ABSTRACT

The thesis presented for the degree of Doctor of Philosophy Ph. D. on the speciality 6D070400 -“Computer Systems and Software” by Talasbek Assem Lesbekkyzy

“Profession inclination identification using machine learning”

General characteristics of research. The given work is devoted to the research and development of an application that suggests recommendations for future profession selection based on the personal characteristics of a person by identifying professional inclinations.

Relevance. Currently, the Kazakhstan market has virtually no systems for profession inclination identification. Modern society makes new demands on performance and professionalism. However, high levels of professionalism suggest a full disclosure of the potential of the individual, which is impossible without taking into account the personal characteristics of an individual. Many of the questionnaires conducted by organizations do not sufficiently define and describe the type of person for appointment, selection of personnel for certain special programs, and do not give a reliable result about the person in question whether the person will cope with certain official duties.

Career counseling aims to help people learn how to make career-related decisions wisely and confidently. This decision should be based on proper self-knowledge and careful consideration of a wide variety of alternatives. Furthermore, people should feel satisfied with their decisions, function successfully in their chosen jobs, and feel prepared for changes in career paths or adjustments in the future. Personality is a combination of a person's characteristics and attitudes in dealing with different social situations as in kindergarten, school, university, family, working team, etc . Humans are addicted to biases and prejudices that might affect their judgment accuracy. Personality can be taken as an assessment in various fields such as selection of staff, choice of profession, relationship, and health counseling. There is a great effect of personality on the learning capabilities of humans. For instance, in learning performance, we may see significant differences between persons who belong to extroverts and the ones belonging to introverts. One of the main reasons why students drop their studies in universities is poor academic performance (AP), but personality also affects AP at the same level as intellectual abilities, self-esteem, motivation, etc.. Some studies show that personality can be taken as an effective measurement in predicting academic performance, especially at the university level.

Prediction of personality type, profession inclinations are one of the modern tasks of researchers. The growth of social network usage such as Twitter, Facebook, Instagram attracted researchers for automated personality prediction and classification tasks. The core theory of these research works is Big Five Factor

Personality Model, NEO-Personality-Inventory Revised, Ten Item Personality Inventory , Myers- Briggs Type Indicator (MBTI) , etc. The existing works in this field are based on supervised learning algorithms applied on benchmark datasets; however, the major issue of them is the data, to be more precisely imbalanced classes to traits. This issue makes the task of personality prediction and classification more difficult.

The research aim. It is to develop an application that identifies the profession inclination of a person based on personality classification by using different data and applying machine learning techniques to reach a high accuracy level.

Objectives of the research. Following the aim, the following objectives are identified to be solved in this work:

- to study and analyze existing personality and its inclination prediction and classification techniques and methods
- to study and analyze the correlation between personality types and profession
- to gather and analyze data from social network accounts to apply machine learning algorithms
- to conduct experiments and implement models to predict identify profession inclinations

The object of research. The study focuses on automated methods of profession inclination and personality type identification.

Research methods. The objectives assigned were solved by carrying out theoretical and empirical research. As part of the research, we used conceptual positions of AI classical ML theories and algorithms, deep learning models, studies of leading foreign and domestic scientists in the field of recommendation systems, personality classification, probability theory, mathematical statistics, numerical analysis, data analysis, in computer science, psychology and education fields.

The scientific novelty of the work. The novelty of the dissertation is to design an automated method for profession inclination identification by taking into account the psychological characteristics of a person. The results obtained from various experiments implemented by using Instagram posts and combining models of recurrent neural network (RNN) and convolutional neural network (CNN) were first proposed in this research.

The following scientific statements are to be defined:

- Methods and algorithms for data collection;
- Methods and algorithms for identification of profession inclination;
- Designed models for automated personality classification from different types of data gathered from Instagram;
- Provided experiments, results, and discussion.

The practical significance of the research results. The practical value of the thesis is the improvement of services in the career counseling field, academic performance, and the possibility of applying the results of the research on different recommendation systems for school and university graduates that helps to improve the systems that help to identify the psycho type and inclination of students, employees, criminals, etc.

Publications. 6 works are published, including:

- 2 published papers to Q2 journal International Journal of Emerging Technologies in Learning (iJET) indexed by SCOPUS Percentile 66;
- 1 article that meets the requirements of the higher Attestation Commission of the Ministry of Education of Science of the Republic of Kazakhstan;
- 3 papers in the proceedings of international conferences.

Structure and scope of the dissertation. The thesis is presented on 82 pages of typewritten text. It consists of normative references, definitions, a list of abbreviations, an introduction, four main chapters, a conclusion, references, and an appendix. The dissertation includes 26 tables, 36 figures. The list of references consists of 104 titles.

The first chapter presents the introduction part, describes the problems and content of the research work.

The second chapter provides a literature review of existing works, describes the methods of career counseling and personality inclination identification. First, it covers MBTI theory that classifies a person into sixteen personality types. The theory includes eight scales combined in pairs: extraversion (E) - introversion (I), sensing (S) – intuiting (N), thinking (T) – feeling (F), judging (J) - perceiving (P). Functional portrait of each personality type is described one by one, as a result, it gives inclination to a particular profession.

Secondly, the chapter reviews the approaches that are used in the personality type prediction and classification tasks. It divides methods into four categories: supervised, unsupervised, semi-supervised, and deep learning techniques. A literature review of modern research works for each category is described. The comparison and analysis of results for each category' works are presented.

The third chapter describes all applied experiments one by one, and analyses their limitations. The first experiment of this research work was dedicated to personality classification by applying *k*-Means clustering. It consists of three parts: data collection, data preparation, and implementation of k-Means clustering model with hyper-parameter tuning. To collect data automated Google form was used. By choosing three clusters and inertia was 700. Inertia property is the sum of squared distances of samples to the closest cluster center. The less inertia property, the better work. After hyper-parameter tuning and changing the number of clusters to 16, our inertia parameter was reduced to 107. As the next step after training and hyper-parameter tuning steps, the testing of the model was done by passing some random input data. As a result, it identifies an index of the cluster it belongs to.

The results of experiment №2 showed that people tend to participate in surveys if it is in a native language that is simple to understand questions itself to answer.

In experiment №3 Naive Bayes, XGBoost, and Recurrent Neural Network models were implemented. The performances of classic and deep learning algorithms were compared. The evaluation metrics, such as accuracy, precision, recall, and f-measure were calculated to describe the performances of each model.

The results of experiment № 4 show that CNN gives high performance in images as it has filters and “condition detectors”. As with all work related to machine learning, the limitation of this work is the lack of data. Another thing that should be noticed is that CNN gives some results but does not give us the analysis of features and how they affect prediction itself.

The experiment №5 was performed on Apache Spark on Google clusters by using Google Colab and importing machine learning libraries. To analyze the speed and efficiency of the Multinomial Naive Bayes model a part of experiment №3 was rewritten on Pyspark by using the MapReduce paradigm where labels are the keys and posts are the values.

In experiment №6 text written in Kazakh language was used. As a sample 450 bachelor’s degree students of the Computer Science Department of Suleyman Demirel University were taken and LSTM model was implemented. The results showed an efficiency of proposed model.

The fourth chapter describes the entire architecture of the proposed method for the application of profession inclination identification. Application has two modes to identify MBTI type of person and their inclination to profession. The first mode is from text data and the second one is prediction by uploading image data. The screens of the Android application with the step-by-step illustration are shown in thesis.

The fifth chapter presents testing stages, experimental results, and their comparison between each other and existing works in the field of title of research work.

The conclusion discusses the analysis and outcomes of current work, its future directions. This work contributes to professional inclination identification based on personality traits, and the results of this thesis can be applied to further research works in Face Recognition, Natural Language Processing, and overall Computer Science in Education fields.