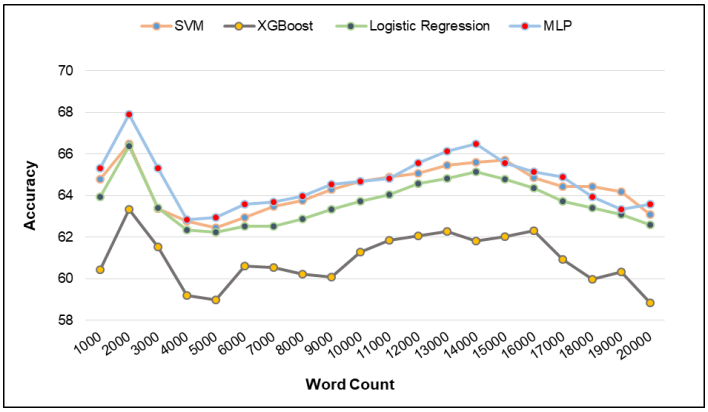


(e) Extraversion dimension

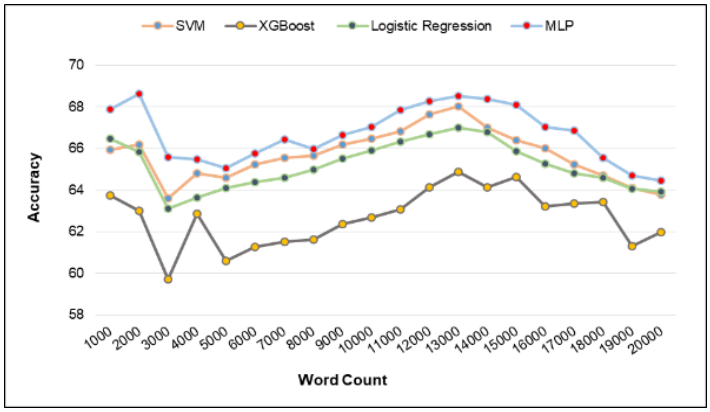
Figure 3. Results of DENOVA_5way in predicting five personality dimensions with respect to the varied number of input features.



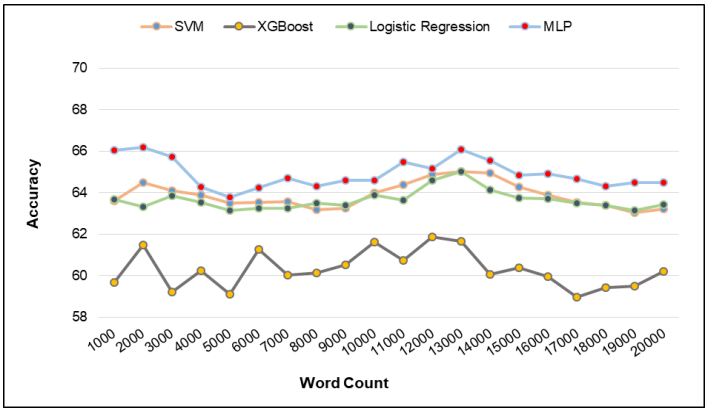
(a) Openness to experience dimension



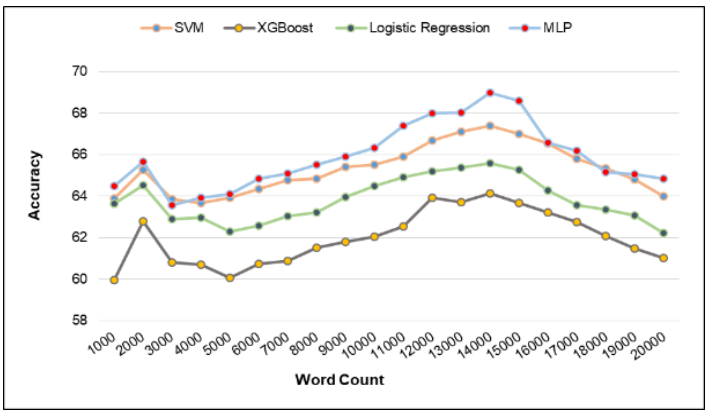
(b) Conscientiousness dimension



(c) Agreeableness dimension



(d) Neuroticism dimension



(e) Extraversion dimension

Figure 4. Results of DENOVA_Rest in predicting five personality dimensions with respect to the varied number of input features.

Table 3. Comparing the results of applying four classifiers on the Essay dataset to the state-of-the-art method (the highest accuracy values are shown in bold style).

	DENOVA_5Way				DENOVA_Rest				
	MLP	**Logistic Regression**	**XGBoost**	**SVM**	**MLP**	**Logistic Regression**	**XGBoost**	**SVM**	**State-of-the-art [8]**
Opn	62.32	61.74	60.02	62.09	**66.46**	65.89	64.31	66.33	62.68
Con	59.90	60.33	57.58	59.89	**64.70**	63.68	60.92	64.33	57.30
Agr	60.47	60.13	58.06	60.58	**66.71**	65.19	62.67	65.70	56.71
Neu	58.20	57.85	56.35	58.30	**64.92**	63.65	60.30	63.88	59.38
Ext	60.26	59.74	57.18	59.43	**65.90**	63.81	61.98	65.29	58.09
Average per classifier	60.23	59.96	57.84	60.06	**65.74**	64.44	62.04	65.11	58.83
Average per method	59.52				64.33				58.83

As shown in Table 3:

- In the DENOVA_5Way method, the SVM, MLP, and logistic regression methods outperformed the state-of-the-art method with respect to the average accuracy in the agreeableness, extraversion, and conscientiousness dimensions, respectively.

- In the DENOVA_5Way method, the average accuracy of MLP, SVM, and logistic regression classifications, on average 1.4%, 1.23%, and 1.13% are higher than the maximum accuracy of the state-of-the-art method, respectively.

- In the DENOVA_Rest method, all of the classification methods in all the five dimensions significantly outperformed the state-of-the-art method with respect to the average accuracy. However, the MLP classifier has the highest results.

- In the DENOVA_Rest method, the average accuracy of the MLP, SVM, logistic regression, and XGBoost classifications, on average, 6.91%, 6.27%, 5.61%, and 3.21% is higher than the maximum accuracy of the state-of-the-art method, respectively.

- In general, the DENOVA_5Way and DENOVA_Rest methods outperform, on average, the state-of-the-art method 0.69% and 5.5%, respectively.

According to the numerical evaluation of our proposed DENOVA method, the DENOVA_Rest method has better results in predicting all the five personality dimensions than the DENOVA_5Way method. The DENOVA_Rest method examines each personality dimension separately, and extracts the words that define each dimension more accurately. However, the DENOVA_5Way method only analyzes the words that can separate five dimensions, and this reduces the accuracy of the prediction.

As shown in Table 3, the best prediction method is the MLP method whose input vector is built according to the ANOVA_Rest approach. Figure 5 compares the accuracy of the DENOVA_Rest method achieved by the MLP classification to the accuracy of the state-of-the-art method. MLP outperforms the state-of-the-art method, on average, 6.91%.

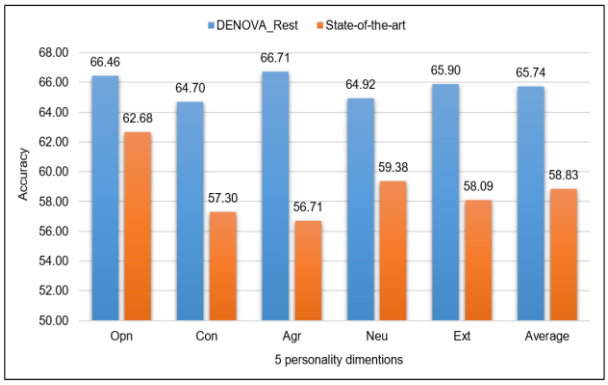


Figure 5. Comparing the results of the MLP classifier whose input features are selected by ANOVA_Rest, with the state-of-the-art method in comparing the five personality dimensions. MLP outperforms the state-of-the-art method, on average, 6.91%.

In order to investigate the positive or negative effect of the stop words in this work, we reviewed our proposed method without removing the stop words. Figures 6 and 7 show the results of the DENOVA_5Way and DENOVA_Rest methods for the presence of stop words compared to the removal of stop words in the pre-processing step.

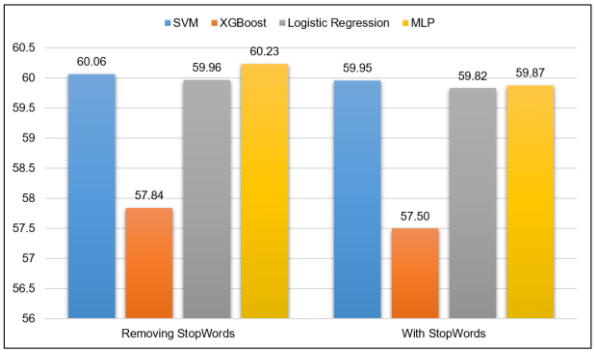


Figure 6. Average accuracy of four classifiers by not removing stop words compared to removing stop words in the DENOVA_5Way method.

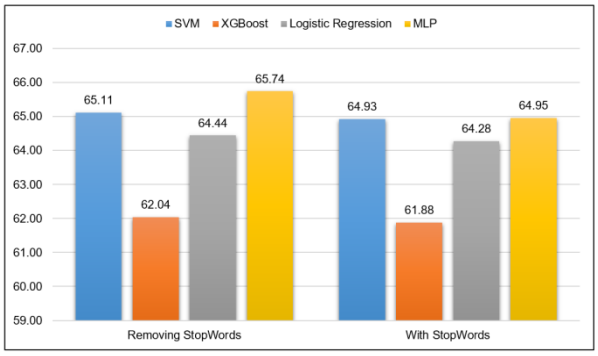


Figure 7. Average accuracy of four classifiers by not removing stop words compared to removing stop words in the DENOVA_Rest method.

As it is shown in Figure 8, all four classifiers had higher accuracy in both methods (DENOVA_5Way and DENOVA_Rest) in the case of removing the stop words.

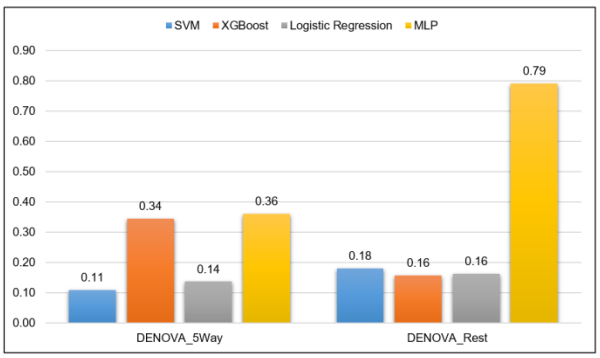


Figure 8. Difference between the average accuracy of the four classifiers by removing stop words compared to not removing stop words in the DENOVA_5Way and DENOVA_Rest methods.

5. Conclusions and Future Works

In the recent years, predicting the personality dimensions has attracted a lot of attention in varied applications, mainly in the recommender systems. To the best of our knowledge, the previous methods considered varied ranges of input types from standard questionnaires to textual/image/video features of people for this purpose. Among the mentioned input features, the textual data is the most available input feature, and in this work, we focused on this type of data. Accordingly, we proposed DENOVA, a method that predicts the five personality dimensions (or five-factor model (FFM), in other words) using deep learning based on the analysis of variance (ANOVA). The experimental results obtained show that DENOVA outperforme, on average, 6.91%, the state-of-the-art method with respect to accuracy. Regarding the future research works, we aim to apply other feature extraction methods (BERT, GloVe, etc.) along with this statistical analysis.

References

[1] M. S. Salem, S. S. Ismail, M. Aref, "Personality Traits for Egyptian Twitter Users Dataset," in ACM, 2019.

[2] P. H. Arnoux, A. Xu, N. Boyette, J. Mahmud, R. Akkiraju, V. Sinha, "25 Tweets to Know You: A New Model to Predict Personality with Social Media," in *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[3] E. P. Tighe, J. C. Ureta, B. A. Pollo, C. K. Cheng, R. de Dios Bulos, "Personality Trait Classification of Essays with the Application of Feature Reduction," in *4th Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2016), IJCAI 2016*, New York City, USA, 2016.

[4] N. Taghvaei, B. Masoumi, M. R. Keyvanpour, "A Hybrid Framework for Personality Prediction based on

Fuzzy Neural Networks and Deep Neural Networks," *Journal of AI and Data Mining*, 2021.

[5] S. Arjaria, A. Shrivastav, A. S. Rathore, V. Tiwari, "Personality Trait Identification for Written Texts Using MLNB," in *Data, Engineering and Applications*, Singapore, Springer, 2019, pp. 131-137.

[6] D. R. Moreno, J. C. Gomez, D. L. Almanza-Ojeda, M. A. Ibarra-Manzano, "Prediction of Personality Traits in Twitter Users with Latent Features," in *IEEE*, 2019.

[7] M. A. Nunes, R. Hu, "Personality-based recommender systems: an overview," in *Proceedings of the sixth ACM conference on Recommender systems*, 2012.

[8] R. R. McCrae, O. P. John, "An introduction to the five-factor model and its applications," *Journal of personality*, vol. 60, pp. 175-215, 1992.

[9] T. T. Nguyen, F. M. Harper, L. Terveen, J. A. Konstan, "User personality and user satisfaction with recommender systems," *Information Systems Frontiers*, 2018.

[10] G. Nave, J. Minxha, D. M. Greenberg, M. Kosinski, D. Stillwell, J. Rentfrow, "Musical preferences predict personality: evidence from active listening and facebook likes," *Psychological Science*, 2018.

[11] R. Gao, B. Hao, S. Bai, L. Li, A. Li, T. Zhu, "Improving user profile with personality traits predicted from social media content," *Proceedings of the 7th ACM conference on recommender systems*, 2013.

[12] C. Aydogmus, S. M. Camgoz, A. Ergeneli, O. T. Ekmekci, "Perceptions of transformational leadership and job satisfaction: The roles of personality traits and psychological empowerment," 2018.

[13] P. Steel, J. Schmidt, F. Bosco, K. Uggerslev, "The effects of personality on job satisfaction and life satisfaction: A meta-analytic investigation accounting for bandwidth{fidelity and commensurability," *Human Relations*, 2019.

[14] K. M. Beaver, J. C. Boutwell BB, Barnes, M. G. Vaughn, M. DeLisi, "The association between psychopathic personality traits and criminal justice outcomes: Results from a nationally representative sample of males and females," *Crime Delinquency*, 2017.

[15] S. G. Van de Weijer, E. R. Leukfeldt, "Big five personality traits of cybercrime victims," *Cyberpsychology, Behavior, and Social Networking*, 2017.

[16] P. T. Costa, R. R. McCrea, "Revised neo personality inventory (neo pi-r) and neo five-factor inventory (neo-ffi)," *Psychological Assessment Resources*, 1992.

[17] O. P. John, E. M. Donahue, R. L. Kentle, "The big five inventory—versions 4a and 54," 1991.

[18] L. R. Goldberg, J. A. Johnson, H. W. Eber, R. Hogan, M. C. Ashton, C. R. Cloninger, H. G . Gough, "The international personality item pool and the future of public-domain personality measures," *Journal of Research in personality*, 2006.

[19] N. Majumder, S. Poria, A. Gelbukh, E. Cambria, "Deep learning-based document modeling for personality detection from text," *IEEE Intelligent Systems*, 2017.

[20] E. Tighe, C. Cheng, "Modeling Personality Traits of Filipino Twitter Users," in *2nd Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES 2018), NAACL HLT 2018, New Orleans, Louisiana, USA*, 2018.

[21] I. Wilf, Y. Michaeli, S. Gilboa, D. H. Gavriel, G. Bechar, "Method and system for predicting personality traits, capabilities and suggested interactions from images of a person," *United States patent application US*, 2019.

[22] M. Cristani, A. Vinciarelli, C. Segalin, A. Perina, "Unveiling the multimedia unconscious: Implicit cognitive processes and multimedia content analysis," *Proceedings of the 21st ACM international conference on Multimedia*, 2013.

[23] F. Valente, S. Kim, P. Motlicek, "Annotation and recognition of personality traits in spoken conversations from the ami meetings corpus," *Thirteenth annual conference of the international speech communication association*, 2012.

[24] T. Polzehl, S. Möller, F. Metze, "Automatically assessing personality from speech," in *IEEE Fourth International Conference on Semantic Computing*, 2010.

[25] S. I. Levitan, Y. Levitan, G. An, M. Levine, R. Levitan, A. Rosenberg, J. Hirschberg, "Identifying individual differences in gender, ethnicity, and personality from dialogue for deception detection," *Proceedings of the second workshop on computational approaches to deception detection*, 2016.

[26] L. F. Gallardo, B. Weiss, "Towards Speaker Characterization: Identifying and Predicting Dimensions of Person Attribution," *INTERSPEECH*, 2017.

[27] Y. Mehta, N. Majumder, A. Gelbukh, E. Cambria, "Recent trends in deep learning based personality detection," *Artificial Intelligence Review*, 2019.

[28] D. J. Hughes, M. Rowe, M. Batey, A. Lee, "A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage," *Computers in Human Behavior*, 2013.

[29] S. Poria, A. Gelbukh, B. Agarwal, E. Cambria, N. Howard, "Common sense knowledge based personality

recognition from text," in *Mexican International Conference on Artificial Intelligence*, 2013.

[30] R. R. McCrae, P. T. Costa, "Validation of the five-factor model of personality across instruments and observers," *Journal of personality and social psychology*, 1987.

[31] J. W. Pennebaker, L. A. King, "Linguistic styles: Language use as an individual difference," *Journal of personality and social psychology*, 1999.

[32] S. M. Mohammad, P. D. Turney, "Crowdsourcing a word–emotion association lexicon," *Computational Intelligence*, 2013.

[33] F. Mairesse, M. A. Walker, M. R. Mehl, R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of artificial intelligence research*, 2007.

[34] J. W. Pennebaker, R. L. Boyd, K. Jordan, K. Blackburn, "The development and psychometric properties of LIWC2015," 2015.

[35] M. Coltheart, "The MRC psycholinguistic database," *The Quarterly Journal of Experimental Psychology Section A*, 1981.

[36] C. Havasi, R. Speer, J. Alonso, "ConceptNet 3: a flexible, multilingual semantic network for common sense knowledge," *Recent advances in natural language processing 2007*, 2007.

[37] S. Poria, A. Gelbukh, A. Hussain, N. Howard, D. Das, S. Bandyopadhyay, "Enhanced SenticNet with affective labels for concept-based opinion mining," *IEEE Intelligent Systems*, 2013.

[38] T. Tandera, D. Suhartono, R. Wongso, Y. L. Prasetio, "Personality Prediction System from Facebook Users," *Procedia Computer Science*, 2017.

[39] M. Kosinski, S. C. Matz, S. D. Gosling, V. Popov, D. Stillwell, "Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines," *American Psychologist*, 2015.

[40] K. Moffitt, J. Giboney, E. Ehrhardt, J. K. Burgoon, J. F. Nunamaker, "Structured programming for linguistic cue extraction," *The Center for the Management of Information*, 2010.

[41] M. M. Tadesse, H. Lin, B. Xu, L. Yang, "Personality predictions based on user behavior on the facebook social media platform," *IEEE Access*, 2018.

[42] S. Nowson, J. Oberlander, "The Identity of Bloggers: Openness and gender in personal weblogs," in *AAAI spring symposium: Computational approaches to analyzing weblogs*, 2006.

[43] IB. Drexel, "Feature Engineering and Word Embedding Impacts for Automatic Personality Detection on Instant Message," in *2019 International Conference on Information Management and Technology (ICIMTech)*, 2019.

[44] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, 2017.

[45] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, T. Mikolov, "Fasttext. zip: Compressing text classification models," *arXiv preprint*, 2016.

[46] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint*, 2016.

[47] J. Philip, D. Shah, S. Nayak, S. Patel, Y. Devashrayee, "Machine learning for personality analysis based on big five model," in *Data Management, Analytics and Innovation*, Singapore, Springer, 2019, pp. 345-355.

[48] C. Fellbaum, "WordNet," *The encyclopedia of applied linguistics*, 2012.

[49] H. Zheng, C. Wu, "Predicting Personality Using Facebook Status Based on Semi-supervised Learning," in *Proceedings of the 2019 11th International Conference on Machine Learning and Computing*, 2019.

[50] E. Loper, S. Bird, "NLTK: the natural language toolkit," *arXiv preprint cs/0205028*, 2002.

[51] J. Perkins, Python 3 text processing with NLTK 3 cookbook, Packt Publishing Ltd, 2014.

[52] N. Green, P. Breimyer, V. Kumar, N. Samatova, "WebBANC: Building Semantically-Rich AnnotatedCorpora from Web User Annotations of Minority Languages," in *WebBANC: Building Semantically-Rich AnnotatedCorpora from Web User Annotations of Minority Languages*, 2009.

[53] T. Bergmanis, S. Goldwater, "Context sensitive neural lemmatization with lematus," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018.

[54] L. St, S. Wold, "Analysis of variance (ANOVA)," Elsevier, 1989.

[55] "Google Code Archive," Google, 2013. [Online]. Available:https://code.google.com/archive/p/word2vec.

[56] M. Mihaltz, 2016. [Online]. Available: https://github.com/mmihaltz/word2vec-GoogleNews-vectors.

[57] C. Cortes, V. Vapnik, "Support-vector networks," *Machine learning*, 1995.

[58] RE. Wright, "Logistic regression".

DENOVA: Predicting Five-Factor Model using Deep Learning based on ANOVA

مطهره نصیری و حسین رحمانی*

دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران، ایران.

ارسال ۲۰۲۱/۰۱۰/۱۹؛ بازنگری ۲۰۲۱/۰۵/۰۷؛ پذیرش ۲۰۲۱/۰۶/۱۱

چکیده:

تعیین ابعاد شخصیتی افراد در تحقیقات روان‌شناختی بسیار مهم است. مشهورترین نمونه ابعاد شخصیتی، مدل پنج عاملی است. برای تعیین ابعاد شخصیتی دو رویکرد وجود دارد: ۱- دستی و ۲- خودکار. در رویکرد دستی، روان‌شناسان ابعاد شخصیتی را از طریق پرسش‌نامه‌های شخصیت کشف می‌کنند. در روش خودکار، انواع مختلف ورودی (متنی، تصویری، ویدئویی و صوتی) از افراد برای این منظور جمع‌آوری و تحلیل می‌شوند. در این پژوهش، ما یک روش خودکار به نامِ DENOVA را ارائه می‌دهیم که به صورت خودکار و با دقت بالا (در مقایسه با روش‌های پیشین) پنج بعد شخصیتی افراد را پیش‌بینی نماید. مهم‌ترین نوآوری ما در این پژوهش، ترکیب یادگیری عمیق با روش آماری تجزیه و تحلیل واریانس (ANOVA) در انتخاب مهم‌ترین و متمایزکننده ترین ویژگی‌ها (کلمات) بوده است. در این پژوهش، ما روشی به نام DENOVA را ارائه می‌دهیم که پنج بُعد شخصیتی افراد را با استفاده از ترکیب یادگیری عمیق و تجزیه و تحلیل واریانس کلمات پیش‌بینی می‌کند. برای این هدف، DENOVA ابتدا ANOVA را برای انتخاب جداکننده‌ترین کلمات در هر سند اعمال می‌کند. سپس، DENOVA از Word2Vec برای استخراج بردار ویژگی‌های هر سند استفاده می‌کند. سرانجام، DENOVA از ۴رده‌بند SVM، Logistic Regression، XGBoost و MLP برای پیش‌بینی پنج بُعد شخصیتی استفاده می‌کند. نتایج ما نشان می‌دهد که DENOVA با توجه به Accuracy به طور متوسط، ۶,۹۱ درصد، نسبت به متدهای رقیب، در پیش‌بینی پنج بُعد شخصیتی عملکرد بهتری دارد.

کلمات کلیدی: ابعاد شخصیتی، مدل پنج عاملی، تجزیه و تحلیل واریانس، یادگیری عمیق، تعبیه کلمه، متن‌کاوی.