Ministry of Education and Science of the Republic of Kazakhstan

Suleyman Demirel University

UDC: 004.8 On manuscript rights

TALASBEK ASSEM LESBEKKYZY

Profession inclination identification using machine learning

6D070400 - Computing Systems and Software

Dissertation submitted in fulfillment of the requirements for the degree Doctor of Philosophy (Ph.D.)

Scientific advisor
Assist. Professor, Ph.D.
Meirambek Zhaparov

Foreign scientific advisor
Assoc. Professor, Ph.D.
Seong-Moo (Sam) Yoo

Republic of Kazakhstan
Kaskelen, 2021

CONTENTS

NORMATIVE REFERENCES

This thesis uses references to the following standards:

- ”Instructions for the preparation of a dissertation and author’s abstract” Ministry of education and science of the Republic of Kazakhstan, 377-3 Zh.
- GOST 7.32-2001. Report on research work. Structure and design rules.
- GOST 7.1-2003. Bibliographic record. Bibliographic description. General requirements and compilation rules.
- GOST 7.32-2017. System of standards of information, librarianship, and publishing. Research report. Structure and design rule

LIST OF ABBREVIATIONS

A -agreeableness
AI - artificial intelligence
API-application programming interface
BiLSTM-Bidirectional Long short-term memory
C-consciousness
CNN - convolutional neural network
CPU - a central processing unit
DL-Deep Learning
E-extraversion
E-extraversion
EmoLex-Word-Emotion Association Lexicon
ERCC- Ensemble of Regressor Chains Corrected
F-feeling
GRU-Gated recurrent units
HDFS -Hadoop Distributed File System
I-introversion
J-judging
LIWC-Linguistic Inquiry and Word Count
LR-Logistic Regression
LSTM - Long-Short-Term memory
MBTI- Myers–Briggs Type Indicator
ML-Machine Learning
MLP-Multilayer Perceptron
MMPI2-Multiphasic Personality Inventory-2
N-intuiting
N-neuroticism
NB- Naive Bayes
NEO PI-R- Revised NEO Personality Inventory
NLP - Natural Language Processing
NN - neural network
O -openness
P-perceiving
RAM- Random-Access Memory
RDD-Resilient Distributed Datasets
RGB -Red, Blue, Green
RNN - recurrent neural networks
S-sensing
SVM - support vector machine
T-thinking
TF-IDF- Term Frequency Inverse Document Frequency
XGBoost- Gradient Boosting

1 INTRODUCTION

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. [1]. 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 [2]. 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. [3]. Some studies show that personality can be taken as an effective measurement in predicting academic performance, especially at the university level [4].

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 [5], NEO-Personality-Inventory Revised [6], Ten Item Personality Inventory [7], Myers- Briggs Type Indicator (MBTI) [8], 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 [9]. 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 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
- Experiments, results, and discussion are provided

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. Two published papers to Q2 journal International Journal of Emerging Technologies in Learning (iJET):

- Talasbek, A., Serek, A., Zhaparov, M., Yoo, S.M., Kim, Y.K. & Jeong, G.H. (2020). Personality Classification Experiment by

Applying k-Means Clustering. International Journal of Emerging Technologies in Learning (iJET), 15(16), 162-177. Kassel, Germany

- Serek, A., Zhaparov, M., Yoo, S.M., Talasbek, A., Kim, Y.K. & Jin, M.W. (2020). Best Practices in Running IT Hackathons Based on Paragon University Dataset. International Journal of Emerging Technologies in Learning (iJET), 15(19), 231-238. Kassel, Germany

Acknowledgments. I would like to express my deep gratitude to Dr. Seong-Moo Yoo, Dr. Meirambek Zhaparov for their guidance, encouragement, my colleagues for their useful criticism, and my family for their support.

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.

The third chapter provides methods**,** describes applied experiments one by one, and analyses their limitations.

The fourth chapter describes the entire architecture of the proposed method for the application of profession inclination identification.

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.

2 LITERATURE REVIEW

According to the National Chamber of Entrepreneurs, 60 percent of graduates do not work in their specialty, as shown in Figure 2.1. The reasons why such a situation has developed in our country turned out to be the lack of demand for a candidate in the labor market, lack of knowledge of one's true inclinations for certain professions, pressure from society in the form of the desires of relatives, as well as other reasons of persons [10].

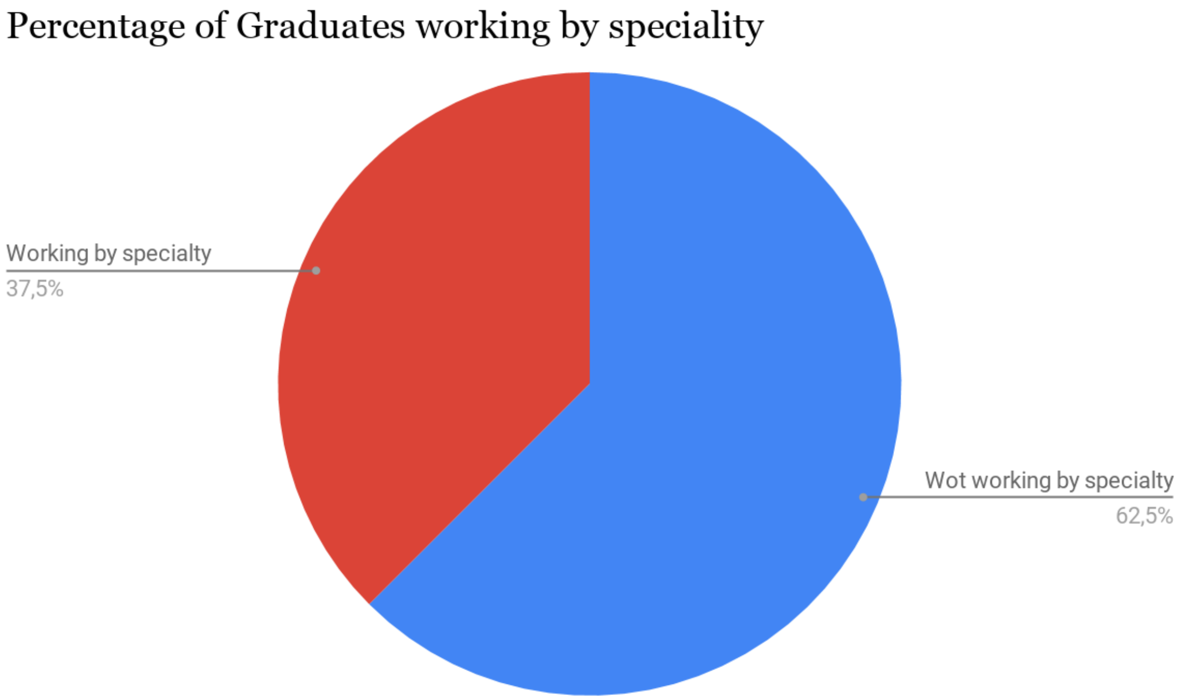
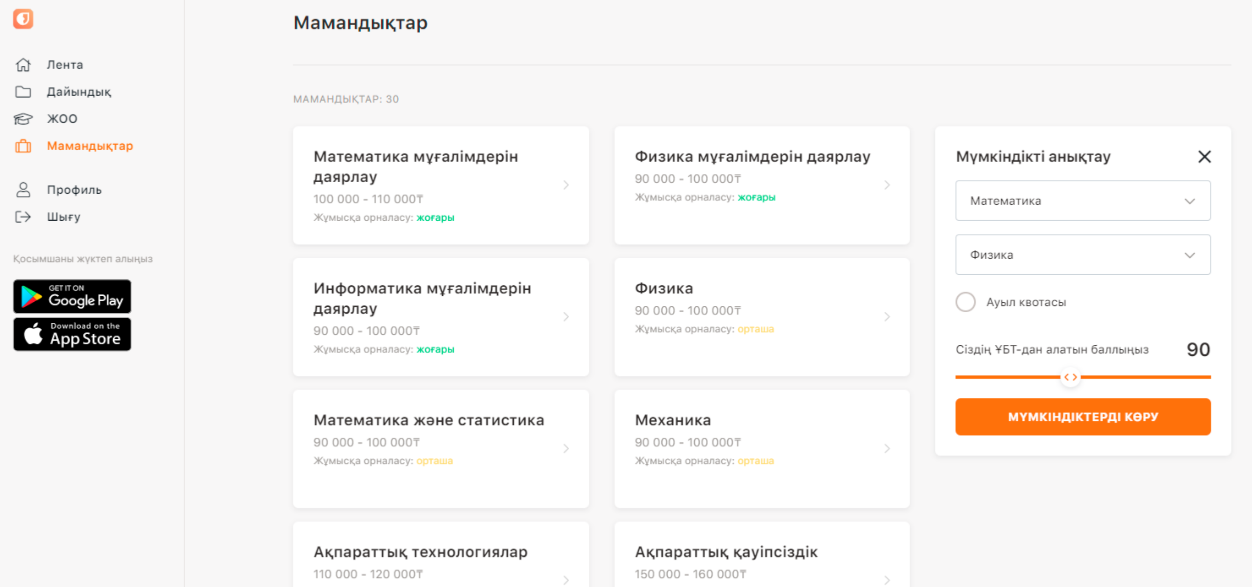


Figure 2.1 - Statistics of graduates [10].

Nowadays a few official web sources and several startups are dedicated to improve the situation on the labor market and help applicants (graduates of school) to choose the right university and specialty to get government grants according to their results of the unified national examination. One of them is www.joo.kz. According to data available in this source in the territory of Kazakhstan, about 105 universities and 103 specialties are available to choose. This source describes each specialty and list of professions available from gaining a particular specialty, the list of universities that provide such specialty diplomas, statistics of wages, and demand for this profession in the labor market [11]. Figure 2.2 represents the screen of the platform.

Another project such as www.vipusknik.kz also provides brief information according to universities and available specialties in these universities. This source provides information more about the territorial location of universities in Kazakhstan and a description of each university by describing their facilities, contact information, etc. [12] Figure 2.3 represents the screen of the web source.

www.Univision.kz is a web source that provides more statistical information about points to get grants to specialty and information about educational programs of the government, university ranking [13]. Figure 2.4 represents the screen of the web source.



Мамандықтар

МАМАНДЫҚТАР: 30

Математика мұғалімдерін даярлау
100 000 - 110 000₸
Жұмысқа орналасу: жоғары

Физика мұғалімдерін даярлау
90 000 - 100 000₸
Жұмысқа орналасу: жоғары

Информатика мұғалімдерін даярлау
90 000 - 100 000₸
Жұмысқа орналасу: жоғары

Физика
90 000 - 100 000₸
Жұмысқа орналасу: орташа

Математика және статистика
90 000 - 100 000₸
Жұмысқа орналасу: орташа

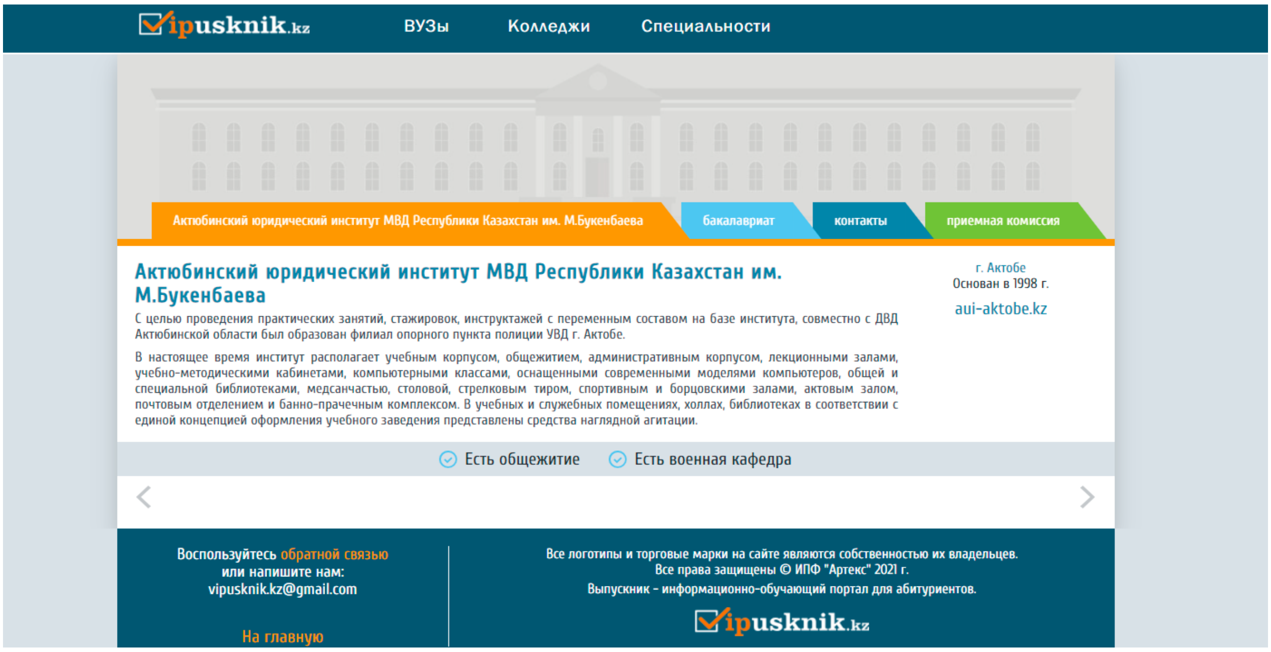
Механика
90 000 - 100 000₸
Жұмысқа орналасу: орташа

Ақпараттық технологиялар
110 000 - 120 000₸

Ақпараттық қауіпсіздік
150 000 - 160 000₸

Мүмкіндікті анықтау
Математика
Физика
○ Ауыл квотасы
Сіздің ҰБТ-дан алатын баллыңыз 90
МҮМКІНДІКТЕРДІ КӨРУ

Figure 2.2 - Screen of www.joo.kz [11].


Vipusknik.kz
ВУЗы
Колледжи
Специальности
Актюбинский юридический институт МВД Республики Казахстан им. М.Букенбаева
бакалавриат
контакты
приемная комиссия

Актюбинский юридический институт МВД Республики Казахстан им. М.Букенбаева

С целью проведения практических занятий, стажировок, инструктажей с переменным составом на базе института, совместно с ДВД Актюбинской области был образован филиал опорного пункта полиции УВД г. Актобе.

В настоящее время институт располагает учебным корпусом, общежитием, административным корпусом, лекционными залами, учебно-методическими кабинетами, компьютерными классами, оснащенными современными моделями компьютеров, общей и специальной библиотеками, медсанчастью, столовой, стрелковым тиром, спортивным и борцовскими залами, актовым залом, почтовым отделением и банно-прачечным комплексом. В учебных и служебных помещениях, холлах, библиотеках в соответствии с единой концепцией оформления учебного заведения представлены средства наглядной агитации.

г. Актобе
Основан в 1998 г.
aui-aktobe.kz

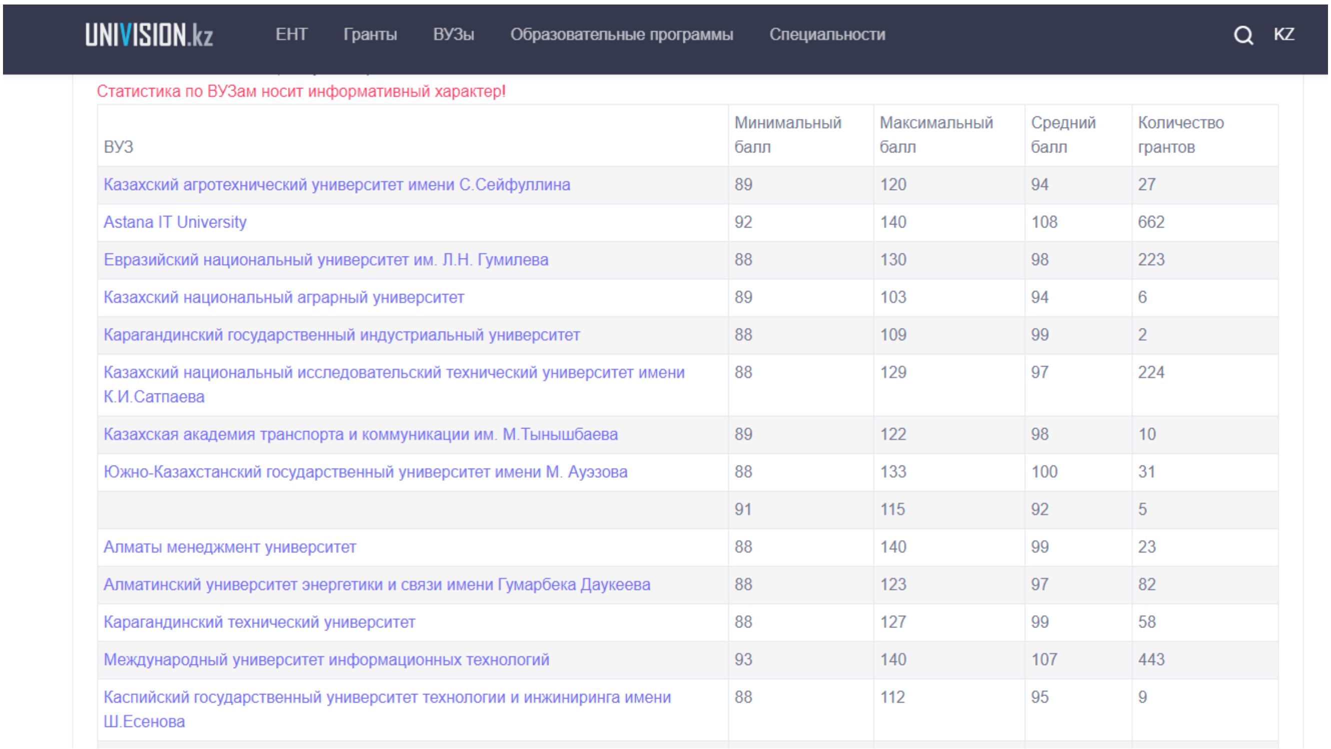
✓ Есть общежитие ✓ Есть военная кафедра

Воспользуйтесь обратной связью
или напишите нам:
vipusknik.kz@gmail.com

На главную

Все логотипы и торговые марки на сайте являются собственностью их владельцев.
Все права защищены © ИПФ "Артекс" 2021 г.
Выпускник - информационно-обучающий портал для абитуриентов.

Figure 2.3 - Screen of www.vipusknik.kz [12].



Статистика по ВУЗам носит информативный характер!

ВУЗ	Минимальный балл	Максимальный балл	Средний балл	Количество грантов
Казахский агротехнический университет имени С.Сейфуллина	89	120	94	27
Astana IT University	92	140	108	662
Евразийский национальный университет им. Л.Н. Гумилева	88	130	98	223
Казахский национальный аграрный университет	89	103	94	6
Карагандинский государственный индустриальный университет	88	109	99	2
Казахский национальный исследовательский технический университет имени К.И.Сатпаева	88	129	97	224
Казахская академия транспорта и коммуникации им. М.Тынышбаева	89	122	98	10
Южно-Казахстанский государственный университет имени М. Ауэзова	88	133	100	31
	91	115	92	5
Алматы менеджмент университет	88	140	99	23
Алматинский университет энергетики и связи имени Гумарбека Даукеева	88	123	97	82
Карагандинский технический университет	88	127	99	58
Международный университет информационных технологий	93	140	107	443
Каспийский государственный университет технологии и инжиниринга имени Ш.Есенова	88	112	95	9

Figure 2.4 - Screen of www.univision.kz [13].

The available current-time web sources provide recommendations to choose specialties and universities for future professions and educational programs. This thesis is dedicated to research experiments that provide revised ways to make recommendations to professions according to the individual characteristics of a person. It means that, graduates of the school chose the profession first, specialty and university next. To suggest a specific profession or career, this research work proposes a method where the inclination and interests of the person, employees, student, or applicant should be considered instead of points scored on a unified national test, which decides whether a person will apply for a specialty or not.

The Slovak psychologist Jan Raiskup writes, “Some professions are associated with types of work that impose increased or specific requirements on the human psyche. Any employee should not be allowed to do such work, without taking into account the appropriate individual prerequisites for this particular profession.” [14]

The study of these issues is very important for the comfortable existence of a person since a significant part of our life is devoted to professional activities. It can even be said that a person's happiness, his/her position in society, material wealth, physical and mental health depend on the correct choice of this activity and its effective implementation. Meanwhile, a practice shows that the first choice of a profession made by a person at a young age, often on a whim, turns out to be unsuccessful. Many do not dare to change their profession, especially if it is associated with obtaining a second higher education, and are forced for decades to do work that they do not like.

A comparison of the personal characteristics of first-year students from various departments of the university, carried out by A.I. Seravin and I.A. Firsova (1999), showed the following. Philosophy students showed characteristics of isolation, introversion, a tendency to reflection, and passivity. History students were distinguished by the same characteristics, as well as judiciousness and aggressiveness. Law students were sociable, extroverted, prone to impulsive, thoughtless conclusions and actions, aggressive, and active. Student-journalists were sociable and frivolous. Psychology students were characterized by the highest sociability, low aggressiveness, and a tendency to think about what they say and do. Thus, there are obvious differences between students: philosophers and historians on the one hand, and lawyers, journalists, and psychologists on the other. Those who choose a profession of the "person-person" type are sociable, while those who have chosen the profession "person-sign" are prone to isolation. At the same time, there are also some personality differences between law students, journalists, and psychologists [14].

This is also confirmed by the data obtained by M.S.Koryagina (2004): psychology students were distinguished by the expressiveness of empathy, and law students were distinguished by their high excitability. They value material well-being, strong will, and in the last places they have values such as "the happiness of others" and "honesty." Among psychology students, “good friends”, “health”, “honesty” and “happy family life” are high in the hierarchy of values. Values such as "strong will" and "material well-being" are not fundamental for psychologists [14].

So, abilities and inclinations are not directly related to each other, but indirectly, through typological features.

The first aspect of career inclination identification study is the conjunction of different factors that affect human beings and their career choices and path. According to work [15], various studies confirm the relationship between psychological type and job type. This idea is based on the nature of psychological type construction that consists of different dimensions, where "an individual can self-select a particular occupation or function" [16]. Different types of questionnaires were developed to suggest career paths such as Minnesota Multiphasic Personality Inventory-2 (MMPI-2) [17, 18], The Sixteen Personality Factor Questionnaire (16PF) [19], The Revised NEO Personality Inventory (NEO PI-R) [20, 21], The Myers–Briggs Type Indicator (MBTI) [22, 23].

MMPI-2 self-assessment test contains more than 500 questions. It includes questions related to the psychological characteristics of a person to questions on political and social topics [17]. MMPI-2 can identify quite a lot of human characteristics, for example, character traits, emotionality, personality needs, leadership abilities, stress resistance, professional qualities, and much more. This test is used in many areas of life, such as health care, education, forensic examinations, recruiting, etc. The psychologist who conducts the MMPI testing must be very professional since the correct compilation of the personality profile is very important [18].

16PF is used to assess personal potential in terms of managerial properties, stress resistance, creativity, communication characteristics, the propensity to take risks, and the level of anxiety. In particular, many applicants face it when passing an interview (including in government and law enforcement agencies). At the moment, several forms of the Kettell multivariate questionnaire have been developed. Their diversity is associated not only with subsequent adaptations and revisions of the "initial" variants but also with the fact that different lists of questions were required for subjects of different ages and with different levels of education. 16PF (version A) consists of 187 questions and assesses personality on 16 factors [19].

NEO PI-R questionnaire is used for Big Five traits assessment (openness, conscientiousness, extraversion, agreeableness, and neuroticism) [20]. The test helps the employer understand the identity of the job seeker or current employee and decide whether the person is suitable for further cooperation. The five-factor personality model allows making predictions whether an employee will interact with the team or choose a single job, realize talents or set up a routine. The test helps to know what character traits of the candidate are suitable for a particular position. The questionnaire consists of 240 questions and is intended for adults (over 18 years old) men and women without mental pathologies [21].

MBTI is formed in three variations: 94 questions, 144 questions, and 167 questions [22]. With the help of the Myers-Briggs typology, it is possible to determine the tendency to the type of human activity, the nature of the solution of questions, and other features of behavior [23].

All of these tests are available on the Internet, and they can be tested online or manually. It means that there are varieties of questionnaires that can be used to determine psychological types of personality and make recommendations of profession according to them, but the problem is that they contain a lot of questions that require a lot of time and other costs. Since our time and the computer technologies of our time allow us to automate these processes to facilitate and improve the quality of life of people, the conjunction of Computer Science and Psychology in the form of a solution using artificial intelligence, machine learning is needed.

As a base for the identification of professional inclination, Myers-Briggs Type Inventory (MBTI) questionnaire was used in this research.

2.1 Carl Jung theory and Myers–Briggs Type Indicator (MBTI)

Myers-Briggs Type Inventory (MBTI) is a personality typology based on Carl Jung's theory - the concept of a psychological attitude, that can be extraverted or introverted and on the predominance of one of the main mental functions - thinking, feeling, sensation, or intuition [22]. According to its authors, the main applied areas of application of the Myers-Briggs typology are self-knowledge and personal growth, career growth and counseling, development of organizations, management, and leadership pieces of training, solution of problems, family consultations, education and curriculum development, and training of interpersonal interaction.

MBTI allows us to determine stable personality traits associated from the point of view of the authors, with temperament, to creation of a holistic type of functional portrait of a person. It is mostly advisable to be used in the selection for positions with a high level of uncertainty of the labor algorithm, that is, those where the role of the employee's personality is great (for example, positions of different levels, positions associated with work in extreme physical or social conditions). Descriptions of personality types in MBTI are quite specific and understandable to a person without psychological preparation [22].

MBTI is designed to identify one of 16 personality types. It includes eight scales combined in pairs. Table 1 shows MBTI personality types.

Table 2.1- Combination of dichotomies [23]

ISTJ	INTJ	ESTJ	ENTJ
ISTP	INTP	ESTP	ENTP
ISFJ	INFJ	ESFJ	ENFJ
ISFP	INFP	ESFP	ENFP

- The extraversion (E) - introversion (I) is the superior direction of MBTI typology. Extroverts concentrate on ideas and events in the outside world, while introverts are more concentrated on internal thoughts. If extroverts like to share their ideas with the outside world, introverts will think about it very carefully before they say or do something. Their interests differ in-depth, but not in breadth. Before discussing a problem, they need to think it over time [23].
- The sensing (S) – intuiting (N) direction is about how a person receives information from the external world. While the intuits clearly see the possibilities and think globally, noticing the consequences and the relationship between events, the sensors can pay attention to details, marking the nuances [22].
- The thinking (T) – feeling (F) direction describes the way of making decisions. Thinkers come to a decision logical and rational, not relying on emotions and sensations. In the process of making decisions, they try to

"rise above the situation", look from the outside and ignore the personal interests of people. Feelers make decisions by following their hearts and tend to be more emotional, soft hearts. An individual approach is important to them, and in the process of making decisions, they put themselves in the place of others and do not accept the same rules “for everyone”. Making decisions is one of the most difficult preferences to recognize [23].

- The judging (J) - perceiving (P) direction reveals that how people orient themselves in the world is related to the way of behavior in conditions of uncertainty. Judgers prefer to organize and make plans to achieve goals, while perceivers solve problems and make choices only when it is necessary. Thinkers love certainty and are result-oriented, while perceivers are process-oriented rather than result-oriented [23, 24].

Table 2.2 represents the functional portraits of different personality types.

Table 2.2 - Functional portraits of MBTI types [14] [23-26]

№	MBTI type	Functional portrait
1	ESTJ	Nature with a highly developed sense of duty. They like to do everything according to a predetermined plan, prefer order and clarity, a set goal. Reasonable and circumspect. Easily get along with people, open-minded, sociable. Very practical, prefer to rely more on reason than on feelings. Easily solve everyday problems and in everyday situations have a clear advantage over representatives of other personality types. In terms of character, it belongs to the administrative type, since it has a symbiosis of such qualities as diligence, responsibility, and the ability to work with people. They are well aware of the mechanisms of the "boss-subordinate" relationship, and they are equally successful in both roles.
2	ISTJ	Calm, restrained, reliable, thorough, and punctual. Somewhat withdrawn and not very emotional. They have a logical mindset, trust reason and facts more than intuition and feelings. Uncommunicative, have few, but loyal and trusted friends. Nature does not like disorganization, uncertainty, fuss. In the working environment, they are inclined to maintain hierarchical relations, preferring clear mechanisms of the “boss-subordinate” relationship, the planned organization of labor.

2.2 - table continuation

3	ESFJ	A practical, sane person possesses everyday sharpness and wisdom, understands a lot about life, and knows how to benefit from many situations. Sociable, open; many friends, acquaintances, try to constantly expand their circle, are cheerful, do not remember offenses for a long time, and their optimism attracts people. They have a distinct commercial ability. Work, service, production - everything is considered in terms of calculations and estimates. Principles that are inconsistent with their vision of the business are of little value.
4	ISFJ	Keepers of traditions, foundations, and well feel the connection of times and the continuity of generations. Calm, restrained. A good owner. Work carefully, thoroughly, somewhat slowly, but reliably. At work, they prefer a clear organization, a stable planned structure. Reliable and responsible performer in all organizational hierarchies. Practical. They have a very developed sense of duty, the desire to take a worthy place in a certain social structure.
5	ESTP	A very energetic and active person, for them life is a game; they are always in search of thrills, in pursuit of risk, whose luck. They have many friends and acquaintances, are experienced in dealing with people, and tend to benefit from relationships. Optimism and wit attract others to them. Monotonous and over-organized activity is not for them. They prefer to work in conditions of risk, on the brink of disaster. However, in any extreme situation, they do not lose composure, the brain gains clarity and calmly looks for a way out.
6	ISTP	People of this type often drop out of school, they are bored, they try to learn everything on their own and do well in this. Monotonous activity requiring punctuality and patience is contraindicated for them. They must realize their energy potential. They are attracted to independence, novelty, and traveling.

2.2 - table continuation

7	ESFP	Optimists, quickly forget about failures, ignore everything gloomy, and go through life with a smile and confidence in the future. Life for them is a continuous adventure, they are interested in everything. Easily seduced, generous, open, and hospitable. Love people, cannot stand loneliness. They quickly and easily find a common language with different people, and they immediately gain confidence in them. This is facilitated by their kindness, cheerfulness, and sense of humor. They perfectly understand the urgent needs of a person, picking up the shades of their feelings and relationships.
8	ISFP	They are very sensitive, impulsive, acutely senses being, current minutes, subtly distinguish between tones and semitones. Independent, restless, seeking to get away from all sorts of restrictions, are ready to do much to achieve personal freedom. They prefer an epicurean way of life, indulges their weaknesses, and tries to get maximum pleasure from life. Love interesting and beautiful people, beautiful things. They are characterized by great sensuality, dependence on mood. They are interested in art: music, painting, but the subtleties of oral and written speech do not particularly attract them. For successful work, they need the appropriate mood, inspiration.
9	ENTJ	A logical mindset, clearly and in detail thinks through goals and ways to achieve them. In any field of activity, they look for patterns, build schemes and models that describe all the variety of connections in the system; develop technology in the most general sense of the word: it can be a technology for communicating with people, playing cards, etc. They do not like ambiguity, if it arises, they must resolve it. Active, self-confident, and power-hungry, strive to be the master of their life. They have the talent of a leader who knows how to formulate tasks for subordinates and to achieve their strict execution. Very efficient and energetic.
10	INTJ	Calm, independent, self-confident, know how to control their emotions and own themselves. Well-developed logical thinking, the ability to analyze and theorize.

2.2 - table continuation

		All their interests are directed to the future. The titles and authority of the position are not essential; their interests are directed towards the search for scientific truth. In their activities, they are attracted by creativity, independence of decision-making, and independence. They work better alone, are far from the problems of relationships in a team, and are not attracted to administrative and organizational work. Able to occupy high management positions that require solving complex problems, but are not related to direct management of the team.
11	ENTP	Enthusiasts that have broad interests in all areas of activity. But they are attracted not by the idea itself, but by its implementation in practice. They do not like banal, routine operations, monotonous work kills them. Interested in everything new, unusual, know how to look at familiar things outside the box, find a zest in everything. Excellent intuition, grasps everything on the fly, quickly understands the essence of technical devices. They are interested in how and where the acquired knowledge can be used. They enjoy the moment when a new idea comes to mind. Sociable, they are interested in people, life in itself, which they seek to make better and more interesting. Often the first to make acquaintances, they have many acquaintances with various interests. A charming companion with a developed sense of humor. This personality type allows one to engage in many interesting professions that require a non-standard approach to solving the problem, as well as the ability to communicate with people.
12	INTP	Intellectuals with a complex inner world, with a desire for in-depth self-knowledge and knowledge of the laws of nature. Intuition is well developed, a synthetic mindset instantly evaluates any situation; possesses an inexhaustible fountain of new ideas. They have a fine sense of beauty and harmony, read a lot, and are interested in art.
13	ENFJ	Possess an original, non-standard mind, good memory, and attention. Sociable, charming, attentive to people, sensitive to their feelings and needs. Cheerful and optimistic, quickly

2.2 - table continuation

		forget offenses. People are always happy with them, they have many friends, and they are the soul of any company. Generous in inventions, never sit idle; know how to organize any events, form groups, distribute roles in a group, create a favorable atmosphere in the team, and take responsibility. Know how to respect someone else's individuality, and have authority with the people around them.
14	INFJ	Read a lot; strive for self-knowledge and self-development of the individual. They have a rich imagination; know how to appreciate a beautiful style, poetry, subtlety of feelings; love metaphors. For them, it is important not only what is said, but also how it is said. Understand all shades of human feelings; for them, first of all, the human soul, the refinement of human relations is important. Easily hurt, need love and recognition.
15	ENFP	Possess an inexhaustible supply of energy, enthusiasm, and optimism. They are interested in everything: interesting people, events, stories, especially attracts everything new and unusual. They approach everything creatively, seeking harmony in the world and people, which somewhat detaches them from reality. They have a rich imagination.
16	INFP	The main thing for them is to be themselves, to have value in their own eyes. All the time they are in search of "the meaning of life", looking for a hidden meaning in all-natural phenomena. Spirituality is a property that is inherent in them to the highest degree.

Paul D. Tieger and Barbara Barron-Tieger in their study describe the importance of personality type identification and provide information about how it can help in the career counseling field [27]. Many studies have conducted that most ENTJs “prefer to be elementary teachers or work in related fields”, ESTJs are the best sellers, and INTJs are good in accounting [26] [28] [29].

According to the studies [23-26] those who define themselves as "I" more prefer to work as counselors, if as ‘‘E’’ more prefer to be executive coaches.

In the study about teachers and their MBTI types [30], Lawrence conducted that the combination of Extraversion, Sensing, Feeling and Judging

(ESFJ) is more frequently common and preferable rather than a combination of other dichotomies. Table 2.3 represents the professions according to MBTI typology.

Table 2.3 - Professions according to MBTI [14] [31].

MBTI type	Description	Profession
ESTJ	“Administrator”	manager; secretary; forwarder; state employee; banker; accountant; controller; tax inspector; administrator.
ISTJ	“Guardian”	clerk; accountant; tax inspector; controller; Bank employee; economist; serviceman; state employee; operator; corrector.
ESFJ	“Consul”	businessman; seller; broker; provider; broker; financier; traveling salesman; conductor; advertising worker; trading floor administrator; commercial director; waiter.
ISFJ	“Defender”	civil servant; tax inspector; the prosecutor; controller; engineer; technical manager; accountant; cashier; bank employee; expert criminalist; judge; clerk; serviceman; archive
ESTP	“Entrepreneur”	rescuer; sailor; trainer; tester; geologist; businessman; investigator; policeman; fireman; driver.
ISTP	“Virtuoso”	firefighter, polar explorer, tester, driver.
ESFP	“Entertainer”	business; trade; organizational, administrative, cultural work.
ISFP	“Artist”	artist; conductor; fashion designer; musician; designer; tailor; jeweler; the hairdresser; artist; sculptor; photographer; art critic.
ENTJ	“Commander”	administrative, military activities.

2.3 - table continuation

INTJ	“Scientist”	physicist; mathematician; chemist; biologist; chess player; archaeologist; philologist; prospector; engineer.
ENTP	“Inventor”	doctor; constructor; magician; teacher; educator; programmer; administrator; practical psychologist; businessman; engineer; inventor.
INTP	“Architect”	mathematician; architect; philosopher; biologist; archaeologist; geneticist; historian; builder; constructor; fashion designer; designer; gardener; restorer.
ENFJ	“Educator”	educator, teacher, trainer, tutor, practical psychologist. If desired, they can also succeed in politics.
INFJ	“Writer”	doctor; writer; psychologist; producer; screenwriter; literary critic; theater expert; film critic; poet.
ENFP	"Journalist"	journalist; businessman; guide; politician; manager; teacher; playwright; cultural worker; sociologist; translator.
INFP	“Mediator”	philosopher; philologist; historian; architect; psychologist; a biologist, but not a businessman.

2.2 Personality Classification using Machine learning algorithms

Background research of personality classification tasks using machine learning (ML) algorithms shows that it can be divided into four groups: supervised algorithms, unsupervised algorithms, semi-supervised, and deep learning techniques [32].

The supervised learning method classifies and predicts personality type using labeled data (independent variable) from which algorithms learn their futures. In works [32-35], authors tried to predict personality types using MBTI and text data by getting posts of users and their labels from social networks such as Twitter and Reddit and faced various limitations. The feature vectors were constructed using TF/IDF, LIWC, and EmoLex.

TF/IDF is a term frequency-inverse document, where frequency is one of the statistical values that measures how the word is relevant to a specific document in a set of documents. To measure the relevance of the word, the multiplication of two metrics such as frequency of the word that appears in a document (term frequency) and the word's inverse document frequency over a collection of documents (inverse document frequency) are used.

Linguistic Inquiry and Word Count (LIWC) is a technique to calculate the percentage of words in a given text that belongs to a specific category. For instance, LIWC can be used to identify the positive or negative degree of emotion in a given text. By using LIWC, users can create their dictionary [36].

Word-Emotion Association Lexicon (EmoLex) is a tool that contains categories such as positive and negative sentiments and emotions such as anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. EmoLex has a huge impact on Data Science, Natural Language Processing, Psychology, and other fields [37].

In work [32] after preprocessing text by using methods mentioned above, the Support Vector Machine (SVM) algorithm produced a high accuracy score across all dimensions rather than Neural Net and Naive Bayes.

SVM is the one of supervised learning algorithms used for classification and regression problems. The main goal of SVM is to create a line (decision boundary) in n-dimensional space to classify given data into correct categories [38]. Extreme points are also called support vectors and are used to create the best line which is also called the optimal hyperplane shown in Figure 2.5.

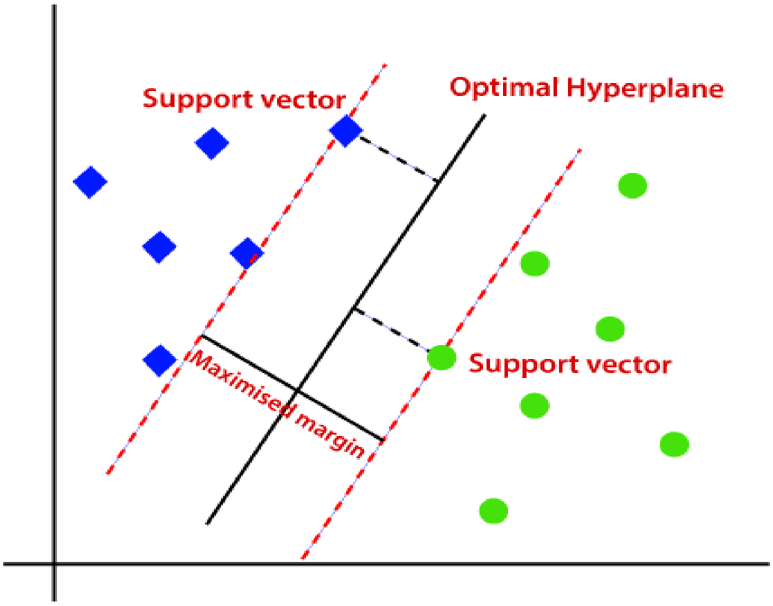


Figure 2.5- Support Vector Machine [38].

In work [33] authors collected the MBTI dataset from the Reddit social network and applied SVM, Logistic Regression, and Multilayer Perceptron (MLP) algorithms. The best result, which is 42% of accuracy, is achieved by the MLP algorithm.

MLP is the class of feed-forward algorithms that has three layers: input, hidden, and output as shown in Figure 2.6. The data in MLP flows from input to output layers. An arbitrary number of hidden layers of MLP is the place where the main computation happens. MLP is mostly applied in tasks such as classification, regression, approximation, and prediction [39]. The major limitation of this paper is a large number of words in posts that are not informative in terms of personality prediction tasks.

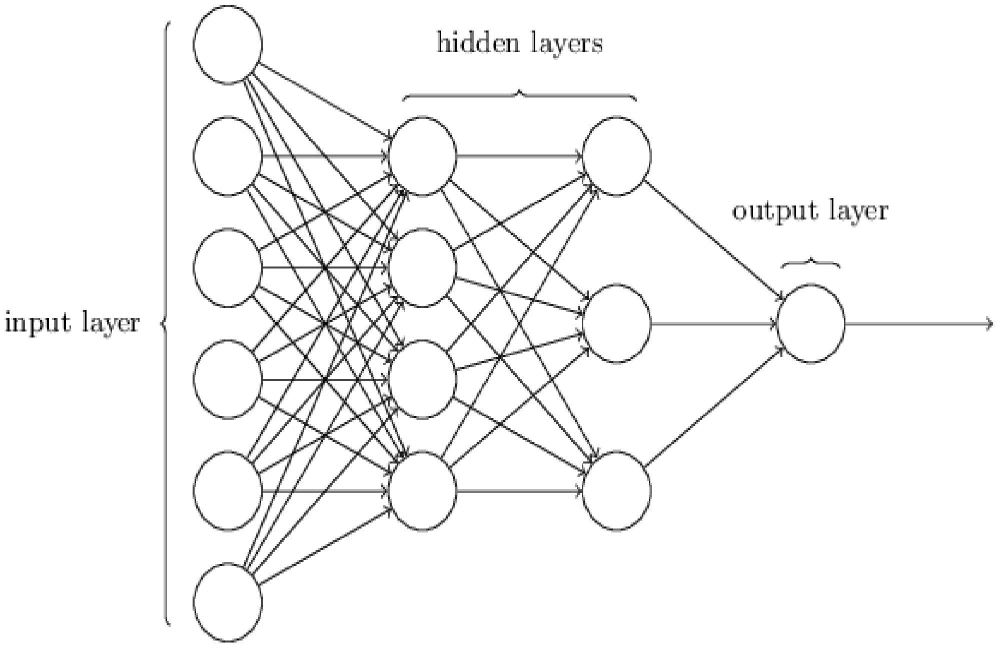


Figure 2.6: MLP structure [40].

In work [41], as a feature selection technique authors used binary word n-gram, a model based on Logistic Regression, and showed improvement in I-E and T-F dimensions while in other dimensions slightly dropped. Logistic Regression is the algorithm that uses probabilistic values from 0 to 1 to classify given labeled data into classes [42]. As a base logistic regression uses sigmoid function shown in Figure 2.7 that is calculated as follows:

$$S(z) = \frac{1}{1-e^{-z}} \tag{2.1}$$

where:

- S is probability estimate;

- *e* is base of natural log; and
- *z* is input to the function.

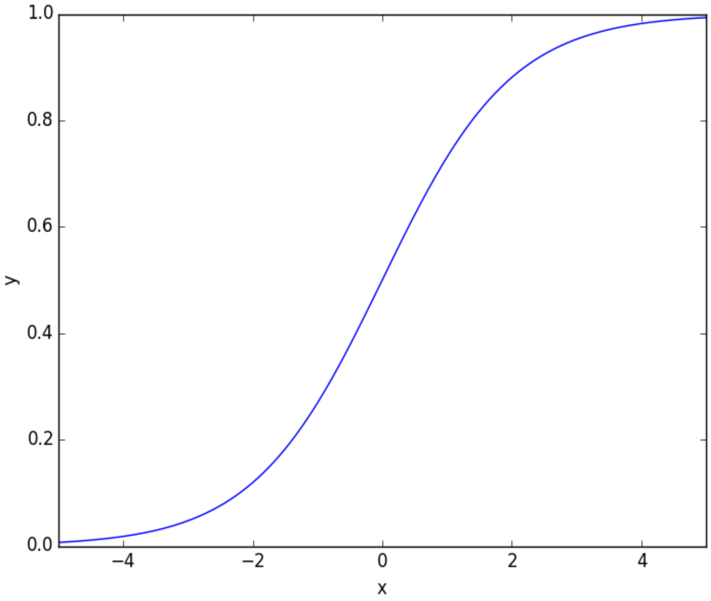


Figure 2.7 - Sigmoid function [43].

The performance of various algorithms such as Naive Bayes, SVM, Logistic Regression (LR), Random Forest was evaluated in [44], and the authors predicted that MBTI type from the online text and again LR algorithm showed higher accuracy results rather than other algorithms which are exactly 66.5% for all MBTI types.

In the study [45] by using the MBTI type indicator, researchers presented TwiSTy, a corpus of tweets for gender and personality prediction that covers Dutch, German, French, Italian, Portuguese, and Spanish languages [45]. “Linear SVM is used as a classifier and results are also tested on Logistic Regression”, [42]. For gender and personality prediction, it outperformed other techniques for two dimensions: I-E and T-F, but for other dimensions, no improvement was shown. Table 2.4 represents a brief literature review summary of later works that belong to the personality prediction field by using supervised methods.

Table 2.4 - Summary of works based on supervised methods for personality prediction [46].

#	Research paper	Authors	Aim and objectives	Methods	Results and Limitations
1	“Persona Traits Identification based on Myers-Briggs Type	S.Bharadwaj, S. Sridhar, R. Choudhary	Personality prediction from text	SVM, Neural Net and Naïve	SVM showed better results than other

2.4 - table continuation

	Indicator(MBTI) - A		data	Bayes	algorithms
	Text Classification Approach, 2018” [47]	and R. Srinath		TF-IDF, Emolex, LIWC, ConceptNet	The limitation of the research was less weightage of the gravity of the word
2	“Reddit: A Gold Mine for Personality Prediction, 2018” [48]	M. Gjurković and J. Šnajder	Personality classification of Reddit user’s posts	SVM, Logistic Regression and MLP with linguistic features	MLP showed better results than other algorithms. The results of the T-F dichotomy should be improved.
3	“Personality traits on Twitter—or—how to get 1,500 personality tests in a week. 2015” [49]	B. Plank, and D. Hovy	Personality and gender prediction from tweets.	Logistic regression Model and Binary word n-gram	Accuracy for personality prediction: I/E = 72.5% S/N = 77.5% T/F = 61.2 % J/P = 55.4%
4	"Personality classification based on Twitter text using Naive Bayes, KNN and SVM, 2015 “[50]	B. Y. Pratama and R. Sarno	Personality type prediction based on Big-5 model from tweets posted in English and Indonesian lang.	KNN, Naive Bayes, SVM	Accuracy KNN = 58% NB = 60% SVM = 59% Limitation: lack of data

2.4 - table continuation

5	"Personality prediction based on Twitter information in Bahasa Indonesia," 2017 [51]	V. Ong et al.	Personality type prediction based on Big5 Model for Bahasa Indonesian Twitter users	XGBoost SVM	Accuracy XGBoost = 97.99% SVM = 76.23% Limitation: lack of data
6	"Personality traits recognition on social network-facebook," 2013 [52]	F. Alam, E. A. Stepanov and G. Riccardi	Prediction based on Big 5 model using Facebook profiles	Multinomial Naive Bayes, Logistic Regression (LR), and SMO for SVM	MNB = 61.79% BLR = 58.34% SMO = 59.98% MNB outperformed other methods.
7	"Towards User Personality Profiling from Multiple Social Networks, "2016 [53]	K. Buraya, A. Farseev, A. Filchenkov, and T. S. Chua	Usage of NUS-MSS for personality profiling.	Supervised	The concatenation of various data sources in one feature vector, improved performance by 17%.
8	"A comparative study of different classifiers for automatic personality prediction," 2016 [54]	N. R. Ngatirin, Z. Zainol and T. L. C. Yoong	Personality prediction using Twitter	Naïve Bayes, Simple logistic, SMO, JRip, OneR, ZeroR,	OneR with F1_Score = 0.837 outperform among all.

2.4 - table continuation

				J48, Random Forest, Random Tree, and Ada BoostM1	
9	"A Comparative Study of Different Classifiers for Myers-Brigg Personality Prediction Model," 2018 [55]	S. Chaudhary, R. Sing, S. T. Hasan and I. Kaur	Prediction of personality based on MBTI from online text	Naive Bayes, SVM, LR and Random Forest	NB = 55.89% LR = 66.59% SVM = 65.44%
10	“Comparing the Behavior of Oversampling and Under sampling Approach of Class Imbalance Learning by Combining Class Imbalance Problem with Noise,” 2018 [56]	P. Kaur and A. Gosain	Imbalanced dataset for comparison of oversampling and under sampling techniques	Decision tree algorithm C4.5	The oversampling method is better than under sampling methods
11	“Exploring Personality Prediction from Text on Social Media: A Literature Review,” 2017 [57]	V. Ong, A. D. Rahmanto, Williem and D. Suhartono	Personality type classification	Review of all papers related to personality prediction from text data	Limitations: less amount of data and features.

2.4 - table continuation

12	"Predicting Personality from Twitter," 2011[58]	J. Golbeck, C. Robles, M. Edmondson and K. Turner	Prediction of personality type from Twitter users posts based on Big Five model	ZeroR and GP	Higher for Open = 75.5% Lower for Neuro =42.8% Limitation: lack of data
13	"Our Twitter Profiles, Our Selves: Predicting Personality with Twitter," 2011[59]	D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft	connection of BigFive and Twitter user’s account	M5 rules with 10-fold cross-validation.	O = 0.69 C = 0.76 E = 0.88 A = 0.79 N = 0.85
14	“Twisty: a multilingual twitter stylometry corpus for gender and personality profiling” 201[60]	B., Verhoeven, W. Daelemans and B. Plank	Prediction of gender and personality using TwiSTy corpus	SVM and logistic Regression	F1 score I/E =77.78 S/N =79.21 T/F = 52.13 J/P = 47.01 for Italic language

Unsupervised learning allows algorithms to act on given data without any labels (independent variables) and additional guidance. There are a few works related to unsupervised approaches to classify people into personality types categories from input text mostly based on the Big Five model. In work [] researchers predicted personality types from Twitter posts by unsupervised score-based approach and achieved a mean accuracy score of 0.665 and mean validity of 0.699. The authors claimed that additional data from Twitter may improve the results.

In [61] unsupervised Adawalk strategy was used for group-based personality identification and as an outcome, Macro F1 score is above 97.74% and again authors claim that increased dataset will improve the performance of the model.

In [62] to recognize the network’s visitors' personality, an unsupervised *k*-means clustering algorithm was used. As a result, *k* = 10 is a more accurate score and a greater number of websites will give better performance of the model.

Findings in [63] are based on English text from Twitter and may be performed in other languages too. For word-embedding, the GloVe model is used from user tweets, and the accuracy of the model is equal to 68.5 %. Table 2.5 represents the summary of the research works in personality prediction areas performed by applying unsupervised machine learning algorithms.

Table 2.5- Summary of works based on unsupervised methods for personality prediction [46].

#	Research paper	Authors	Methods	Results	Limitation s
1	"Mining user personality in twitter", 2011 [64].	F.Celli	Score-based personality classification from individual's writing pattern	Mean Accuracy =0.6651 Mean validity= 0.6994	Lack of data
2	"Group-level personality detection based on text generated networks", 2019 [65]	X.Sun et al.	personality identification by applying Adȧwalk	F1 97.74%	Lack of appropriat e data
3	"The role of emotional stability in Twitter conversations", 2012 [66]	F. Celli and L. Rossi	Statistical impact of linguistic characteristics on personality identification	Accuracy 78.29%	Lack of data
4	"Identify Website Personality by Using Unsupervised Learning Based on Quantitative Website Elements", 2015.[68]	S.Chishti and A. Sarrafzadeh	Personality identification of network visitors' personality by applying k-means clustering	optimal K=10	Needs addition of new elements of websites

2.5 – table continuation

5	“Unsupervised personality recognition for social network sites”, 2012 [69]	F. Celli	score-based Impact of linguistic characteristics on Big Five traits	Accuracy 81.43%	Needs the extension of corpus
6	“25 Tweets to Know You: A New Model to Predict Personality with Social Media,” 2017[70]	P.H.Arnoux et al.	Prediction of Big-Five traits by using word-embedding	Accuracy 68%	tested only on English text

Semi-supervised ML methods are experiments that use a mix of linguistic and lexicon features, supervised machine learning methodologies and different feature selection algorithms. Authors of research [71] proposed a multilingual predictive method to identify personality type, age, and gender from Twitter accounts. For identification of age and gender, they have used an SGD classifier, while for personality prediction LIWC with regression model they have used ERCC.

ERCC, which stands for Ensemble of Regressor Chains Corrected, is a multivariate regression method that uses the result of prediction of the previous trait to predict the next trait. The main disadvantage of ERCC is the randomness of the chain sequence determination that causes differences in the prediction performance of the model [72]. As a result of [71], 68.5% of accuracy was achieved, and the model can be enhanced to other languages.

Lexicon-based and linguistic rule-driven machine learning approach was used to predict personality traits from Twitter in the Indonesian language in research [73]. 2,500 tweets per user were analyzed among 97 selected user accounts. The collection of machine learning algorithms as WEKA was used to classify and train datasets. The achieved accuracy for the I-E trait is equal to 80%, for S-N, T-F and J-P accuracy is 60%. Authors claim that by increasing the dataset, accuracy also will be improved.

In work [74] by using only the Word count approach on text data from social media, higher accuracy for the "openness" trait of Big Five was achieved, while for MBTI, accuracy for the S-N dimension is greater than all dimensions. Authors claim that by adding machine learning algorithms and other features results can be improved in the future. Table 2.6 represents the summary of the research works in personality prediction areas performed by applying semi-supervised methods.

Deep learning is the branch of machine learning where algorithms learn without human interaction, by training themselves on unstructured unlabeled data, performing tasks repeatedly, and improving results on each iteration. A few works using deep learning were carried out.

In research [75] deep learning approaches such as Recurrent Neural Network (RNN), Long short-term memory (LSTM), Gated recurrent units (GRU), Bidirectional Long short-term memory (BiLSTM) models were tested in real-life examples such as Donald Trump's Twitter account to predict his actual MBTI type. Various RNN layers were tested, after comparing all applied algorithms. LSTM is the type of RNN that is capable to learn long-term dependencies which show better performance. But still, results need to be improved, after classifying achieved accuracy is very low and equal to 0.028.

Various classification methods such as Softmax as the baseline, SVM, Naïve Bayes, and deep learning algorithms were applied in work [76]. Deep learning methods showed better results than other classic algorithms, but accuracy is still lower than 50%. In studies [77,78] based on Big Five theory, the Convolutional Neural Network (CNN) model was implemented to classify personality traits in both kinds of research and "openness" traits gave more than 50% of accuracy which is better than other traits.

More features and data set improvements are needed in all of these research works. Table 2.7 represents a summary of research works that applied deep learning methods to predict and classify personality traits.

Table 2.6 - Summary of works that applied semi-supervised methods for personality identification

#	Research paper	Authors	Methods	Results	Limitations
1	“Age, gender and personality recognition using tweets in a multilingual setting”, 2015 [71]	M.Arroju et al.	SGD classifier with n-gram features, LIWC with ERCC	accuracy above 68%	accuracy can be increased by using various personality models
2	"Social media user personality classification using computational linguistic", 2016 [73]	L.C. Lukito et al.	Lexicon Based and linguistic rules-driven ML algorithms on data from Twitter	I/E trait = 80%	Limitations of corpus

2.6 - table continuation

3	"Estimating Personality from Social Media Posts", 2017[74]	N. Alsadhan and D. Skillicorn	WordCount method to predict personality from social media	Big Five "openness" trait only with good results, MBTI S/N dimension	only WordCount approach was used, should add machine learning algorithms

Table 2.7- Summary of works that applied deep learning methods for personality identification.

#	Research paper	Authors	Methods	Results	Limitations
1	"Predicting Myers-Briggs type indicator with the text", 2017 [75]	R.K. Hernandez and L. Scott	RNN, LSTM, GRU, BiLSTM on text data	I/E= 67.6% S/N=62.0% T/F=77.8% J/P=63.7%	lack of data
2	"Survey Analysis of Machine Learning Methods for Natural Language Processing for MBTI Personality Type Prediction", 2018 [79]	B. Cui, and C. Qi	Multilayer LSTM	accuracy= 38% I/E= 89.51% S/N=89.84 % T/F=69.09 % J/P=69.37 %	improvement of word embedding techniques

2.7 - table continuation

3	"Deep learning-based personality recognition from text posts of online social networks", 2018[76]	D. Xue et al.	Deep learning algorithms to predict BigFive	O= 0.3577 C= 0.4251 E= 0.4776 A= 0.3864 N= 0.4273	improvement of performance of model is needed
4	"Deep Learning Based Document Modeling for Personality Detection from Text ", 2017 [80]	N.Majumder et al.	CNN to BigFive traits	O= 62% C= 56% E= 58% A= 56% N= 59%	Addition of LSTM is needed to improve performance of model.

3 PERSONALITY CLASSIFICATION EXPERIMENTS

After summarizing all previous studies, six experiments were conducted in this research work to perform better personality classification as this is one of the basic steps in the profession inclination identification task. Here, conducted experiments, one by one, will be explained. As a base of each experiment, MBTI typology, BigML service [81] was used.

3. 1 Experiment №1

The first experiment of this research work was dedicated "to personality classification by applying *k*-Means clustering"[82]. Figure 3.1 represents the general view of experiment №1.

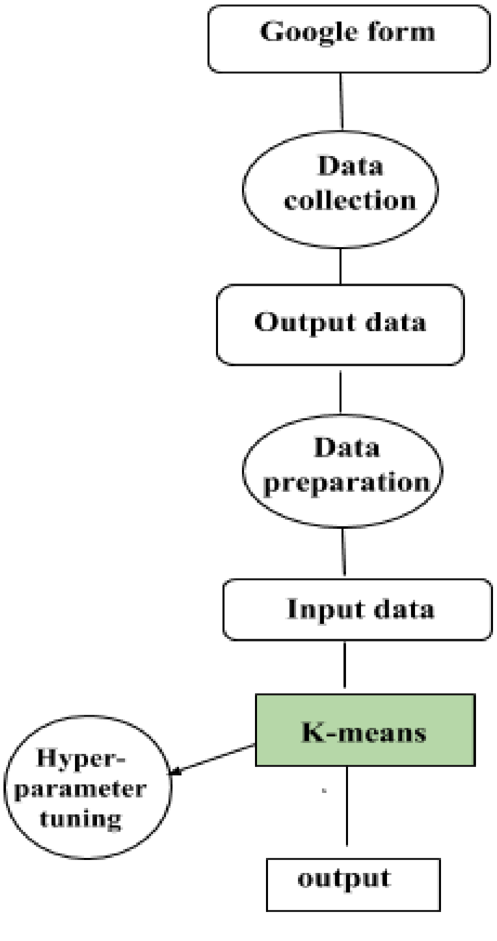


Figure 3.1- Personality classification by applying k-Means clustering [82].

As shown in Figure 3.1 experiment №1 "consists of three parts: data collection, data preparation, and implementation of k-Means clustering model with hyper-parameter tuning"[82].

3.1.1 Dataset

To experiment №1, first, we collected data. By using Google Script we have created an automated Google form that collects data from responders, calculates their MBTI personality type, and generates answers of questionnaires as a portable document format (pdf) file to send the results to each responder one by one via email automatically.

For each of four dichotomies of MBTI typology, there are allocated five questions resulting in an overall number of 40 questions with "Yes" and "No" answers. For instance, questions like "Do you communicate openly without censoring?" show a level of extraversion or introversion. Statements like "You show evidence (e.g., facts, details, examples, etc.)." belong to sensing-intuiting traits that identify the type of obtaining information from the outside world, etc. Table 3.1 shows the details of dataset 1.

Table 3.1- Dataset 1.

Timestamp	ID	1. Do you respond quickly without long pause to think?	2. Do you communicate openly without censoring?	3. Do you show energy and enthusiasm?	…
11/5/2018 12:42:03	16495234	Yes	Yes	No	...
11/5/2018 12:46:23	38494057	Yes	No	No	...

“After the data preprocessing stage, we perform some simulations by applying the k-Means clustering algorithm on a given data. The main limitation of this research is that collected data did not give us exact output because more than 35% of participants found survey questions difficult to answer and the results were controversial. As an instance, one person may belong to an introvert and extravert 50% to 50% or several types at once.

As the next stage to resolve this issue we made some changes in the questionnaire by using Likert scale from 1 to 5 instead of “Yes” and “No” answers, where 1 is "strongly disagree" and 5 is "strongly agree". As a sample, we took 105 bachelor's degree students of the Computer Science Department of Suleyman Demirel University”[82]. Table 3.2 represents the details of dataset 2.

Table 3.2- Dataset #2

Timestamp	ID	Name & Surname	email	Do you respond	Do you communi	...

3.2-table continuation

				quickly without long pause to think?	cate openly without censoring?	
11/23/2018 10:17:18	160103035	…	…@stu.sdu.edu.kz	2	4	...

3.1.2 k-Means clustering

“Given that we did not have labels (independent variables) to train our model, we decided to use unsupervised learning. Out of all unsupervised learning algorithms, we choose the k-Means clustering algorithm. Figure 3.2 illustrates how the k-Means clustering algorithm works. K-Means clustering continue by switching back and forward between *assignment* and *update* steps. Using this algorithm, we cannot be sure that the most optimal result will be found. The convergence of algorithms can be stopped by using various functions, for instance, by applying Manhattan distance.

- *The assignment step* of k-Means clustering sets each observation to the cluster with the smallest Euclidean distance to present how observations are close to each other.
- *The update step* of k-Means clustering calculates the new mean value of clusters that will become centroids in the new clusters.

Manhattan distance is calculated as follows:

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i| \qquad (3.1)$$

where:

- x and y are data points;
- n is the number of dimensions.

Manhattan distance is also called City Block Distance or *L1* norm, *L1* distance in machine learning which is used for regularization problems.

Euclidean distance is calculated as following:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}((x_i - y_i)^2} \qquad (3.2)$$

where:

- x and y are samples in n-dimensional feature space.

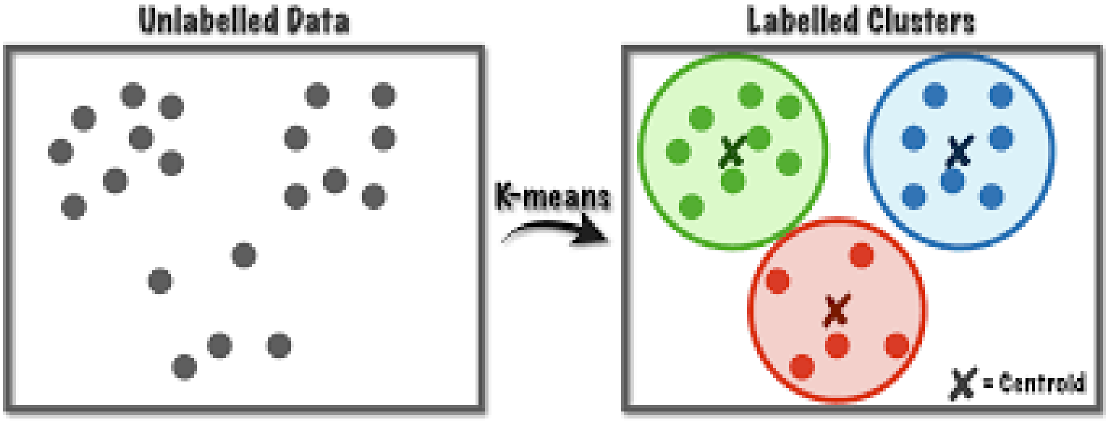


Figure 3.2 - k-Means clustering.

To simplify the process of the machine learning to apply the k-Means algorithm on our dataset, we used BigML.com [81] platform that automates the various process, for example, the process of data processing, analysis parts, machine learning, supervised and unsupervised learning parts.

In terms of data preprocessing, the values of categorical variables have been transformed from textual representation into numeric one, so that the k-means algorithm will be able to work with it. As for hyper-parameter tuning, we used the same infrastructure of BigML.com to find the most optimum inertia value. Inertia value is one of the most important parameters of k-means, which defines how far the data points are from their centroids. That can be manifested itself as a standard deviation regarding the mean notion. Therefore, the corollary of that is the less is the inertia, the better is our k-means model in terms of its cluster division" [82].

3.1.3 Results

"After collecting dataset 2, to make personality type prediction by classifying data into 16 personality types, we performed a second simulation using the k-Means clustering algorithm. We started from three clusters and our inertia was 700. Inertia property is the sum of squared distances of samples to the closest cluster center. The less inertia property, the better our 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, we started to test our model bypassing some random input data. As a result, it identifies an index of the cluster it belongs to. Table 3.3 represents the results of experiment №1.

Table 3.3 - Cluster instances based on collected data [82].

Cluster name	Instances	%	Mean distance
Cluster 00	30	28.6	0.133
Cluster 01	5	4.72	0.140
Cluster 02	9	8.49	0.144
Cluster 03	4	3.77	0.147
Cluster 04	2	1.89	0.098
Cluster 05	3	2.83	0.100
Cluster 06	5	4.72	0.111
Cluster 07	1	0.94	0
Cluster 08	17	16.04	0.140
Cluster 09	2	1.89	0.103
Cluster 10	2	1.89	0.103
Cluster 11	3	2.83	0.118
Cluster 12	5	2.83	0.113
Cluster 13	16	15.09	0.138
Cluster 14	1	0.94	0
Cluster 15	3	2.83	0.098
Total	106	100	0,164

According to the results of experiment №1, the most respondents of the survey belong to cluster 00, which is 28.6 percent (30 participants). The smallest number of participants, only one participant in each cluster belongs to clusters 07 and 14 which is 0.94 percent of the total, and the mean distance to centroids of clusters 07 and 14 is equal to 0. Figure 3.3 illustrates 16 clusters that were obtained by applying k-means clustering on data from the survey” [82].

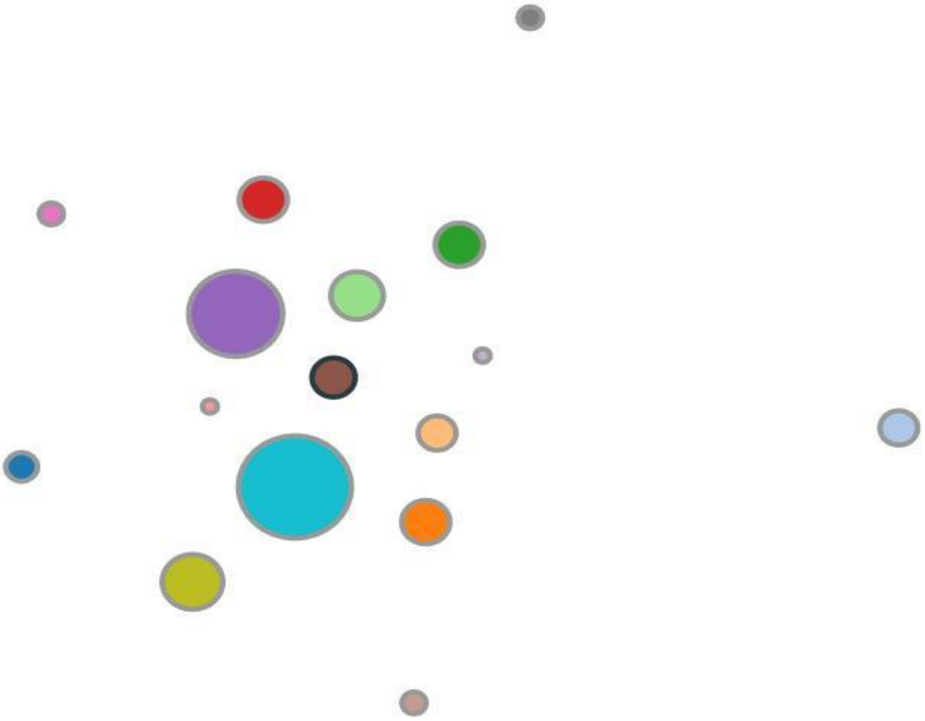


Figure 3.3 - The results of experiment №1 [82]

As shown in Figure 3.3, in general, clusters are close to each other, which means that people who participated in experiment №1 have similar personality traits, but still, there are some outliers - small clusters that are placed far from others.

The main limitation of experiment №1 is that the output of the k-Means clustering algorithm gives us a particular cluster to which person belongs without a description of personality. It means that the help of additional professionals as psychologists are needed to identify personality type according to MBTI typology.

Table 3.4 presents the association analysis, which shows the relationship between statements such as "how you communicate" and "what you are communicating". Here, we found that if a person is brief and concise, the person is often calm and reasonable.

Table 3.4- Association analysis [82]

Antecedent	**Consequence**
You aware that how you communicate is as important as what you're communicating > 4	You take care about how the idea will affect people and what people's reaction would be > 4

3.4-table continuation

You take care about how the idea will affect people and what people’s reaction would be > 4	You aware that how you communicate is as important as what you’re communicating > 4
You are brief and concise > 4	You are often calm and reasonable > 4
You are often calm and reasonable >4	You are brief and concise >4

3. 2 Experiment №2

After the conduction of experiment №1, we decided to change the questions of the survey into a minimized full version of the MBTI test adapted to the population of post-soviet Union countries.

To experiment №2 new "pclassification@gmail.com" google mail account was created. Google form was automated by using Google ads on features, to calculate and send the results of surveys to participants. The survey of experiment №2 is still available by link: https://forms.gle/ef9HXZk4H6H2YSEJ9 More than 400 respondents participated in a survey of experiment №2. And k-means clustering algorithm was used to classify the gathered data into 16 classes as in experiment №1.

3.2.1 Dataset

The survey of experiment №2 consists of 167 questions such as "What will you do if ...?". Participants were asked to choose adjectives according to their feelings and to describe themselves by using them. The questions were asked in the Russian language to see how the language itself will affect the number of participants in the survey.

Additionally, participants were asked to provide some text about themselves (10 sentences), their personal Instagram accounts, if they agree to use their data in our future experiments related to this topic. Table 3.5 represents the detailed view of dataset 3.

Table 3.5 -Dataset 3.

Отметка времени	Адрес электронной почты	ID	Ссылка на ваш Instagram	Напишите 10 предложении о себе и своих увлечениях. (на казахском используя кириллицу)	Укажите ваш пол	Какой ответ ближе всего подходит для описания того, как Вы обычно думаете, чувствуете или действуете?
11.12.2019 23:55:14	.. @stu.sdu.edu.kz	….	...	рисование, дизайн, маркетинг, экономика, биржа	б) женский	а) привлекает Вас

3.2.2 Results

Table 3.6 and Figure 3.4 represent the results of applying k-means clustering on collected data from survey 2.

Table 3.6 - The results of experiment №2 [82]

Cluster name	Instances	%	Mean distance
Cluster 00	52	12.7	0.043
Cluster 01	58	14.15	0.043
Cluster 02	57	13.9	0.038
Cluster 03	40	9.7	0.043
Cluster 04	24	5.8	0.044

3.6 - table continuation

Cluster 05	25	6.10	0.046
Cluster 06	12	2.9	0.042
Cluster 07	26	6.34	0.042
Cluster 08	21	5.12	0.046
Cluster 09	9	2.19	0.044
Cluster 10	11	2.68	0.045
Cluster 11	10	2.44	0.045
Cluster 12	10	2.44	0.045
Cluster 13	16	3.90	0.045
Cluster 14	8	1.95	0.043
Cluster 15	31	7.56	0.045
Total	410	100	0.049

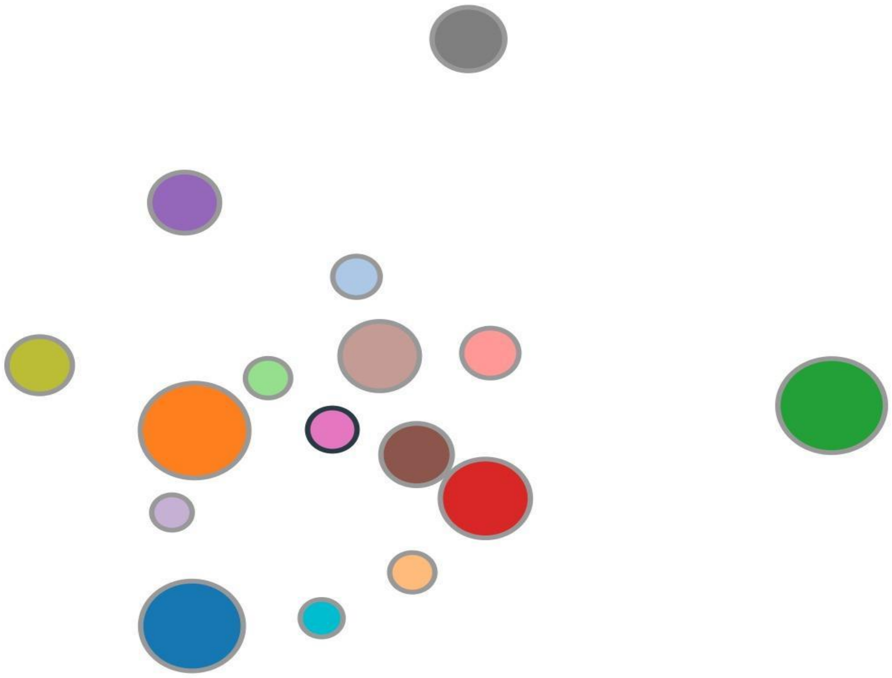


Figure 3.4- The experiment №2 results [82].

The results of experiment №2 show that people tend to participate in surveys if it is in a native language that is simple to understand questions itself to answer. The main advantages of the k-Means clustering algorithm also are the main disadvantages:

- Euclidean distance is applied as the metric, dispersion, and scattering measure also.
- The number k of clusters is the input value; thus, incorrect choice of k can be the reason for non-positive results and needs some testing.
- Convergence can cause many contradictions in calculations.

The results of experiment №1 and experiment №2 are published in papers [82, 83].

3.3 Experiment №3

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.

3.3.1 Dataset

As a result and a future work of experiment №2, we have collected data from respondents of survey 2. The dataset contains 500 rows, per each row 10 sentences about respondents, their MBTI types which are our labels (independent variables) and Instagram accounts. To collect more data using Instagram API we started to scrap Instagram posts of each participant of our survey 2 and combine them with the open-source data set with labels from the Kaggle website.

As a result, our dataset contains 8675 rows, and each row contains type (one of 16 MBTI types) and posts (last 50 posted texts, separated by three pipe characters "|||"). Table 3.7 represents the structure of dataset 4.

Table 3.7: Dataset 4.

```
type ,posts
INFJ,"'http://www.youtube.com/watch?v=qsXHcwe3krw|||http://41.media.tu
mblr.com/tumblr_lfouy03PMA1qa1rooo1_500.jpg|||enfp and intj moments
https://www.youtube.com/watch?v=iz7lE1g4XM4  sportscenter, not top ten
plays  https://www.youtube.com/watch?v=uCdfze1etec  pranks|||What has
been the most life-changing experience in your
life?|||http://www.youtube.com/watch?v=vXZeYwwRDw8
http://www.youtube.com/watch?v=u8ejam5DP3E  On repeat for most of
today.|||May the PerC Experience immerse you.|||The last thing my INFJ
friend posted on his Facebook before committing suicide the next day. Rest in
peace~   http://vimeo.com/22842206|||Hello ENFJ7. Sorry to hear of your
distress. It's only natural for a relationship to not be perfect all the time in
every moment of existence. Try to figure the hard times as times of growth,
as...|||84389  84390  http://wallpaperpassion.com/upload/23700/friendship-
```

3.3.2 Data Preprocessing and Feature extraction

Data Preprocessing and Feature extraction steps:

- Conversion of all letters to lowercase
- Removing of 3 pipe characters '|||' to separate sentences, punctuation, Unicode emojis, URLs, and links
- Tokenization
- Dropping Stop Word
- Lemmatizing - we remove the endings from the words and return the basic form of the word, known as the lemma.
- Count Vectorizerization
- Term Frequency Inverse Document Frequency extraction

To make data more convenient to use, text strings were transformed to vector numbers using Count Vectorizer, then reconstructed to TF-IDF Vectorizer view. As we want to compare several algorithms, we tokenize our data to get a bag of words. The dataset was divided into two sets: one for training that is 80 % and another for testing 20 %.

Count Vectorizer is the tool that transforms given text into vectors based on the frequency of each word occurrence in the entire text, and it is used to extract features from the given text to apply the machine learning models. It is available on sci-kit-learn library in Python [84].

TF-IDF Vectoriser (Term Frequency Inverse Document Frequency) is an algorithm to transform data in a form of text into a representation of numbers (weights of words) that is used to fit the algorithm for future prediction [85].

TF-IDF is calculated as following:

$$w_{i,j} = tf_{i,j} \times idf(w) \tag{3.4}$$

where:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_{k}^{j} n_{i,j}} \tag{3.5}$$

- n is the number of words in a document.

$$idf(w) = log\ (\frac{N}{df_t}) \tag{3.6}$$

- N is the number of documents.
- df_t is the number of documents that contain this word.

After the feature extraction step XGBoost, Naive Bayes, and Recurrent

Neural Network models were implemented one by one on a given preprocessed dataset.

3.3.2 XGBoost

Extreme Gradient Boosted Trees (XGBoost) is a decision tree-based algorithm that uses a gradient enhancement environment and belongs to gradient boosting. This is a universal and flexible tool that is used for most of the problems associated with regression, classification, and ranking. In tasks related to unconstructed data, artificial neural networks created currently surpass all other algorithms or structures [86].

Gradient boosting works in such way:

1. Fits the first model using the original dataset;
2. Fits the second model using the residuals of the first model;
3. Creates the third model using the dot product of the first and the second models.

First of all, XGBoost converts data sets into DMatrix format to make classification, prediction, etc. tasks. The features of XGBoost are:

- regularization that prevents overfitting,
- the algorithm can be run in parallel,
- provision of cross-validation in each iteration,
- handling of missing values automatically,
- use of deeper and optimized trees [87].

For a probability distribution over predicted classes, Softmax function can be used in implementation of XGBoost model and each element after applying it can be in interval (0, 1). Standard Softmax function is defined by formula [88]:

$$\sigma(\vec{z}_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \qquad (3.7)$$

where:

- σ = softmax;
- $(\underline{z})$ = input vector;
- e^{z_i} = standard exponential function for input vector;
- K= number of classes in the multi-class classifier;
- e^{z_j} = standard exponential function for output vector.

3.3.3 Naive Bayes classifier

The Naive Bayes classifier is an algorithm based on the Bayes theorem, which states that functions are independent of each other [89]. The conditional probability of the Naive Bayes theorem is calculated as follows:

$$\mathrm{P(A|B)}= \frac{P(B|A)P(A)}{P(B)} \qquad (3.8)$$

Bayes' theorem helps us find the probability that A happened, given that B happened. A is a hypothesis, and B is proof of this hypothesis. This hypothesis suggests that the features are independent.

3.3.4 Recurrent Neural Networks (RNN)

A Recurrent Neural Network (RNN) is the most common form of the neural network [90]. In neural networks, usually, the input and output data are independent of each other, but when for example you need to predict the next word, you need words that come before that, and, accordingly, you need to remember the previous words. The main and most crucial attribute of RNN is the hidden layer, which is remembered by some sequence information. That is, RNN has a "memory", which is filled with the results obtained when we train the model [90]. For each word, we used an RNN cell - basic LSTMCell, added dropout for inputs and outputs DropoutWrapper, multiple layers to stack cells - MultiRNNCell, and filled initial state with zeros. Since each post has a different length size, we used dynamic RNN. To generalize and regulate inputs, outputs also have been added a dense layer that performs matrix-vector multiplication [91]. To proceed with how the model is learning, we rate with Adam optimizer minimizing mean errors [92].

3.3.5 Results of evaluations of the models

After implementing all the above-mentioned models, we have compared their accuracies using a table for training and test data based on 16 personality types. Table 3.8 shows the highest accuracy of 23.75% and an average of 18.95%, which means the model's poor efficiency.

Table 3.8- Accuracies of algorithms.

Algorithm	Training set, %	Test set, %
Naïve Bayes Classifier	43.91	10.23
XG Boost Classifier	46.62	22.86
RNN	30.0	23.75

That is why we decided to perform classification to each letter of MBTI type; hence, there will be four classifiers: Introvert-Extrovert, Intuition-Sensing, Thinking-Feeling, and Judging-Perceiving. Accuracy sets for training and test set based on four pairs are calculated and now the accuracy is much higher. Classifying four times between two types, we predict each letter and just append them to get MBTI type.

Knowing if a person is an introvert or extrovert is more efficient and easier than the doubtful predicted MBTI type.

As a result of the comparison of the three algorithms, higher accuracy (average 67.4%) is obtained in RNN. The following three tables (Table 3.9, Table 3.10, and Table 3.11) present the accuracy of the train and test for three algorithms, and Figure 3.5 illustrates a diagram of results that were obtained.

Table 3.9 - Accuracy of Naïve Bayes Classifier

	IE,%	NS, %	TF, %	JP, %
train	81.32	70.11	81.24	79.69
test	57.32	55.64	58.47	54.12

Table 3.10- Accuracy of XG Boost Classifier

	IE, %	NS, %	TF, %	JP, %
train	76.48	84.53	73.11	64.12
test	63.17	60.62	78.98	60.09

Table 3.11 - Accuracy of RNN

	IE, %	NS, %	TF, %	JP, %
train	82.01	69.95	81.08	78.40

3.11- table continuation

test	66.69	61.89	79.12	61.9

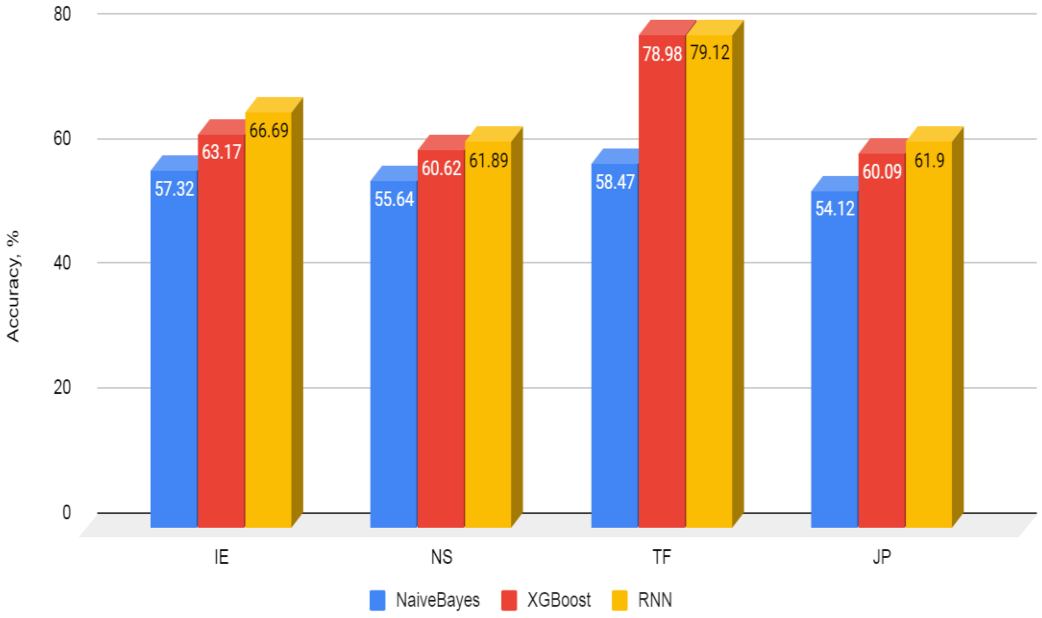


Figure 3.5 - Accuracy of models by traits.

As our data is imbalanced text data, the efficient metric to be considered is the F1 score. Table 3.12 presents the F1 scores of each algorithm by each dichotomy. Metrics such as positive and negative precision and recall were considered to get a maximum clear description of results. The metrics are shown in Table 3.13 and 3.14.

Table 3.12 - F1 scores of models according to dichotomies

dichotomies models	IE,%	NS,%	TF,%	JP,%
Naive Bayes	55	48	59.1	54
XG Boost	23.05	3.37	71.57	74.47
RNN	78	78	78	78

Table 3.13- Precision and Recall of Naive Bayes model

dichotomies metrics	IE,%	NS,%	TF,%	JP,%
positive precision	64	61	63	52
positive recall	67	70	49	58
negative precision	46	37	69	56
negative recall	44	27	62	50

Table 3.14 - Precision and Recall of XG Boost model

dichotomies metrics	IE,%	NS,%	TF,%	JP,%
recall	5	0.5	67	87
precision	57	66.6	69	66

It is noticeable, according to obtained results, that Recurrent Neural Network's performance is much better than other classic algorithms such as Naive Bayes Classifier and XGBoost. During experiment №3 we realized that it is better to classify all traits separately rather than perform classification into 16 classes at once; it leads to more accurate predictions of personality type.

3.4 Experiment № 4

After obtaining the results of experiment №3, we decided to work with image data, too. By using Instagram API, after scrapping Instagram accounts of experiment №2 participants, we gather more than 500 human faces. This data set was taken to experiment №4.

3.4.1 Dataset

To make the data collection process more accurate and faster, only the last 12 posts from each user were scrapped and only photos that are closer to the photo from the passport were chosen. The Instagram images scraper algorithm detects only the person's face and saves only this part as is shown in Figure 3.6. The structure of dataset 5 used in experiment №4 is represented in Table 3.15.

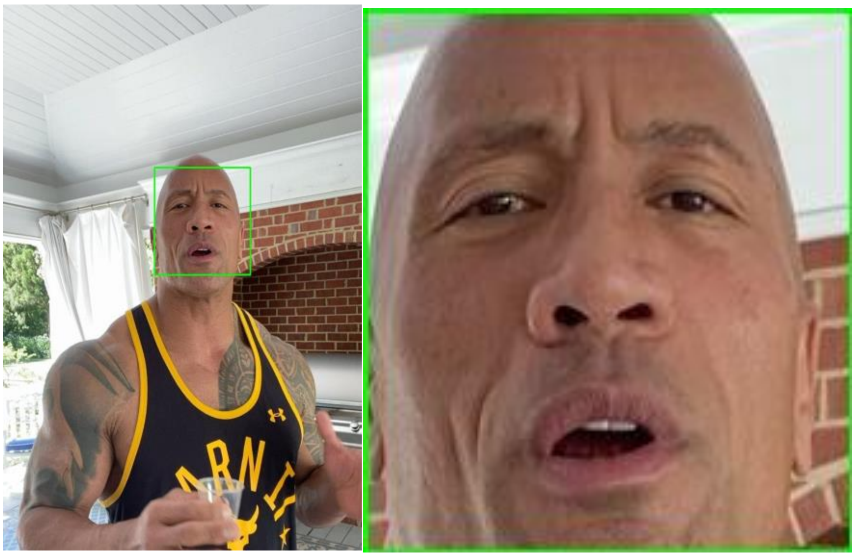


Figure 3.6 -Image preprocessing [93].

Table 3.15 - Dataset 5.

Image	**Type**
cristiano.jpg	INTP
arianagrande.jpg	INFP
therock.jpg	ENFP
selenagomez.jpg	INFJ
kyliejenner.jpg	INFP
…	…
joannagaines.jpg	INFP
iza.jpg	INTP
serenawilliams.jpg	INFJ
ileana_official.jpg	INFP
irfanhakim75.jpg	INTP

Each image of our dataset, part of which is shown in Table 3.15, was converted to Grayscale to avoid complexities; the colored image is in a three-dimensional array whereas gray is one only. To normalize the pixel values of an image, it was divided by 255, since this is the largest value that an array can take, to get a result from 0 to 1. For each pixel, an array of three digits describes the color scheme: RGB (Red, Green, Blue).

3.4.2 Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) is one of the most used neural network models for the task of classifying images [94]. The CNN spends less time and space since it takes not every pixel individually, but a certain square then rotates the pixel that is most suitable for the criteria. Alternatively, with a fully connected weight network from each pixel, CNN has sufficient weights to look at a tiny piece of the image. The architecture of CNN is shown in Figure 3.7.

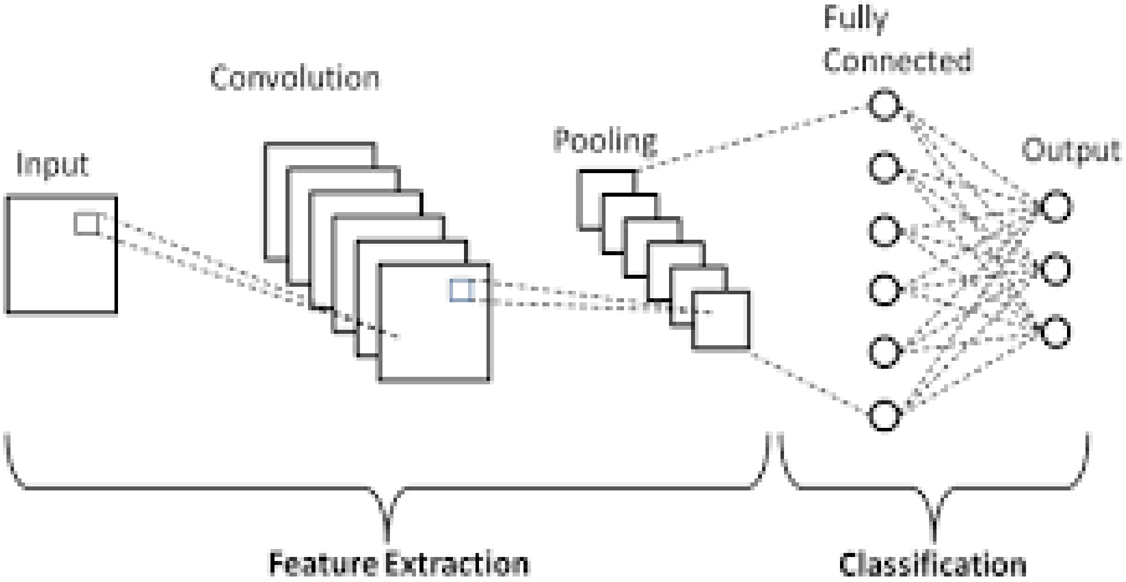


Figure 3.7- Convolutional Neural Network architecture [95].

Since input data is an image, the following properties were defined:

- the shape of Input Layer 64 × 64,
- three convolutional layers with the increasing filter size
- left the same values for kernel size as 5, with 1 stride length,
- padding ‘same’
- Activation function ‘ReLU’, graph of which is represented in Figure 3.8.

ReLU - rectified linear activation function that will output the input directly if it is positive, otherwise, it will output zero, the default activation function for the variety of neural networks because of better performance and simplicity [96]. Despite its name, the function is not linear.

ReLU is calculated by following equation:

$$f(x) = max(0, x) \tag{3.9}$$

where:

- x can be infinite.

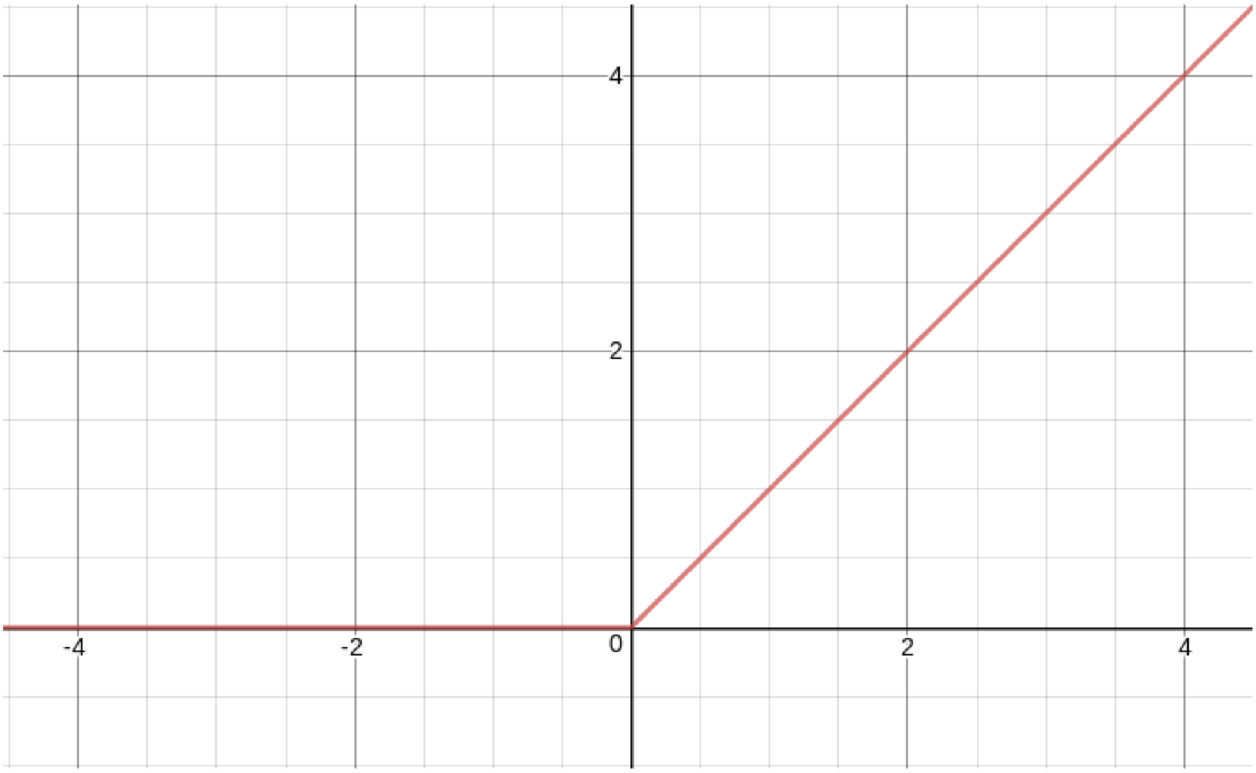


Figure 3.8 - ReLU [43].

A filter is a matrix for determining features from an image, a combination of connections. Filter size was increased because of feature extraction from noisy data. Each time combinations are getting complex, it becomes abstract, and hence we can work with a larger size. Kernel size is the size of these convolutional filters, 5 × 5 square-shaped. The padding determines the size for output volumes based on input. To keep the same output volume size, as the input is 'same padding', it will add zeros around to the input volume. Immediately after the convolutional layer, the MaxPool2D pooling layer was used. In our case, it reduces the spatial dimension of output volume.

After convolutional layers, we use one dropout layer to keep the network generalized and avoid overfitting data (caused by a complex model, performs well on training data but poor for unseen test data). To make classification, we need fully connected layers without any complex structure, just a large piece of ready output data. To convert neurons of the convolutional layer it needs to be flattened, using the dense layer to transfer a three-dimensional array to one-dimensional. Softmax activation function, the graph of which is represented in Figure 3.9, is intended for categorical targets [97] and the formula of which is explained above in Equation 3.7.

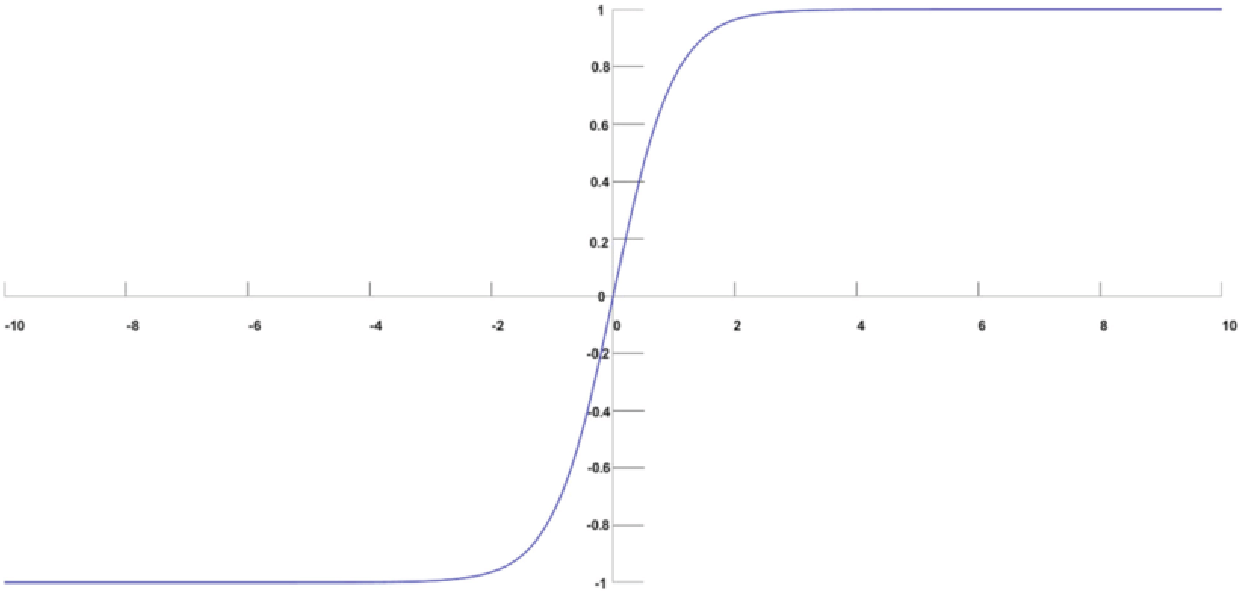


Figure 3.9 - Softmax activation function [43].

The optimizer uses Adam optimizer with a learning rate of 0.001 and determines the learning proceeds. After the model is completed training, the accuracy of 55.07% and the loss value of 2.61 on average were achieved.

Figure 3.10 illustrates custom random images in JupyterLab with the predicted MBTI type and accuracy of them.

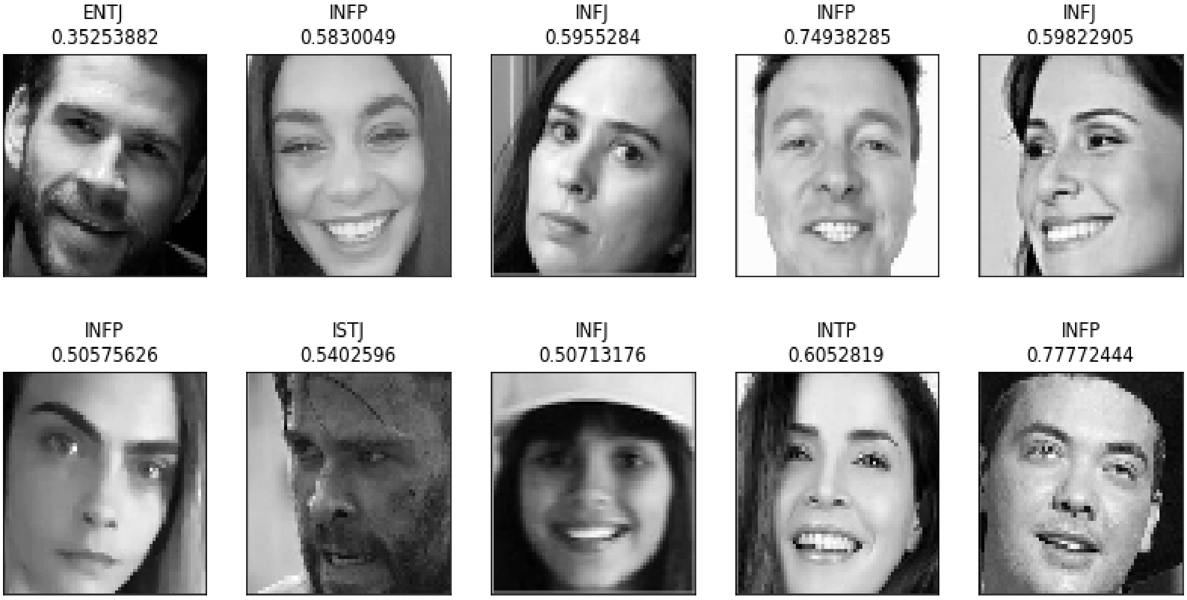


Figure 3.10 - Results of datasets on Jupyter.

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.

3.5 Experiment № 5

After conducting all previous experiments, we decided to analyze the performance of algorithms by running them on Apache Spark running on HDFS (Hadoop Distributed File System). The philosophy of HDFS runs on MapReduce Paradigm and divides work among nodes of the cluster. HDFS has one Master Node (Name Node) for managing the file system and Slave Nodes (Data Nodes) for storing data [98]. Data Nodes have constant communication with Name Node to receive and perform needed tasks from it. Data Nodes also have communication between each other to normally cooperate to perform tasks. The architecture of HDFS is presented in Figure 3.11.

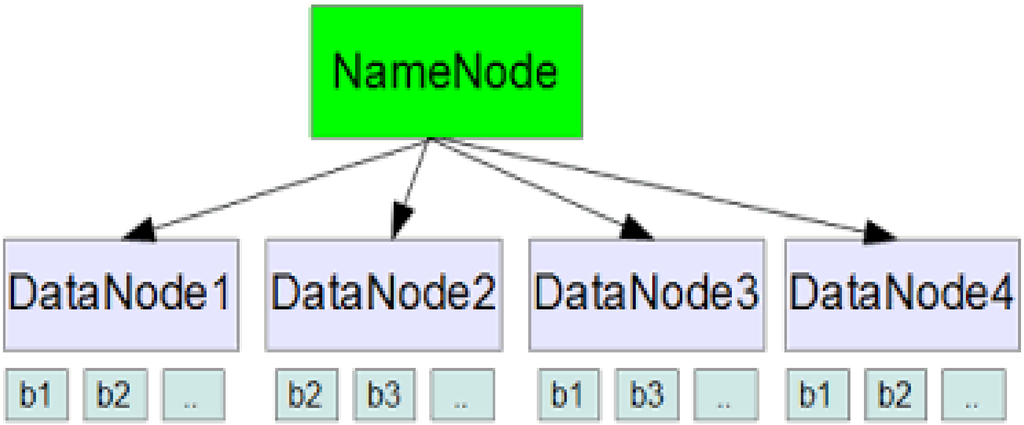


Figure 3.11- Hadoop Distributed File System [98].

The main advantages of HDFS are the low cost, huge storage, fast recovery, portability, and data access. Unlike HDFS, Apache Spark uses Resilient Distributed Datasets (RDD) to store data and RAM to store intermediate data which makes Big Data processing and computation itself faster. Figure 3.12 represents the architecture of Apache Spark.

3.5.1 Apache Spark

Our 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 we rewrote a part of experiment №3 on Pyspark by using the MapReduce paradigm where labels are the keys and posts are the values.

MapReduce is a distributed data processing model proposed by Google for processing large amounts of data on computer clusters. MapReduce is illustrated in Figure 3.13. MapReduce assumes that the data is organized in the form of some records. Data processing in MapReduce takes place in three stages:

1. Stage Map. At this stage, the data is preprocessed using the map function, which is defined by the user. The job of this stage is to preprocess and filter the data. The operation is very similar to the map operation in functional programming languages - a user-defined function is applied to each input record.

The map function is applied to a single input record and produces multiple key-value pairs. It may return only one record, may return nothing, or may return multiple key-value pairs. It is up to the user to decide what will be in the key and the meaning, but the key is a very important thing since data with one key will end up in one instance of the reduced function in the future. All runs of the map function work independently and can work in parallel, including on different machines in the cluster.

2. Stage Shuffle. At this stage, the output of the map function is "parsed into baskets" - each basket corresponds to one output key of the map stage. In the future, these baskets will serve as input to reduce. Shuffle internally represents a parallel sort, so it can also work on different machines in the cluster.

3. Reduce the stage. Each "basket" with values formed at the shuffle stage goes to the input of the reduce function. All runs of the reduce function work independently and can run in parallel, including on different machines in the cluster.

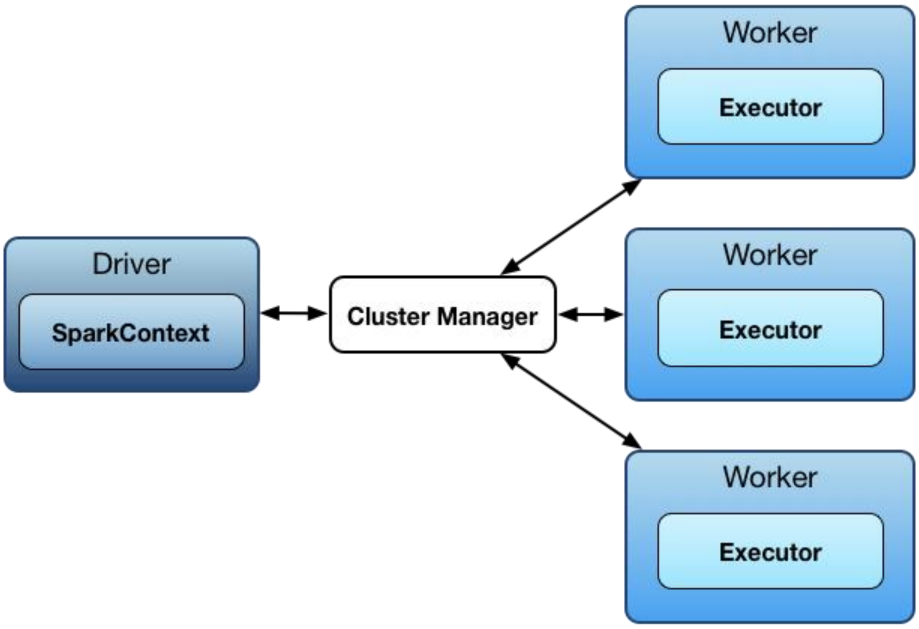


Figure 3.12 - Apache Spark [99].

3.5.2 Results

As the result of experiment №5, the metric F1 score is equal to 0.49. However, one thing which should be noticed is the number of classes was equal to 16. The cluster gives us better performance rather than the performance of the model with 16 classes in experiment №2, but did not affect the speed too much.

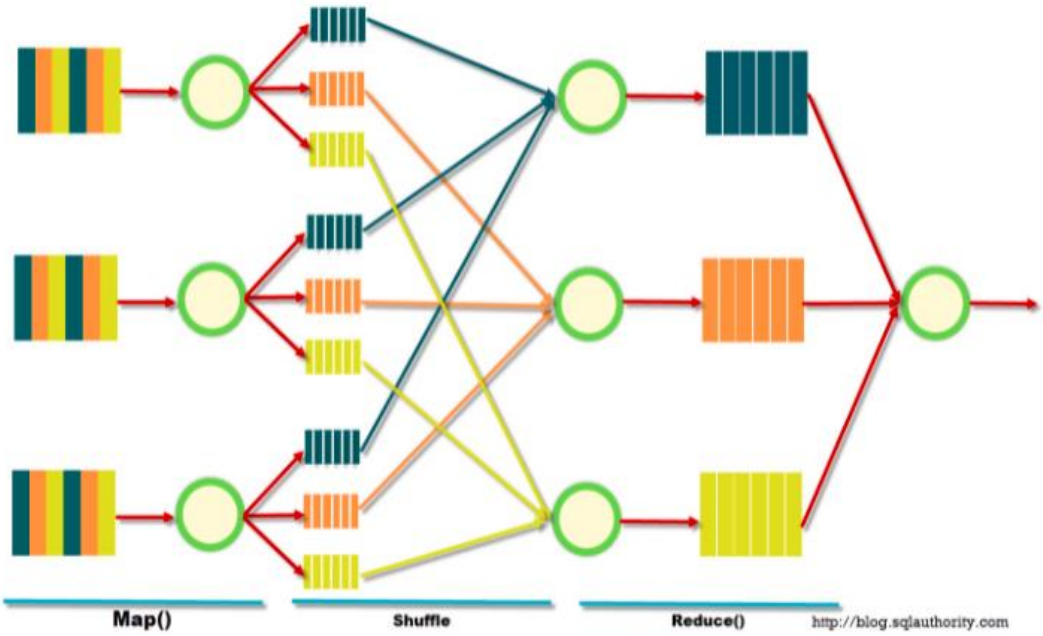


Figure 3.13 - Map-Reduce paradigm [99].

3.6 Experiment № 6

In experiment №6 text written in Kazakh was used. “As a sample we took 450 bachelor’s degree students of the Computer Science Department of Suleyman Demirel University”[83], participants were asked to answer the questions in the automated google form used in survey 2 and provide any text written by them during the last six months. As a text, they may submit their emails, essays, posts.

3.6.1 Dataset

After preprocessing steps the collected dataset consists of 2735 unique rows divided into 2 columns, the first column contains a personality type of a particular participant and the second column contains the participant's text of various sizes. Table 3.16 represents the structure of dataset 6.

Table 3.16: Dataset 6

Type,sentences
ISFP,Қанағаттану
ENFP,"Соңғы кездері ойымды көптенен мәселелер толғандыруда, неден бастарымды да білмеймін. Міне, мектепті де бітірдім, ҰБТ да артта қалды, универге де түстім - бәрі ойдағыдай келе жатқан сияқты негізінде. Бірақ кей кездері әлемде болып жатқан мәселелер қатты уайымдатып жібереді, бәлкім есейгендіктен болар, қарапайым нәрселерді ғана емес, глобальный мәселелерді де ойланады екенсін. Ертеңгі күні не болар екен? Осы сұрақ көкейімді тесіп барады. Қандай қадам жасау керек, қайда бет алып келемін әлі өзіме де түсініксіз. Дәл қазір бар білерім: болашақта мен де еліміздің дамуына үлес қосуға міндеттімін! "…

3.6.2 Data Preprocessing and Feature extraction

To preprocess data and feature extraction fastText library (Word2vec) created by Facebook AI is used. FastText converts a given text into vector representations. To obtain a vector representation of words, the skip-gram and CBOW (Continuous Bag-of-Words) models are applied simultaneously.

Skip-gram is one of the reinforcement learning models that searches for related words for a given word in n-dimensional space. Skip-gram model is presented in figure 3.14.

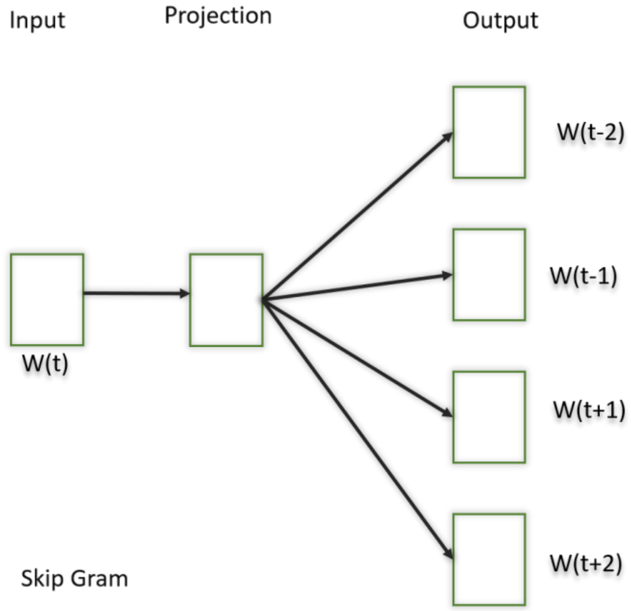


Figure 3.14 - Skip-gram model [100].

CBOW predicts the current word based on its surrounding context, which works in reverse, unlike skip grams. CBOW model is presented in Figure 3.15.

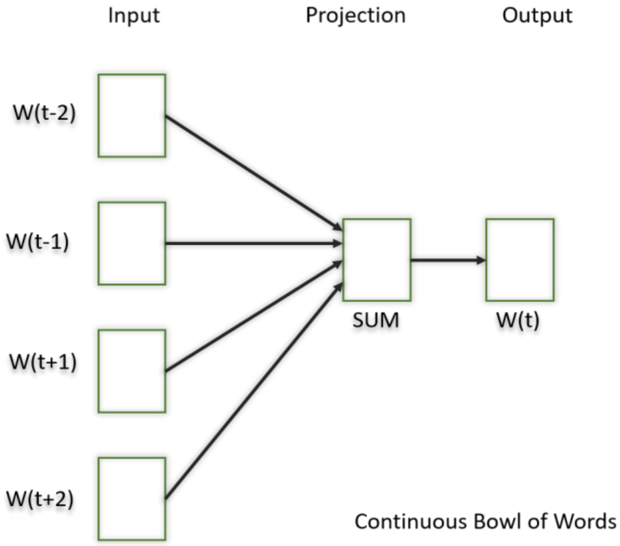


Figure 3.15 - CBOW model [100].

3.6.3 Long Short Term Memory

Long Short Term Memory (LSTM) is a special type of recurrent neural network architecture capable of learning long-term dependencies. The structure of LSTM resembles a chain, which modules contain four layers. LSTM by default has 3 gates: forget gate, input gate, output gate [101].

The first step in LSTM is to determine what information can be thrown out of states. “This decision is made by a sigmoidal layer called the "forget layer". It looks at h_{t-1} and x_t and returns a number between 0 and 1 for each state number in C_{t-1}. 1 means” [101] “keep completely” and 0 means “discard completely”. The structure of forget layer presented in figure 3.16

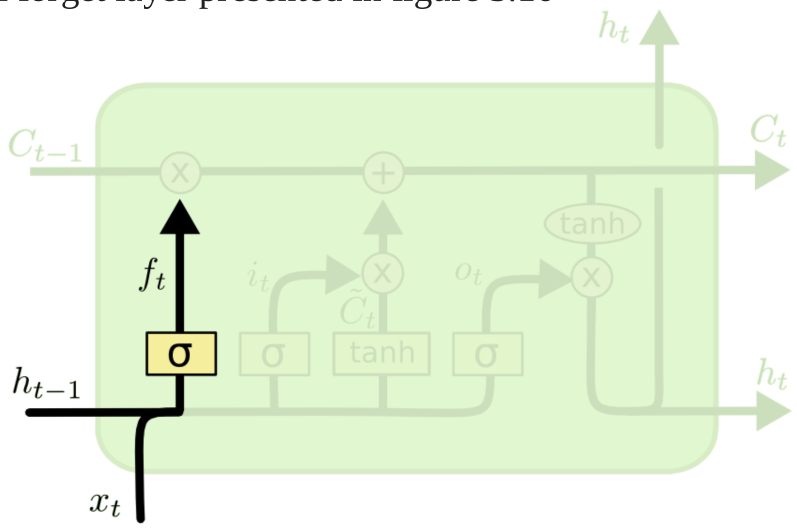


Figure 3.16 - Forget layer [102].

Forget layer is calculated by following equation:

$$f_t = \sigma\left(W_f\left[h_{t-1}\ ,x_t\right]+b_f\right) \tag{3.10}$$

where:

- f_t = *forget gate;*
- σ= *sigmoid function;*
- W_f= *weight of neurons;*
- h_{t-1}= *output vector of previous LSTM unit;*
- x_t = *current vector of LSTM unit;*
- b_f=*biases of representative gates;*
- *t = time step*

The next step is “to decide what new information will be stored in the cell. This stage has two parts. First, a sigmoidal layer called the "input layer" specifies which values should be updated. The *tanh* layer then builds a vector of new candidate values $C`_t$ that can be added to the cell state”[102] . The structure of input layer presented in figure 3.17 and figure 3.18.

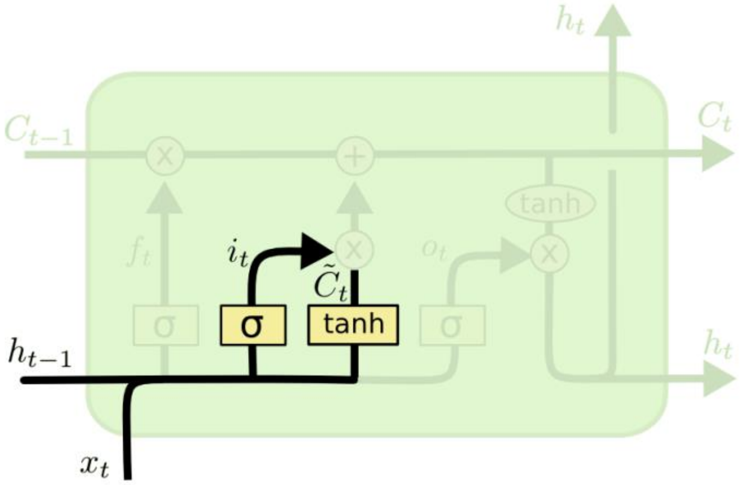


Figure 3.17-Input layer [102].

Input layer is calculated by following equation:

$$i_t = \sigma(W_i[h_{t-1} \ ,x_t]+b_i) \tag{3.11}$$

$$C`_t = tahn(W_C[h_{t-1} \ ,x_t]+b_C) \tag{3.12}$$

where:

- $C`_t = candidate\ for\ cell\ state\ at\ time\ step$
- i_t = *input gate;*
- σ= *sigmoid function;*
- W_i= *weight of neurons;*
- h_{t-1}= *output vector of previous LSTM unit;*
- x_t = *current vector of LSTM unit;*
- b_i=*biases of representative gates;*
- *t = time step*

"To replace the old state of cell C_{t-1} with the new state C_t , we multiply the old state by f_t, forgetting what we chose to forget. Then we add $i_t C`_t$. These are the new candidate values multiplied by t by how much we want to update each of the state values" [102].

The equation for cell state:

$$C_t = f_t \ C_{t-1} \ + \ i_t \ C`_t \tag{3.13}$$

where:

- C_t = cell state;
- $C`_t$ =candidate

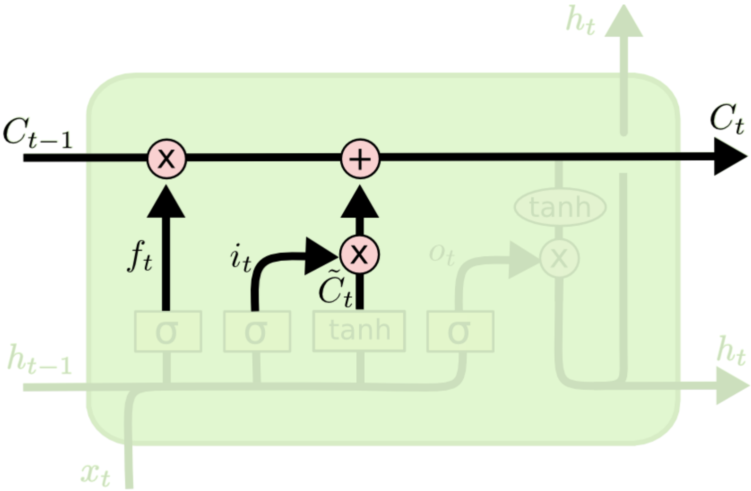


Figure 3.18 - Cell state [102].

The output is based on cell state, with some filters applied on it. “First, a sigmoidal layer that decides what information we will output from the cell state will be applied. The cell state values are then passed through a *tanh* layer to output values in the -1 to 1 range, and are multiplied with the output values of the sigmoidal layer, allowing only the information that is required to be output”[101]. The structure of output layer presented in figure 3.19

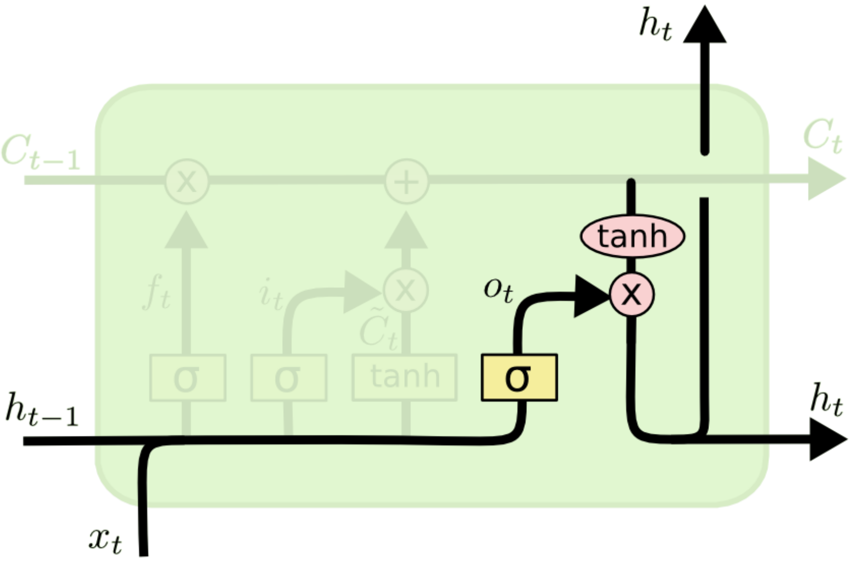


Figure 3.19 - Output layer [102].

Output layer is calculated by following equation:

$$o_t = \sigma(W_o[h_{t-1,}\ x_t]+b_o) \tag{3.14}$$

$$h_t = o_t tahn(C_t) \tag{3.15}$$

where:

- o_t = *output gate;*
- σ= *sigmoid function;*
- W_o= *weight of neurons;*
- h_{t-1}= *output vector of previous LSTM unit;*
- x_t = *current vector of LSTM unit;*
- b_o=*biases of representative gates;*
- *t = time step*

Because data size is still not large in experiment №6 the simplest version of LSTM: four LSTM were applied on it.

3.6.4 Results

Table 3.17 shows the highest F1 score 73%, which means the LSTM model's good efficiency. Also, the results of metrics such as Precision, Recall, Accuracy are shown in table 3.17

Table 3.17- F1 score according dichotomies

metrics	IE,%	NS,%	TF,%	JP,%
F1 score	72	50	72	73
Precision	58	66	60	57
Recall	93	41	90	100
Accuracy	59	66	59	58

4 THE ARCHITECTURE OF THE PROPOSED METHOD

This chapter is dedicated to the architecture and design part of the android application. The android application allows the final users to both write a text and upload a photo of them to get predictions according to submitted data. Figure 4.1 illustrates the architecture of an application.

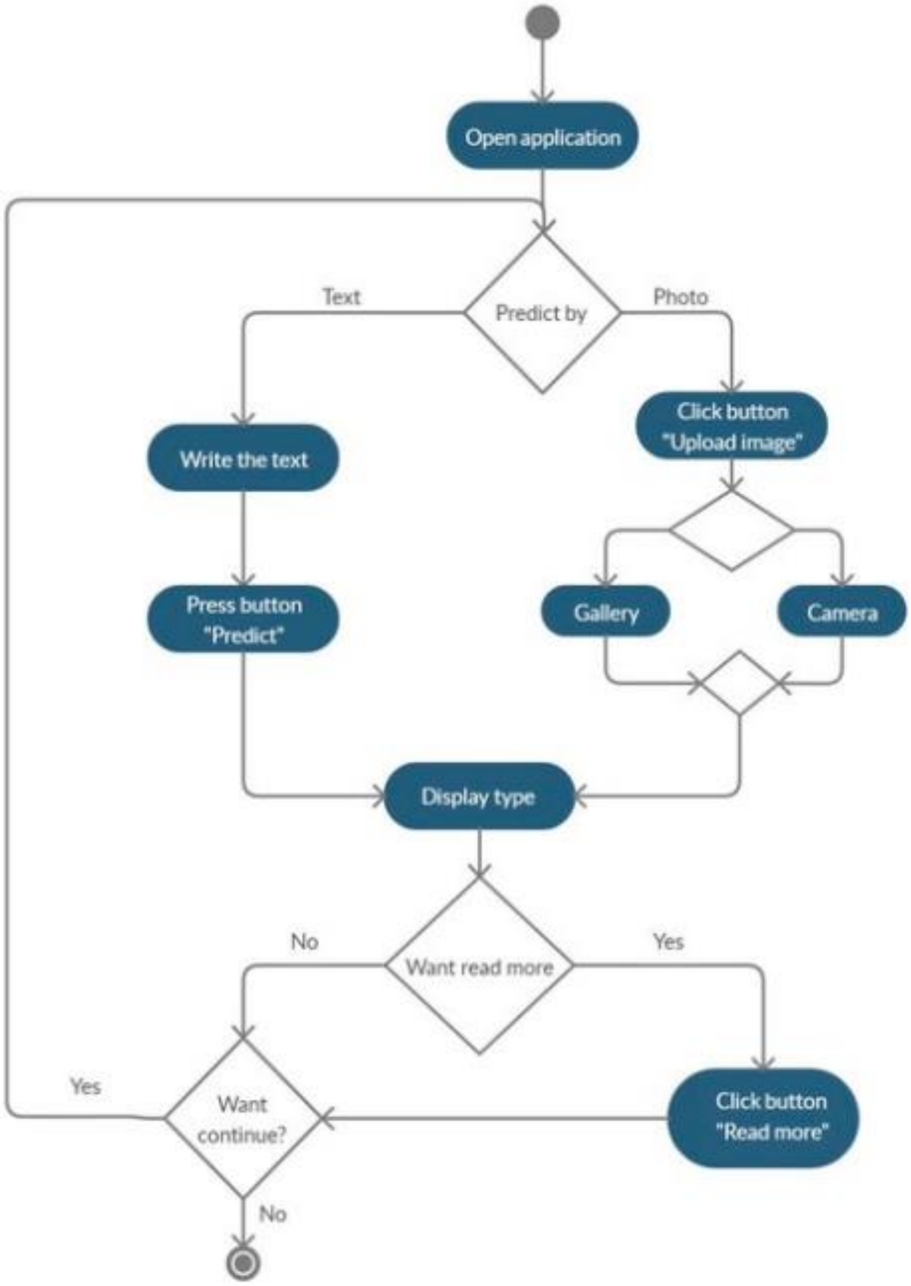


Figure 4.1- UML diagram of an application.

The application consists of two parts as android and machine learning. To connect these machine learning and android parts, Tensorflow and Python Programming Language were used. To implement the model and unfreeze obtained results on Android application, Java Programming Language was used.

TensorFlow is an open-source software library for machine learning, developed by Google to solve the problems of building and training a neural network to automatically find and classify images, achieving the quality of human perception, and for other artificial intelligence algorithms. Tensorflow API is mainly released for Python language, but there are also implementations for C, C++, Java, Swift, and Go languages [103]. In this particular case Tensorflow is used to save and freeze trained models using Python language to apply it on mobile platforms as a file Read from the frozen model and get classification results passing custom test data on Java, Android.

Application has two modes to identify MBTI type of person. The first mode is from text data and the second one is prediction by uploading image data, shown in Figure 4.1. The screens of the Android application with the step-by-step illustration are shown in Figures 4.2-4.8. Figures 4.2-4.4 illustrate the MBTI type prediction based on text data. Figures 4.5-4.8 illustrate the MBTI type prediction based on image data.

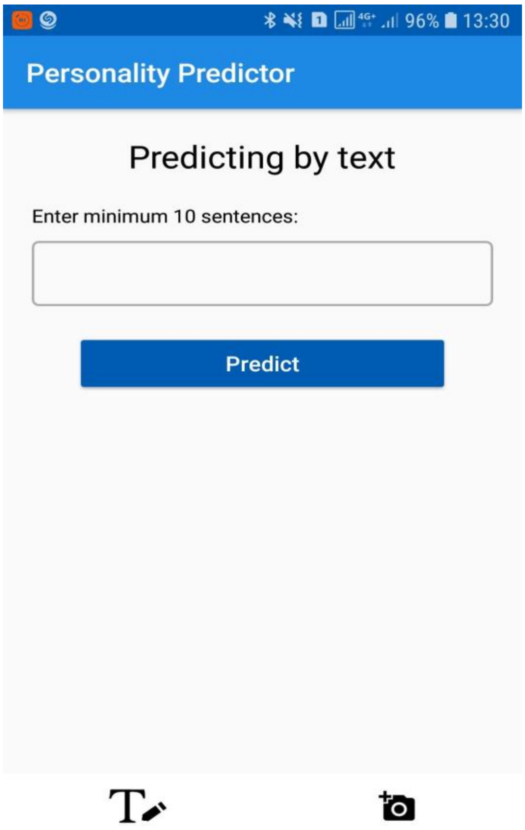
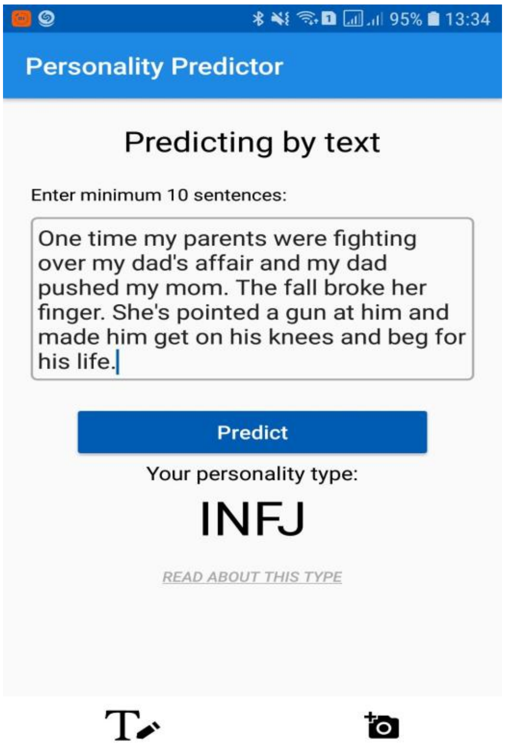


Figure 4.2 - Screen of prediction based on text data. Step 1.



Predicting by text

Enter minimum 10 sentences:

One time my parents were fighting over my dad's affair and my dad pushed my mom. The fall broke her finger. She's pointed a gun at him and made him get on his knees and beg for his life.

Your personality type:

INFJ

Figure 4.3 - Screen of prediction based on text data. Step 2.

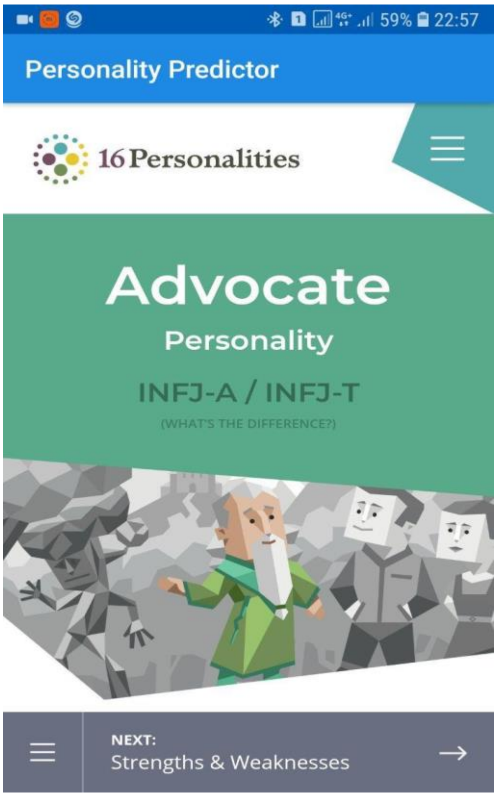


Figure 4.4 - Screen of prediction based on text data. Step 3.

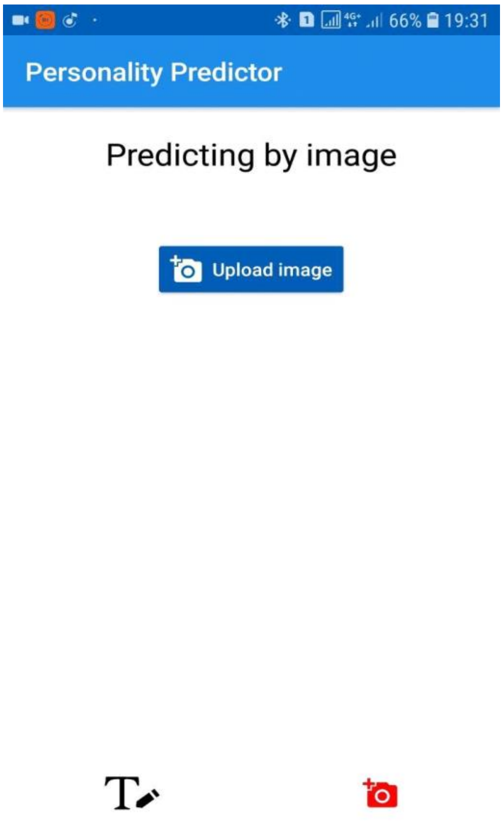


Figure 4.5 - Screen of prediction based on the image data. Step 1.

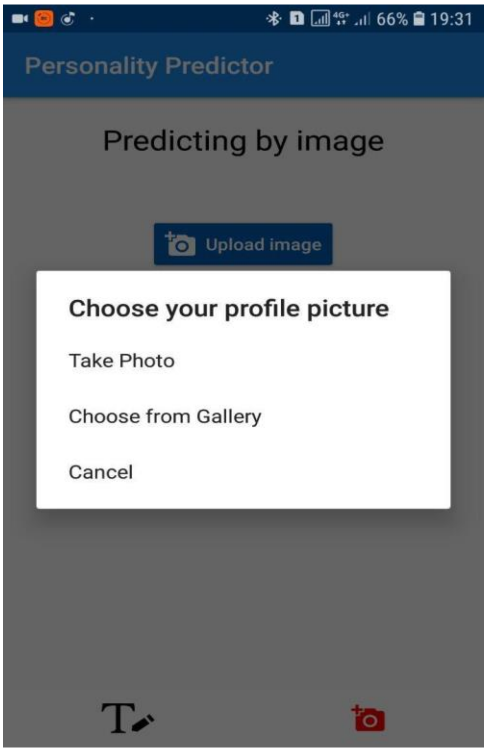


Figure 4.6 -Screen of prediction based on image data. Step 2.

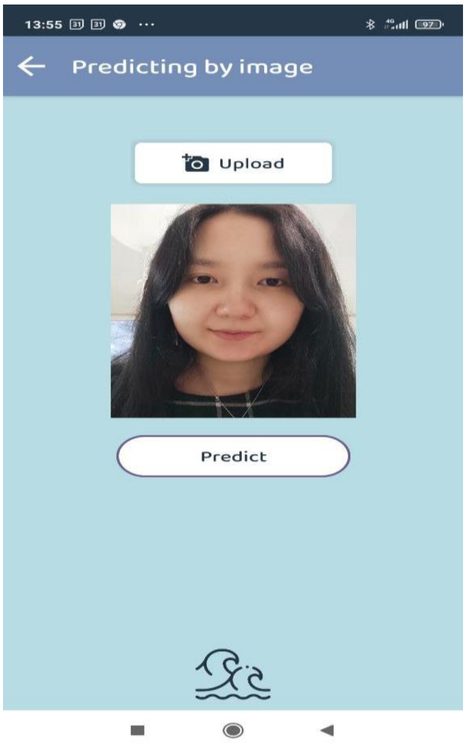
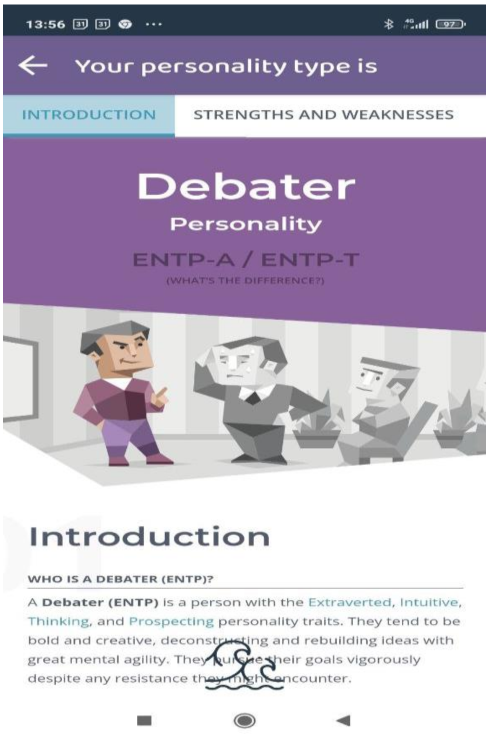


Figure 4.7 - Screen of prediction based on image data. Step 3.



Introduction

WHO IS A DEBATER (ENTP)?

A **Debater (ENTP)** is a person with the Extraverted, Intuitive, Thinking, and Prospecting personality traits. They tend to be bold and creative, deconstructing and rebuilding ideas with great mental agility. They pursue their goals vigorously despite any resistance they might encounter.

Figure 4.8 - Screen of prediction based on image data. Step 4.

5 RESULTS

After the development of an Android application, the last step was system testing. Table 5.1 shows the results of applied testing.

Table 5.1 - Application check

#	Description	Steps	Data	Expected results	Actual results	Status
1	Fill the text field	1. Click on a text field 2. write a description of yourself 3. Click "read"	textual. example: "My name is... I like… I prefer ..."	Application takes the input data, processes it, gives output as a type of personality, and suggests profession according to personal descriptions.	As expected	pass
2	Upload image	1. upload an image or choose from a gallery 2. Click "read"	image file			pass

Table 5.2 presents the comparison of existing research works in the field of personality prediction and classification based on MBTI and the results of models implemented in this research work.

Table 5.2 - Comparison of the results

#	Research Work	Dataset	Algorithms	Results				
				Metrics	I/E	S/N	F/T	J/P

5.2-table continuation

1	“Persona Traits Identification based on Myers-Briggs Type Indicator(MBTI) - A Text Classification Approach, Bharadwaj, et al.” (2018) [32]	Kaggle	Naïve Bayes	accuracy	77%	86.2%	77.9%	62.3%
				F1 score				
			SVM	accuracy	84.9%	88.4%	87.0%	78.8%
				F1 score				
			MLP	accuracy	77.0%	86.3%	54.1%	61.8%
				F1 score				
2	“Reddit: A Gold Mine for Personality Prediction, Gjurković” et al. (2018) [33]	MBTI9k	SVM	accuracy				
				F1	79.6%	75.6%	64.8%	72.6%
			LR	accuracy				
				F1	81.6%	77.0%	67.2%	74.8%
			MLP	accuracy				
				F1	82.8%	79.2%	64.8%	72.6%
3	“Personality Classification from Online Text using Machine Learning Approach, Alam Sher et al. Khan”, (2020) [9]	Kaggle	XGBoost	accuracy	99.37%	99.92%	94.55%	95.53%
				recall	97.16%	100	89.96%	92.66%
				precision	100	99.50%	100	100
				F1	98.56%	99.75%	94.72%	96.19%

5.2-table continuation

4	Experiment №1, Experiment №2	Survey 1, survey 2	k-Means clustering	Inertia value	107			
				k	16			
5	Experiment №3	Instagram posts (text data)mixed with Kaggle dataset	Naïve Bayes	accuracy	57.32	55.67	58.47	54.12
				F1	55	48	59	54
				precision	64	61	63	52
				recall	67	70	49	58
			RNN	accuracy	66	61	79	61
				F1	78	78	78	78
			XGBoost	accuracy	63	60	78	60
				F1	23.05	3.37 71.57	71.57	74.47
				precision	57	66.6	69	66
				recall	5	0.5	67	87
6	Experiment №4	Instagram images mixed with Kaggle dataset	CNN	accuracy	55.07			
				loss value	2.61			

5.2-table continuation

7	Experiment №5	Instagram posts (text data) mixed with Kaggle dataset	Apache Spark Multinomial NB	F1	49			
8	Experiment №6	Kazakh text from survey	LSTM	Accuracy	59	66	59	58
				F1 score	72	50	72	73
				Precision	58	66	60	57
				Recall	93	41	90	100

Figure 5.1 and 5.2 are two plots that describe accuracy and loss estimation for training (red lines) and validation data (blue lines) after each epoch. Figures 5.1-5.2 illustrate the overfitting problem; the training loss line continues to decrease while validation loss decreases until 19 epochs and begins increasing again.

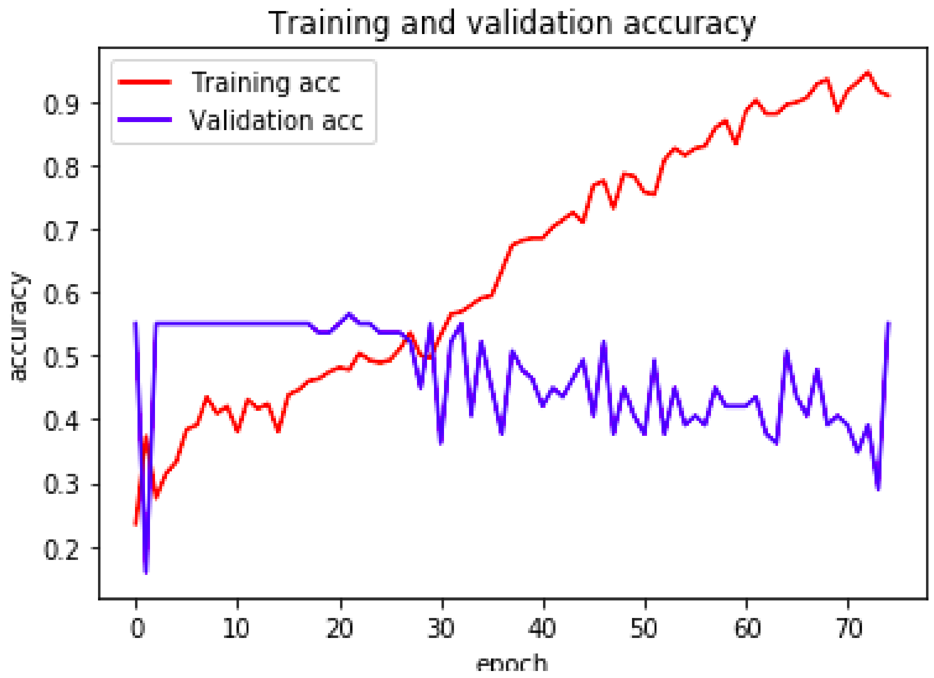


Figure 5.1 - Evaluation of CNN model 1

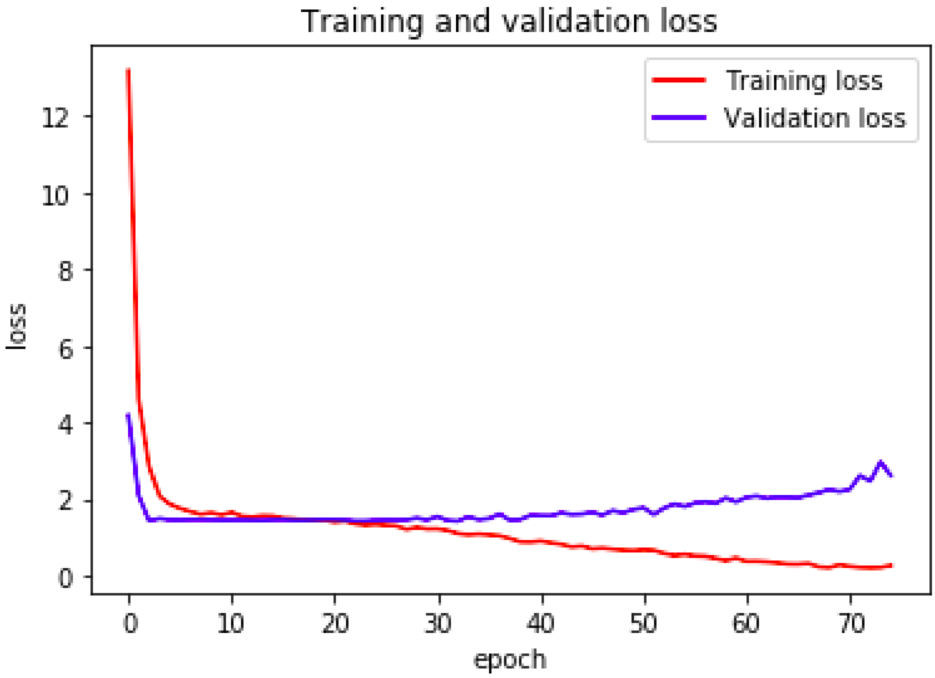


Figure 5.2 - Evaluation of CNN model 2.

CONCLUSION

According to [104] research work, the correct and well-timed professional choice that estimates human desires, inclination, and the psychological type performed at school age improves the labor productivity by 10–15%, decreases staff turnover by 2–2.5 times, and the cost of their training by 1.5–2 times.

This research work considers current issues in profession inclination identification and personality prediction. The study of these spheres allows researchers to apply Computer Science in Education and Psychology spheres.

This thesis proposes and compares supervised, unsupervised, deep learning methods to identify professional inclinations of a person by taking into account personality characteristics and abilities of a person by using different kinds of data such as text, images, and posts that are collected from various sources including Instagram social network.

The results of the research showed that it is better to classify by all character traits separately:

- Introvert-Extrovert,
- Intuition-Sensing,
- Thinking-Feeling,
- Judging-Perceiving;

which allowed us to increase the probability of a more accurate determination of the type of person, than classify by 16 MBTI classes.

The second important point, neural networks are the best algorithms for classifying the personality type in our case, both for text and pictures (CNN, RNN). Convolutional Neural Network gives high performance in images as it has filters and “condition detectors” that work like the human visual system.

Recurrent Neural Network has the so-called "memory", that is, the output is the input for the next layer; it showed such high results when working with Natural Language Processing.

As with all work related to machine learning, we know that an acute problem is a lack of data. Due to the lack of data, as described above, the personality was determined from the text with a 57 percent probability, and then the models to determine from photos were added for training. This was reflected in the inaccuracies in the determination of the image.

The results show that visual and textual data can be used for professional inclination identification based on personality traits and generally work equally well. However, combining them (photos and text) may increase the degree of prediction.

The obtained results have theoretical and practical significance. All of the goals are achieved. Results of all proposed methods are shown and compared between each other and can be useful for further study of issues in the given area.

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.

REFERENCES

1. David C. Funder. Personality. Annual Review of Psychology, 52(1):197–221, 2001
2. Jieun Kim, Ahreum Lee, Hokyoung Ryu. Personality and its effects on learning performance: Design guidelines for an adaptive e-learning system based on user model In 2013 Elsevier. International Journal of Industrial Ergonomics. 2013 p. 1–12. https://doi. org/10.1016/j.ergon.2013.03.001
3. Sawsen Lakhal, Serge Sévigny. Éric Frenette Personality and student performance on evaluation methods used in business administration courses. Educational Assessment, Evaluation, and Accountability. May 2015, Volume 27, Issue 2, pp 171–199 Springer. https://doi.org/10.1007/s11092-014-9200-7
4. Chamorro-Premuzic, T., & Furnham, A. (2003a). Personality traits and academic examination performance. European Journal of Personality, 17(3), 237–250. https://doi. org/10.1002/per.473
5. Oliver P John, Eileen M Donahue, and Robert L Kettle. The big five inventory—versions 4a and 54. University of California, Berkeley, Institute of Personality and Social Research, 1991.
6. Paul T. Costa and Robert R. McCrea. Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI). Psychological Assessment Resources, Odessa, Fla. P.O. Box 998, Odessa 33556, 1992. https://doi.org/10.1037/t03907-000
7. Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. A very brief measure of the big-five personality domains. Journal of Research in Personality, 37(6):504 – 528, 2003. https://doi.org/10.1016/s0092-6566(03)00046-1
8. Charles C. Healy Journal of Career Development In June 2000 Springer, Volume 26, p.295–308.
9. Alam Sher Khan, Hussain Ahmad, Muhammad Zubair Asghar, Furqan Khan Saddozai, Areeba Arif and Hassan Ali Khalid, “Personality Classification from Online Text using Machine Learning Approach” International Journal of Advanced Computer Science and Applications(IJACSA), 11(3), 2020. http://dx.doi.org/10.14569/IJACSA.2020.0110358
10. Статистика. Занятость. Министерство труда. Link: https://www.zakon.kz/4945414-60-vypusknikov-vuzov-v-kazahstane-ne.html Data of access: 31.05.2021
11. Жоғарғы оқу орындары. Link: https://joo.kz/ Date of access: 31.05.2021
12. Путеводитель. Link: https://www.vipusknik.kz/ Date of access: 31.05.2021
13. Гранты. ЕНТ Link: https://univision.kz/ Date of access: 31.05.2021
14. Е.П. Ильин, Дифференциальная психология профессиональной деятельности 2009 link: https://www.litres.ru/evgeniy-

ilin/differencialnaya-psihologiya-professionalnoy-deyatelnosti/ Data of access 20.05.2018

15. Relationships Between MBTI Profiles, Motivation Profiles, and Career Paths, Anna Garden, link: https://www.capt.org/jpt/pdfFiles/Garden_A_Vol_41_3_16.pdf date of access [16.05.2021]
16. Walck, C. L. (1992). Psychological type and manage page 16 Journal of Psychological Type, Vol. 41, 1997 ment research: A review. Journal of Psychological Type, 24, 13-23.
17. Han, K., Moon, K., Lee, J., & Kim, J. (2011). Minnesota Multiphasic Personality Inventory-2 Restructured Form Manual. Seoul, Korea: Maumsarang.
18. Williams, C.L. and Lally, S.J., 2017. MMPI-2, MMPI-2-RF, and MMPI-A administrations (2007–2014): Any evidence of a “new standard?”. *Professional Psychology: Research and Practice*, *48*(4), p.267.
19. Gavrilescu, M. and Vizireanu, N., 2017. Predicting the Sixteen Personality Factors (16PF) of an individual by analyzing facial features. *EURASIP Journal on Image and Video Processing*, *2017*(1), pp.1-19.
20. McCrae, R. R., & Costa, P. T., Jr., (2010). NEO Inventories: Professional manual. Lutz, FL: Psychological Assessment Resources, Inc.
21. Gaughan ET, Miller JD, Lynam DR. Examining the utility of general models of personality in the study of psychopathy: A comparison of the HEXACO-PI-R and NEO PI-R. Journal of personality disorders. 2012 Aug;26(4):513-23.
22. Myers-Briggs Type Indicator (MBTI) Archived 2019-08-26 at the Wayback Machine, *The Myers-Briggs Company*, Sunnyvale, CA, 2019, Retrieved 3 September 2019.
23. Celli, Fabio, and Bruno Lepri. "Is Big Five Better than MBTI? A Personality Computing Challenge Using Twitter Data." In CLiC-it. 2018.
24. R. R. McCrae. Cross-cultural research on the five-factor model of personality. Online Readings in Psychology and Culture, 4(4), 2002. https://doi.org/10.9707/2307-0919.1038
25. V. Rodrı´guez Montequı´n, J. M. Mesa Fernández, J. Villanueva Balsera, A. Garcı´a Nieto. Using MBTI for the success assessment of engineering teams in project-based learning In 2013 Springer. International Journal of Technology and Design Education Vol 23 2013 p.1127–1146. https://doi.org/10.1007/s10798-012-9229-1
26. Charles C. Healy "Interpreting the Myers-Briggs Type Indicator to Help Clients in Understanding Their Strong Interest Inventory" In 2000 Springer Journal of Career Development Vol 26 2000 p 295–308. https://doi.org/10.1177/089484530002600405
27. Tieger, Paul, D. and Barbara Barron-Tieger. "Personality Typing: The First Step to a Satisfying Career." Journal of Career Planning & Employment, Vol. 53, No. 2 (January 1993), pp. 50-56.
28. Pittenger, David J. Measuring the MBTI. . .And Coming Up Short. Journal

of Career Planning and Employment, v54 n1 p48-52 Nov 1993

29. Passmore, J., Holloway, M., & Rawle-Cope, M. (2010). Using MBTI type to explore differences and the implications for practice for therapists and coaches: Are executive coaches really like counsellors? Counselling Psychology Quarterly, 23(1), 1–16. doi:10.1080/09515071003679354
30. Lawrence, G. (1979). Peoples types and tiger stripes: A practical guide to learning styles. Gainesville, FL: Center for Applications of Psychological Type
31. MBTI 16 personality types link: https://www.16personalities.com/infp-personality Date of access [16.05.2021]
32. S. Bharadwaj, S. Sridhar, R. Choudhary, and R. Srinath, "Persona Traits Identification based on Myers-Briggs Type Indicator(MBTI) - A Text Classification Approach," 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, 2018, pp. 1076-1082.
33. M. Gjurković and J. Šnajder, "Reddit: A Gold Mine for Personality Prediction," In Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, pp. 87-97, 2018.
34. B. Plank, and D. Hovy, "Personality traits on Twitter—or—how to get 1,500 personality tests in a week." In Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media Analysis, pp. 92-98, 2015.
35. Bharadwaj, S. Sridhar, R. Choudhary, and R. Srinath, "Persona Traits Identification based on Myers-Briggs Type Indicator(MBTI) - A Text Classification Approach," 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, 2018, pp. 1076-1082]
36. Tausczik, Y.R. and Pennebaker, J.W., 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, *29*(1), pp.24-54.
37. Yavuz, M.C., Analyses of Character Emotions in Dramatic Works by Using EmoLex Unigrams.
38. Support Vector Machine illustration: Link: https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm Date of access 27.05.2021
39. Hastie, Trevor. Tibshirani, Robert. Friedman, Jerome. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, NY, 2009.
40. MLP structure. Link: https://github.com/d-r-e/multilayer-perceptron Date of access 27.05.2021
41. B. Plank, and D. Hovy, "Personality traits on Twitter—or—how to get 1,500 personality tests in a week." In Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media Analysis, pp. 92-98, 2015.

42.Kim A, Song Y, Kim M, Lee K, Cheon JH. Logistic regression model training based on the approximate homomorphic encryption. BMC medical genomics. 2018 Oct;11(4):23-31.

43.Sigmoid function. Logistic Regression in Machine Learning. Link: https://nathanbrixius.files.wordpress.com/2016/06/sigmoid.png Date of access 27.05.2021

44.S. Chaudhary, R. Sing, S. T. Hasan, and I. Kaur, "A Comparative Study of Different Classifiers for Myers-Brigg Personality Prediction Model," IRJET, vol.05, pp.1410-1413, 2018.

45.B., Verhoeven, W. Daelemans and B. Plank, “Twisty: a multilingual twitter stylometry corpus for gender and personality profiling,” In Proceedings of the 10th Annual Conference on Language Resources and Evaluation (LREC 2016)/Calzolari, Nicoletta [edit.]; et al. pp. 1-6, 2016

46.Khan, A., Ahmad, H., Asghar, M.Z., Saddozai, F.K., Arif, A., & Khalid, H.A. (2020). Personality Classification from Online Text using Machine Learning Approach. International Journal of Advanced Computer Science and Applications, 11

47.S. Bharadwaj, S. Sridhar, R. Choudhary and R. Srinath, "Persona Traits Identification based on Myers-Briggs Type Indicator(MBTI) - A Text Classification Approach," 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, 2018, pp. 1076-1082

48.M. Gjurković and J. Šnajder, “Reddit: A Gold Mine for Personality Prediction,” In Proceedings of the Second Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media , pp. 87-97, 2018

49.B. Plank, and D. Hovy, “Personality traits on twitter—or—how to get 1,500 personality tests in a week.” In Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 92-98, 2015.

50.B. Y. Pratama and R. Sarno, "Personality classification based on Twitter text using Naive Bayes, KNN and SVM," 2015 International Conference on Data and Software Engineering (ICoDSE), Yogyakarta, 2015, pp. 170-174.

51.V. Ong et al., "Personality prediction based on Twitter information in Bahasa Indonesia," 2017 Federated Conference on Computer Science and Information Systems (FedCSIS), Prague, 2017, pp. 367-372

52.F. Alam, E. A. Stepanov and G. Riccardi, “Personality traits recognition on social network-facebook,” WCPR (ICWSM-13), Cambridge, MA, USA, 2013.

53.K. Buraya, A. Farseev, A. Filchenkov and T. S. Chua, “Towards User Personality Profiling from Multiple Social Networks,” In AAAI, pp. 4909-4910, 2017.

54.N. R. Ngatirin, Z. Zainol and T. L. C. Yoong, "A comparative study of different classifiers for automatic personality prediction," 2016 6th

IEEE International Conference on Control System, Computing and Engineering (ICCSCE), Batu Ferringhi, 2016, pp. 435-440.

55.S. Chaudhary, R. Sing, S. T. Hasan and I. Kaur, "A comparative Study of Different Classifiers for Myers-Brigg Personality Prediction Model," IRJET, vol.05, pp.1410-1413, 2018.

56. Kaur, P. and Gosain, A., 2018. Comparing the behavior of oversampling and undersampling approach of class imbalance learning by combining class imbalance problem with noise. In *ICT Based Innovations* (pp. 23-30). Springer, Singapore.

57.V. Ong, A. D. Rahmanto, Williem and D. Suhartono, "Exploring Personality Prediction from Text on Social Media: A Literature Review," INTERNETWORKING INDONESIA, vol. 9, no. 1, pp. 65- 70, 2017a.

58.J. Golbeck, C. Robles, M. Edmondson and K. Turner, "Predicting Personality from Twitter," 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, Boston, MA, 2011, pp. 149-156

59.D. Quercia, M. Kosinski, D. Stillwell and J. Crowcroft, "Our Twitter Profiles, Our Selves: Predicting Personality with Twitter," 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, Boston, MA, 2011, pp. 180-185.

60.B., Verhoeven, W. Daelemans and B. Plank, "Twisty: a multilingual twitter stylometry corpus for gender and personality profiling," In Proceedings of the 10th Annual Conference on Language Resources and Evaluation (LREC 2016)/Calzolari, Nicoletta [edit.]; et al. pp. 1-6, 2016

61.F. Celli, "Mining user personality in twitter, " Language, Interaction and Computation CLIC, 2011.

62.X. Sun, B. Liu, Q. Meng, J. Cao, J. Luo, and H. Yin, "Group-level personality detection based on the text generated networks," World Wide Web, pp. 1-20, 2019.

63.S. Chishti, X. Li, and A. Sarrafzadeh, "Identify Website Personality by Using Unsupervised Learning Based on Quantitative Website Elements, " In International Conference on Neural Information Processing, Springer, Cham. pp. 522-530, 2015.

64.P. H. Arnoux, A. Xu, N. Boyette, J. Mahmud, R. Akkiraju, and V. Sinha, "25 Tweets to Know You: A New Model to Predict Personality with Social Media," 2017, arXiv preprint arXiv:1704.05513

65.F. Celli, "Mining user personality in twitter, " Language, Interaction and Computation CLIC, 2011.

66.X. Sun, B. Liu, Q. Meng, J. Cao, J. Luo and H. Yin, "Group-level personality detection based on text generated networks," World Wide Web, pp. 1-20, 2019

67.F. Celli and L. Rossi, "The role of emotional stability in Twitter conversations," In Proceedings of the workshop on semantic analysis in social media, Association for Computational Linguistics, pp. 10-17, 2012.

68.S. Chishti, X. Li and A. Sarrafzadeh, “Identify Website Personality by Using Unsupervised Learning Based on Quantitative Website Elements, “ In International Conference on Neural Information Processing, Springer, Cham. pp. 522-530, 2015.

69.F. Celli, “Unsupervised personality recognition for social network sites,” In Proc. of Sixth International Conference on Digital Society, 2012.

70.P. H. Arnoux, A. Xu, N. Boyette, J. Mahmud, R. Akkiraju and V. Sinha, “25 Tweets to Know You: A New Model to Predict Personality with Social Media,” 2017, arXiv preprint arXiv:1704.05513.

71.M. Arroju, A. Hassan, and G. Farnadi, “Age, gender and personality recognition using tweets in a multilingual setting,” In 6th Conference and Labs of the Evaluation Forum (CLEF 2015): Experimental IR meets multilinguality, multimodality, and interaction, pp. 23-31, 2015.

72.Demirel KC, Sahin A, Albey E. Ensemble Learning based on Regressor Chains: A Case on Quality Prediction. InDATA 2019 (pp. 267-274).

73.L. C. Lukito, A. Erwin, J. Purnama and W. Danoekoesoemo, "Social media user personality classification using computational linguistic," 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE), Yogyakarta, 2016, pp. 1-6.

74. N. Alsadhan and D. Skillicorn, "Estimating Personality from Social Media Posts," 2017 IEEE International Conference on Data Mining Workshops (ICDMW), New Orleans, LA, 2017, pp. 350-356.

75.R. K. Hernandez and L. Scott, “Predicting Myers-Briggs type indicator with text,” In 31st Conference on Neural Information Processing Systems (NIPS 2017), 2017.

76.D. Xue, L. Wu, Z. Hong, S. Guo, L. Gao et al, “Deep learning-based personality recognition from text posts of online social networks,” Applied Intelligence, vol. 48, no. 11, pp. 4232-4246, 2018.

77.Y. Yan, Y. Liu, M. Shyu, and M. Chen, "Utilizing concept correlations for effective imbalanced data classification," Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014), Redwood City, CA, 2014, pp. 561-568.

78.S. Rezaei and X. Liu, "Deep Learning for Encrypted Traffic Classification: An Overview," in IEEE Communications Magazine, vol. 57, no. 5, pp. 76-81, May 2019.

79. B. Cui and C. Qi, “Survey Analysis of Machine Learning Methods for Natural Language Processing for MBTI Personality Type Prediction”.

80.N. Majumder, S. Poria, A. Gelbukh and E. Cambria, "Deep LearningBased Document Modeling for Personality Detection from Text," in IEEE Intelligent Systems, vol. 32, no. 2, pp. 74-79, Mar.-Apr. 2017

81. BigML https://bigml.com/ date of access [21.11.2018]

82.Talasbek, A., Serek, A., Zhaparov, M., Yoo, S.M., Kim, Y.K. and Jeong, G.H., 2020. Personality Classification Experiment by Applying k-Means Clustering. *International Journal of Emerging Technologies in Learning (iJET), 15*(16), pp.162-177.

83. Talasbek, A., Serek, A., Zhaparov, M., Yoo, S.M., Kim, Y.K. and Jeong, G.H., 2020, February. Personality classification by applying k-means clustering. In *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)* (pp. 421-426). IEEE.
84. Kulkarni A, Shivananda A. Converting text to features. In Natural Language Processing Recipes 2019 (pp. 67-96). Apress, Berkeley, CA.
85. Ramos J. Using tf-idf to determine word relevance in document queries. InProceedings of the first instructional conference on machine learning 2003 Dec 3 (Vol. 242, No. 1, pp. 29-48).
86. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y. and Cho, H., 2015. xgboost: extreme gradient boosting. *R package version 0.4-2*, *1*(4).
87. Mitchell, R. and Frank, E., 2017. Accelerating the XGBoost algorithm using GPU computing. *PeerJ Computer Science*, *3*, p.e127.
88. Bouchard, G., 2007. Efficient bounds for the softmax function, applications to inference in hybrid models. In *Presentation at the Workshop for Approximate Bayesian Inference in Continuous/Hybrid Systems at NIPS-07*.
89. McCallum, Andrew. "Graphical Models, Lecture2: Bayesian Network Representation" (PDF). Retrieved 22 October 2019.
90. Mao J, Xu W, Yang Y, Wang J, Huang Z, Yuille A. Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv preprint arXiv:1412.6632. 2014 Dec 20.
91. Kong, C., Kanezashi, M., Yamomoto, T., Shintani, T. and Tsuru, T., 2010. Controlled synthesis of high-performance polyamide membrane with a thin dense layer for water desalination. *Journal of Membrane Science*, *362*(1-2), pp.76-80.
92. Zhang Z. Improved adam optimizer for deep neural networks. In2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS) 2018 Jun 4 (pp. 1-2). IEEE.
93. Duein Jonson, Link: https://www.instagram.com/therock/?hl=en Date of access [20.05.2019]
94. Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences. arXiv preprint arXiv:1404.2188. 2014 Apr 8.
95. Phung, & Rhee, (2019). A High-Accuracy Model Average Ensemble of Convolutional Neural Networks for Classification of Cloud Image Patches on Small Datasets. Applied Sciences. 9. 4500. 10.3390/app9214500.
96. Agarap, A.F., 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.
97. Peng, H., Li, J., Song, Y. and Liu, Y., 2017, February. Incrementally learning the hierarchical softmax function for neural language models. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 31, No. 1).
98. Verma A, Mansuri AH, Jain N. Big data management processing with Hadoop MapReduce and Spark technology: A comparison. In2016

Symposium on Colossal Data Analysis and Networking (CDAN) 2016 Mar 18 (pp. 1-4). IEEE.

99. Verma JP, Patel A. Comparison of MapReduce and spark programming frameworks for big data analytics on HDFS. IJCSC. 2016 Mar;7(2):180-4.

100. CBOW and SKIPgram. Link: https://www.geeksforgeeks.org/word-embeddings-in-nl Date of access[10.05.2021]

101. Sundermeyer, M., Schlüter, R. and Ney, H., 2012. LSTM neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*.

102. LSTM. Link https://habr.com/ru/company/wunderfund/blog/331310/ Date of access: 10.05.2021

103. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., and Kudlur, M., 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)* (pp. 265-283).

104. Bomba A., Kunanets N., Nazaruk M., Pasichnyk V., Veretennikova N. (2020) Model of the Data Analysis Process to Determine the Person's Professional Inclinations and Abilities. In: Hu Z., Petoukhov S., Dychka I., He M. (eds) Advances in Computer Science for Engineering and Education II. ICCSEEA 2019. Advances in Intelligent Systems and Computing, vol 938. Springer, Cham. https://doi.org/10.1007/978-3-030-16621-2_45

APPENDIX A: coding

Code A-1 Face detection and scraping

```
def getByUser(account):
    hostUrl = "https://www.instagram.com/"
    try:
        query = hostUrl + account + "?__a=1"
        responseJson = getResponse(query)
        user = responseJson["graphql"]["user"]
        print(user)
        if user["is_private"]:
            return
        else:
            allMedia = user["edge_owner_to_timeline_media"]["edges"]
            for media in allMedia:
                imageDescription = getDescription(media["node"])
                shortUrlToPhotos = media["node"]["display_url"]
                listA = [account, imageDescription, shortUrlToPhotos]
                listList.append(listA)

          def scrap_image(image_path, account):
          image = cv2.imread(image_path) 36
                    gray_image = cv2.cvtColor(image, cv2.COLOR_BGR2GRAY)
                    face_cascade = cv2.CascadeClassifier
                    (cv2.data.haarcascades + ,
                    → "haarcascade_frontalface_default.xml")
                     detected_faces = face_cascade.detectMultiScale
                    ( gray_image,
                    scaleFactor=2,
                    minNeighbors=3,
                    minSize=(30, 30) )
                    if len(detected_faces) == 1:
                            for (x, y, w, h) in detected_faces:
                                         cv2.rectangle(image, (x, y), (x + w, y + h), (0, 255, 0),
                                         2) cv2.imwrite('images/scrapped_images/' + account +
                                         '.jpg')
                               return True
                    else:
                              return False
```

Code A-2: Naive-Bayes model with RDD

```
rdd = data.rdd
rdd = rdd.map(lambda line: parseLine(line))
df = rdd.toDF()
nb = NaiveBayes(smoothing=1.0, modelType="multinomial")
#model = NaiveBayes.fit(train)
pipeline = Pipeline(stages=[countVectors, label_stringIdx, nb])
modelya = pipeline.fit(train)
```