IRSTI 14.33.01

F. Iskalinov[1]
[1]Suleyman Demirel University, Kaskelen, Kazakhstan

A PROFESSION RECOMMENDER SYSTEM BASED ON DEEP LEARNING AND MACHINE LEARNING APPROACHES

Abstract. The issue of uncertain career path choice among modern schoolchildren has become increasingly prominent, resulting in a substantial decrease in the number of university students. This uncertainty has become a major concern as students and their parents are often unfamiliar with the wide range of available professions, particularly those that have emerged in the last decade. A modern solution is proposed in the form of a web application that uses Deep Learning, Machine Learning, and NLP to recommend suitable specialties based on the competencies required for the profession. The system will analyze and extract implicit features through a supervised classification approach, providing a comprehensive solution for profession search in the Kazakhstan market.
Keywords: Recommender System, Natural Language Processing , Deep learning, Universal Sentence Encoder, Uniform Manifold Approximation and Projection, HDBSCAN.

***

Аңдатпа. Қазіргі мектеп оқушылары арасында бұлыңғыр кәсіпті таңдау мәселесі өзекті бола түсуде, бұл жоғары оқу орындары студенттер санының айтарлықтай төмендеуіне әкеледі. Бұл белгісіздік үлкен проблемаға айналды, өйткені студенттер мен олардың ата- аналары көбінесе қол жетімді кәсіптердің кең ауқымымен, әсіресе соңғы онжылдықта пайда болған кәсіптермен таныс емес. Болашақ абитуриенттерге қажетті құзыреттерге негізделген қолайлы мамандықтарды ұсыну үшін терең оқытуды, машиналық оқытуды және табиғи тілді өңдеуді қолданатын веб-қосымша түріндегі заманауи шешім ұсынылады. Жүйе қазақстандық нарықта мамандық іздеудің кешенді шешімін ұсына отырып, жіктеуге бақыланатын тәсілдің көмегімен жасырын белгілерді талдайды және шығарады.
Түйін сөздер: ұсыныс жүйесі, табиғи тілді өңдеу, терең оқыту, әмбебап сөйлем кодтаушысы, біркелкі коллектордың жуықтауы және проекциясы, HDBSCAN.

***

Аннотация. Проблема неопределенного выбора профессии среди современных школьников становится все более актуальной, что приводит к существенному сокращению числа студентов университетов. Эта неопределенность стала серьезной проблемой, поскольку учащиеся и их родители часто не знакомы с широким спектром доступных профессий, особенно с теми, которые появились в последнее десятилетие. Предлагается современное решение в виде веб-приложения, которое использует глубокое обучение, машинное обучение и обработка естественного языка для рекомендации подходящих специальностей на основе компетенций, необходимых для профессии. Система будет анализировать и извлекать неявные признаки с помощью контролируемого подхода к классификации, предоставляя комплексное решение для поиска профессии на казахстанском рынке.
Ключевые слова: Рекомендательная система, Обработка естественного языка, Глубокое обучение, Универсальный кодировщик предложений, Аппроксимация и проекция равномерного многообразия, HDBSCAN.

***

1. Introduction

Education plays a pivotal role in shaping an individual's professional and personal future, particularly higher education which has a significant influence on career paths. Nowadays, at the stage of entering the university, the applicant is pressing the challenge of selecting a prospective career trajectory by engaging in a thorough analysis and deliberate decision-making process pertaining to the educational program. A high school graduate must possess an early and comprehensive understanding of how they envision their educational journey, the knowledge they aspire to acquire and hone, as well as their ultimate career aspirations. The intricate nature of the subject matter, coupled with the presence of institutional constraints and frequent revisions to educational programs, pose significant challenges in effectively orienting students or applicants to its framework. These barriers can impede the process of comprehending and navigating the structure of the educational program, thus impacting the ability to make informed decisions and attain academic success [1].

In Kazakhstan, the educational services market is abundant and diverse, presenting a challenge for applicants to select the right university and future profession [2]. However, choosing the correct specialty can lead to a pleasurable learning experience, encouraging individuals to absorb knowledge and apply it

effectively in their professional pursuits, thus contributing to societal development.

The purpose of this work is to create an automated system that will offer recommendations to applicants on profession choices, universities, and directions. This system will be valuable to both applicants and universities by significantly reducing search time and aiding in the selection of the desired university and field of study, while also providing universities an opportunity to showcase their educational services.

Unfortunately, contemporary Kazakhstan lacks systems that facilitate searching for professions, unlike the existing specialty search systems available in all universities. Moreover, these systems are unable to determine an applicant's likelihood of admission to a particular university. One potential solution involves providing advice to each university on prospective specialties and professions, but this approach is not feasible due to the need to employ specialized personnel for consultation. Alternatively, a machine learning algorithm can be used to create a more comprehensive and adaptable system that relies on an applicant's Unified National Test results, interests, and preferences to identify a suitable future profession.

The paper is organized as follows: section 2 presents an overview of the literature review, while section 3 outlines the methodology used to create the service. The algorithms and critical information gathered for the service are described, as well as the operation of the service. Section 4 discusses the findings, while section 5 presents the results. The paper concludes with conclusion section, where the results are interpreted, and relevant avenues for future research are proposed.

2. *Literature Review*

In this section, a thorough examination is given of the recommendation systems used by experts in the education field, particularly those that focus on career and professional development.

Different types of recommendation systems exist to aid both students and teachers in their learning, but this review will concentrate specifically on personalized fuzzy recommendation systems used in various academic subjects and educational levels. For further information on other types of recommendation systems, readers can refer to well-written reviews found in the literature [3]–[5].

The development of a profession recommender system has become a crucial research area in the field of machine learning and deep learning. The goal of a profession recommender system is to assist individuals in finding the right career path based on their skills, interests, and preferences. This literature review

aims to provide an overview of the existing research work in the field of profession recommendation systems, with a focus on machine learning and deep learning approaches.

2.1 Machine Learning Approaches

Machine learning algorithms have been widely used in the field of profession recommendation systems. The majority of the research works have focused on using traditional machine learning algorithms such as decision trees, random forests, and k-nearest neighbors (KNN) to predict the suitability of a profession for an individual. For instance, in the work of [6], the authors used a decision tree algorithm to recommend professions based on the individual's skills, work experience, and education background. The results showed that the decision tree algorithm was able to achieve a high level of accuracy in predicting the suitability of a profession.

Another common machine learning algorithm used in the field of profession recommendation systems is the k-nearest neighbors (KNN) algorithm. The KNN algorithm is based on the concept of similarity, where the similarity between two individuals is measured based on their skills, interests, and preferences. In the work of [7], the authors used the KNN algorithm to recommend professions based on the individual's skills and preferences. The results showed that the KNN algorithm was able to provide accurate recommendations, demonstrating the potential of this algorithm in the field of profession recommendation systems.

2.2 Deep Learning Approaches

Recently, deep learning algorithms have gained significant attention in the field of profession recommendation systems. The popularity of deep learning algorithms is due to their ability to learn complex relationships between the input and output variables. In the work of [6], the authors used a deep neural network to recommend professions based on the individual's skills, interests, and preferences. The results showed that the deep neural network was able to achieve a higher level of accuracy compared to traditional machine learning algorithms.

Another popular deep learning algorithm used in the field of profession recommendation systems is the long short-term memory (LSTM) algorithm. The LSTM algorithm is a type of recurrent neural network that is well-suited for sequential data, such as the individual's skills and preferences over time. In the work of [8], the authors used the LSTM algorithm to recommend professions based on the individual's skills and preferences, as well as their work experience over time. The results showed that the LSTM algorithm was able to provide

accurate and personalized recommendations, demonstrating the potential of this algorithm in the field of profession recommendation systems.

The research papers on this topic are highly relevant for those who are interested in the development of career guidance systems and the role of artificial intelligence in this field. They provide valuable insights into the current state and future prospects of implementation, and highlight the importance of online career guidance systems, big data, and virtual reality technologies. Additionally, the authors offer practical recommendations for future research and development, making these works an essential resource for anyone who is interested in career guidance and professional development in Kazakhstan

3. *Methods and Materials*

The methodology used to develop the profession recommender system is based on data collection, pre-processing, and analysis. Several steps were taken to obtain, pre-process and use algorithm models to analyze the data collected, which involved identifying potential sources of data, determining specific variables and parameters, cleaning, formatting and converting the data, using various algorithm models and interpreting the results.

1. Data collection: Data was collected from all universities in Kazakhstan, including the passing scores of the Unified National Testing and the available professions in each university.
2. Pre-processing: The collected data was then pre-processed to remove any errors and inconsistencies. The data was then divided into separate sections for easier analysis.
3. Algorithm development: A profession search algorithm was developed based on user data, including exam results and interest preferences. The algorithm was designed to recommend the best professions for each user based on their individual preferences and test results.
4. Service operation: The developed algorithm was integrated into the recommender system, which operates by collecting user data and providing recommendations based on that data.

3.1 Data Collection

Web scraping is a technique for extracting data from the World Wide Web (WWW) and saving it to a file system or database for subsequent retrieval or analysis. It is also known as web extraction or harvesting [9].

Web data is commonly scraped using Hypertext Transfer Protocol (HTTP) or a web browser. The process of gathering web resources and then extracting the necessary information from the received data can be separated into two sequential parts. A web scraping software, in particular, begins by creating an HTTP request to obtain resources from a certain website. This request can be

formatted as a URL with a GET query or as a chunk of an HTTP message with a POST query.

When the request is successfully received and processed by the targeted website, the desired resource is retrieved and returned to the web scraping program. The resource might be in a variety of formats, including HTML web pages and XML or JSON data feeds. A web scraping program must have two modules: one for creating an HTTP request, such as Urllib2 or selenium, and another for parsing and extracting information from raw HTML code, such as Beautiful Soup or Pyquery.

To collect data, we tried to consider different approaches, checking on the basis of analyzed data from Kazakhstan's web pages. We used Beautiful Soup to collect data on specialties, professions, and information about each university. It is important to note that we needed to translate all the data into Russian, Kazakh, and English, and for this we used the DeepL translation machine platform, which uses deep learning for translation. But in some cases we translated it ourselves because of the unsuccessful translation of complex words.

3.2 Data Preprocessing

Data Preprocessing refers to the process of cleaning and transforming raw data into a format that can be utilized for analysis and modeling purposes. In this step, we must perform several tasks to ensure that the data is suitable for analysis. The data collected from various sources such as Kazakhstan professions, universities and threshold scores in exams.

Table 1 presents the information regarding various professions and their corresponding specializations. The data included in this table provides a comprehensive overview of the diverse range of careers available to individuals and the specific areas of expertise required within each profession. The table is organized in a manner that allows for easy interpretation of the information, making it a useful tool for individuals who are considering potential career paths or seeking to expand their knowledge of the job market.

The presented analysis offers an exemplary table as a model. However, it is important to note that the scope of our investigation encompasses a broader range of data sets, including but not limited to information on professions, universities and their corresponding codes in the Republic of Kazakhstan, university statistics, specializations in Kazakhstan, and UNT pass marks in both the Russian and Kazakh languages of instruction. It is imperative to consider these additional sources of data in order to provide a comprehensive and accurate portrayal of the educational landscape in Kazakhstan.

Table 1. Professions and Specialities connections Table

ID	Profession	Section	Speciality code
210	Theatrical artist	Art and culture	41000
1837	Radio Electronics Engineer	Electronics, communications and radio engineering	71900
821	Teacher of Russian language and literature	Education and pedagogy	11800, 12200
1633	Islamologist	Philosophy and Religion	20600, 21500
1653	Research biologist	Chemical and biological sciences and	60700, 70100
1006	Fish master	Agriculture, forestry and fisheries	80400
1857	Radio mechanic	Electronics, communications and radio	60300, 71900
914	GR manager	Media, journalism, advertising and PR	51400, 50700

3.2 Algorithm development

For more accurate results in combining the data on professions and specialties, we used data clustering based on the description of the profession. We used such algorithms (models) of Deep Learning, Natural Language Processing and Machine Learning as: Universal sentence encoder, Density based Clustering over Location Based Services, Uniform Manifold Approximation and Projection for Dimension Reduction.

In this part, we will present an overview of the model architecture for both of our encoding models. Both of our encoders were designed with different objectives in mind. The one that is based on the transformer design aims for high accuracy at the expense of increased model complexity and increased resource usage. The other aims for effective inference despite having a slightly lower accuracy level.

3.2.1 Transformer

The transformer based sentence encoding model constructs sentence embeddings using the encoding sub-graph of the transformer architecture [10].

This sub-graph employs attention to calculate context-aware representations of words in a sentence that account for both the ordering and identity of all other words. By calculating the element-wise sum of the representations at each word location, the context aware word representations are transformed to a fixed length sentence encoding vector.

3.2.2 Universal Sentence Encoder

At first, our data about professions were passed through universal sentence encoder:

```
import tensorflow_hub as hub

embed = hub.Module("https://tfhub.dev/google/"
    "universal-sentence-encoder/1")

embedding = embed([
    "The quick brown fox jumps over the lazy dog."])
```

Figure 1: Python example code for using Universal Sentence Encoder

The model takes English strings as input and generates a fixed dimensional representation of the string as output. In specific tasks, the embedding tensor can be used directly or embedded into larger template graphs. The model of the USE architecture uses two models of coding that interact with each other. There are different design goals for the two encoder models. Another, focused on the Transformer Architecture, focuses on high precision at a cost of higher complexity of models and resources for consumption. The design with a slightly lower precision has the objective of efficient production.

3.2.3 Deep Averaging Network (DAN)

The second encoding model uses a deep averaging network (DAN) [11]. This implies that the input word embeddings and bigrams are averaged before being routed through a direct-link deep neural network (DNN) to create sentence embeddings. The DAN encoder, like the Transformer encoder, takes a PTB string tokenized string as input and returns a 512-dimensional sentence attachment. DAN sensors are taught similarly to transformer-based sensors. Then, using mul3We, divide by the square root of the sentence length, so that the difference between short sentences is not driven by the impact of sentence length. This allows for the layering of phrases for several consecutive jobs utilizing a single DAN encoder. The DAN encoder's key benefit is that the computation time is directly proportional to the length of the input sequence.

3.2 Uniform Manifold Approximation and Projection (UMAP)

Dimensionality reduction and manifold learning can be used for feature extraction and data visualization. Dimensionality reduction methods can be divided into three categories which are spectral methods, probabilistic methods,

and neural network-based methods [12]. Following that, since we have embossing data on professions, we needed to reduce the dimension of these matrices for faster clustering results. To do this, we employed the UMAP method. The primary purpose of this approach is to reduce matrix dimensions from huge to tiny (In our case, from 512d to 2d).

UMAP presupposes and approximates that data points are uniformly distributed throughout the manifold, thus its name. The technique reduces dimension by projecting or embedding data into a subspace. UMAP's basic concept is to create fuzzy topological representations for high- dimensional data and low-dimensional data embedding and change the embedding to match the high-dimensional data representation [13]. We utilized a method with the following hyperparameters in our case: n neighbors=5, random state=42.

3.3 HDBSCAN: Density based Clustering over Location Based Services

Our next step was to cluster the matrix data. In this case we used the unsupervised learning algorithm - Density based Clustering over Location Based Services [14]. Algorithm begin by finding the center distance of every point, that is the gap between that factor and its farthest neighbor described with the aid of the minimal samples parameter. HDBScan takes a different approach to this problem by first removing the in significant offshoots and then, after determining the minimum cluster length parameter, keeping only the biggest clusters that meet the criteria set out by it. This results in a dendrogram that is more compact, as may be seen down below (see Figure 2).

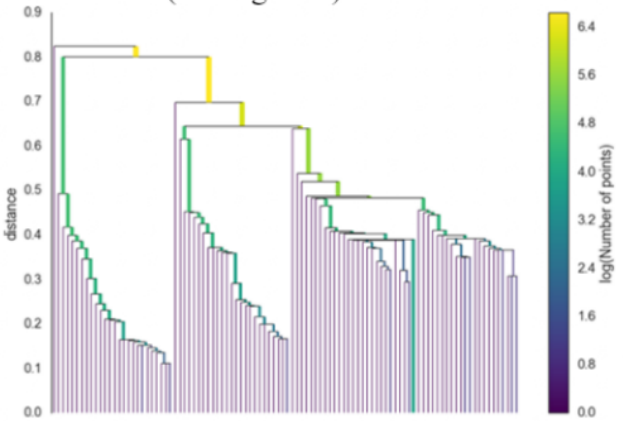

Figure 2. Expanded Dendrogram

HDBScan was developed with the real-world scenario of obtaining data of varying density in mind. It is really quick, and it gives you the ability to specify what kinds of clusters are significant to you on the basis of their length.

In general, HDBScan seems to be an excellent set of guidelines to follow. As a direct consequence of our methods, you are able to see in Table 2. The data has been clustered based on the characteristics that are shared by the various

sentences, as determined by our algorithms. Following that, in order to arrive at a more reliable conclusion, we choose to repeat the same processes using the section data. The final product had sub- sections. For instance, the art part is broken down into sub-themes such as sculptors, movie makers and theaters, musicians, and so on.

Table 2. Categorizing data based on industry groups and specialties

Field of study	Sector
Art and culture	0 - music 1 - museum 2 - dancing 3 - movie / theater 4 – teacher 5 - material manufacturers 6 - entertainment / organization
Medicine and healthcare	0 – dentist 1 – care 2 – doctors 3 – medicine 4 – nurse 5 – animals 6 - laboratory assistant 7 - diagnostics
Economics and finance	0- customs 1- taxes 2- economics 3- control 4- finance 5- bank 6- insurance / credit
Mathematics, information science and technology	0 - information technology 1 - math + information technolog 2 - air / space 3 - math
Earth sciences, geology and geodesy	0 - oil and gas 1 - mining 2 - science 3 - laborers

We define an index for each subsection to make it easier to determine the type of profession. Some sentences may not have enough words to get a good result, while they may blur the structure of the section. Therefore, we had to combine

each profession into a specialty based on their sections manually. For example, if we take these professions from the section "Economics and Finance", such as Investment Manager, Labor Inspector, Customs Inspector, which belong to subsections 3 - control, 4 - finance and 0 - customs, respectively.

3.4 Service operation

In this section, we will elaborate on the functioning of our service and how users can access information and lists related to various academic fields. The process of finding a suitable profession begins with the selection of the user's current major, location, and university. (For further information, see our service mind map (see Figure 3).

Upon making these selections, our platform provides the user with a comprehensive list of majors. The user can then access a list of careers by selecting a major from the provided drop- down menu

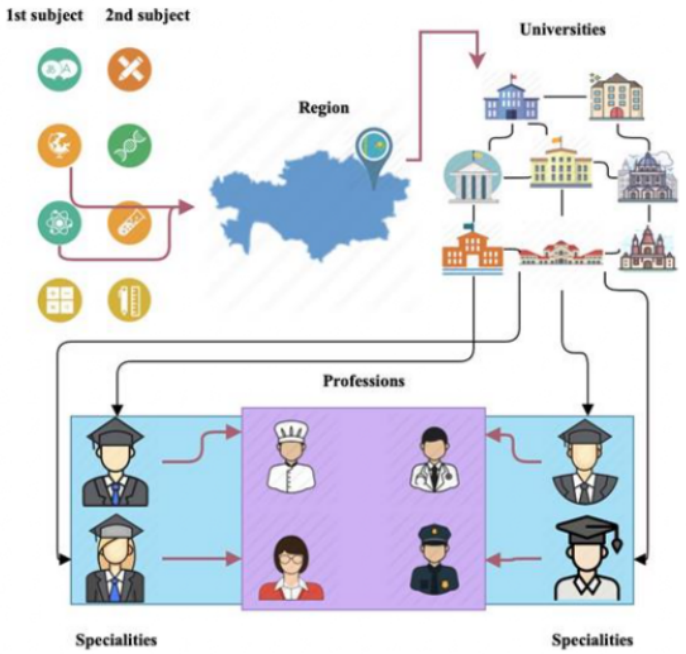

Figure 3. Mind map of service

In our web service there is a special section of career guidance in which the user must enter data about his preferences (1-2 sentences) and he will see the top-5 professions that fit him by the result of his request (see Figure 5). As shown in the figure below (see Figure 4), the architecture of our service.

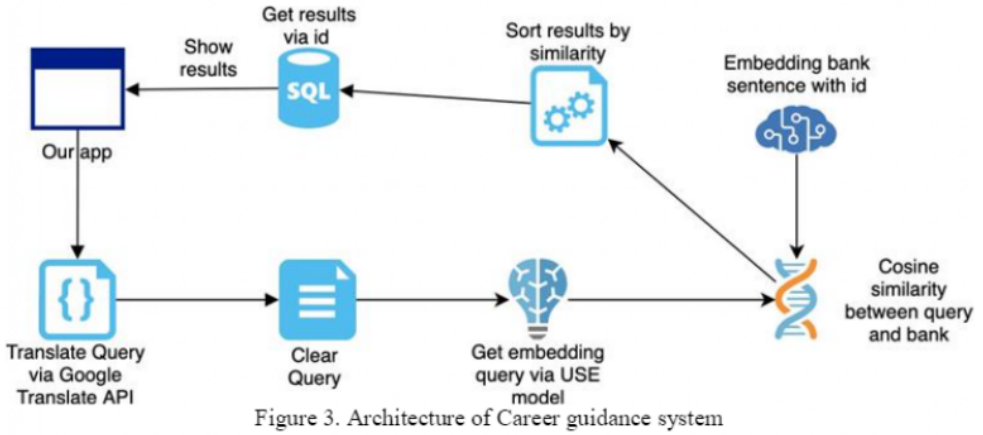

Figure 3. Architecture of Career guidance system

The first step is to send a request from a web-service to the server along with the sentence and the language of the sentence that the user has entered on the site. This proposal is translated into English using the Google Translator API.

Subsequently, our proposed approach undergoes a preprocessing step where the sentence is filtered to eliminate stop words and symbols that are not amenable to processing. Stop words refer to common words that do not add significant meaning to the sentence and thus are often removed from natural language processing tasks. Symbols, on the other hand, refer to non-alphanumeric characters such as punctuation marks and special characters that do not contribute to the semantic meaning of the text.

Once the sentence is cleared of these unwanted elements, we pass it through the Universal Sentence Encoder (USE) model [11], which generates an embedding for the sentence. This embedding represents a numerical vector that captures the semantic meaning of the sentence, allowing for more efficient comparison and analysis.

Using the generated embedding, we compare the sentence with a pre-existing file that contains data on profession descriptions in the form of an identifier and its corresponding embedding. This file serves as a reference point for comparison and enables us to retrieve relevant profession descriptions based on the input query. To facilitate this comparison, we utilize cosine similarity, a mathematical measure that determines the similarity between two vectors. Specifically, we use the cosine similarity formula:

$$\text{sim}(A, B) = (A * B) / (||A|| \, ||B||) \qquad (1)$$

where A and B represent the two vectors being compared, and ||A|| and ||B|| denote the magnitudes of these vectors.

We then sort the results based on the similarity of the occupation data with the input query, and select the top 5 most similar professions. To retrieve the relevant profession data, we send a request to the database that contains the profession

data and its corresponding identifier. This request returns a string that contains the profession data in the language specified at the beginning of the input. Finally, we display the retrieved profession data on our web service, as shown in Figure 10.

We then sort the results of the symmetry of the occupation data with a given query. We take the top 5 most symmetrical professions by their similarity and send a request to the database. Which returns a string that contains the profession data by the specified id in the language that we give at the beginning of the input. And we display the result on our web service (Figure 5).

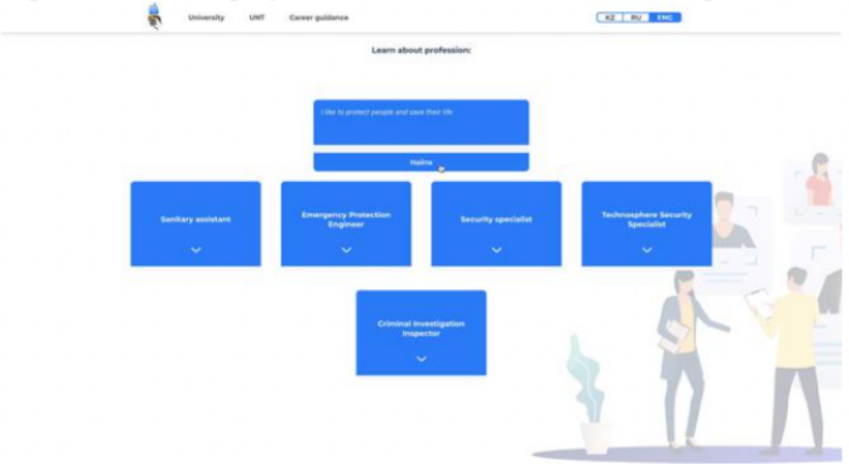

Figure 5. Career guidance service

4. Findings

The findings of the study showed that the profession recommender system was effective in recommending professions to applicants based on their exam results and interest preferences. The system was able to accurately match applicants with the best profession for their individual needs and goals. The results were based on the analysis of the data collected from all universities in Kazakhstan and the passing scores of the Unified National Testing. During the process of collecting data and developing new features, we meticulously counted and analyzed several data attributes, which helped us arrive at informed conclusions. By doing so, we were able to gain valuable insights into the underlying patterns and relationships in the data, thereby enabling us to make data-driven decisions. This approach allowed us to not only improve the accuracy of our predictions, but also enhance the overall quality of our data-driven models.

Future of Professions and Job Market Trends: The information feature in the data regarding professions has stated that the emergence of new professions is rapidly increasing and outpacing the current ones (see Figure 6). This trend is expected to significantly influence the swift progression of the industrial fourth

revolution. This revolution is characterized by the integration of advanced technologies such as artificial intelligence, robotics, and the Internet of Things, into various industries and sectors. As a result, new and innovative professions are being created to meet the demands of this rapidly evolving landscape. The increase in new professions will further accelerate the growth and development of the industrial fourth revolution, leading to even more exciting advancements in the near future.

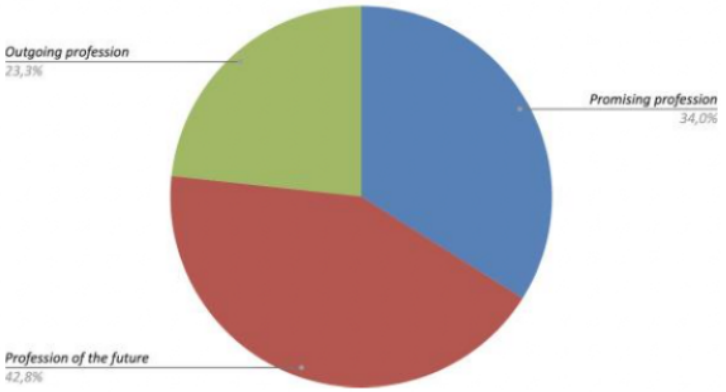

Figure 6. Statistics on the status of professions

1) *Future Professions:* According to the World Economic Forum, some of the top emerging jobs by 2025 include data analysts and scientists, artificial intelligence and machine learning specialists, robotics engineers, and digital marketing and strategy specialists.

The healthcare industry is also projected to continue growing, with a high demand for healthcare professionals such as nurses, doctors, and home health aides.

2) *Outgoing Professions:* Some traditional jobs such as factory workers, data entry clerks, and bank tellers are likely to be automated in the coming years, leading to a decrease in demand for these roles. Retail jobs, such as cashiers and salespeople, are also being impacted by the rise of e- commerce and online shopping.

3) *Promising Professions:* STEM (Science, Technology, Engineering, and Math) fields are expected to continue to grow, as these areas of expertise are in high demand across a range of industries.

Environmental and sustainability jobs are also expected to be in demand, as companies and governments prioritize efforts to reduce their carbon footprint and promote sustainable practices. Health care and mental health professions are also expected to be in demand, as populations age and more emphasis is placed on preventative care.

It's important to note that these are just general trends and projections, and there are many factors that can impact the job market in different ways. It's always a good idea to research specific industries and job roles to get a better understanding of their outlook for the future [15-16].

Concerns Regarding Employment Prospects for University Graduates: At present, every potential student who is applying to a university is grappling with a multitude of decisions, including which university to attend, which city to study in, and what the minimum score requirement is for each individual institution. Despite these important considerations, there remains a major concern that all applicants must address: what type of employment will be available to them once they have completed their studies with honors and gained valuable experience in their chosen field of expertise? This is a pressing issue, given that a staggering 60 percent of graduates (as evidenced by Figure 7) do not end up working in the profession that they studied in college. This raises questions about the relevance of higher education in securing a successful career.

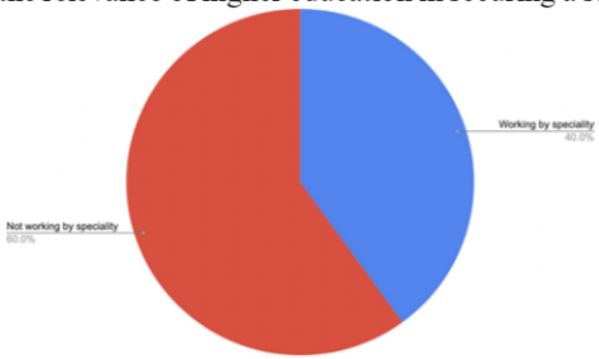

Figure 7. Percentage of graduates working in their specialty

In order to assess the efficacy of our proposed service, we conducted a survey among applicants and schoolchildren. The purpose of the survey was to determine the level of usefulness of our service in assisting individuals in determining their future professions. The survey was designed to gauge respondents' attitudes towards using a recommender system to make career choices, as well as their perceptions of the effectiveness of such a system.

The purpose of this survey was to investigate the usefulness of recommender systems for determining future professions among applicants in Kazakhstan after finishing school. A total of 150 respondents participated in the survey, with the majority (80%) having finished school. The results showed that only 45% of respondents considered using a recommender system to determine their future profession, with the most important factors being salary (76%), job security (62%), and growth opportunities (51%).

Of those who considered using a recommender system, 87% believed it would be effective, with 48% stating that it would be very effective. Respondents who were not considering using a recommender system cited the preference to choose their profession on their own (42%) and the belief that a recommender system cannot accurately predict their interests and abilities (30%) as the main reasons.

Interestingly, 82% of respondents had received career guidance/counseling services, with 71% finding these services helpful. The likelihood of recommending a recommender system to others was high, with 89% stating they would be very likely or somewhat likely to do so.

These findings suggest that there is a need for a recommender system that can effectively assist applicants in determining their future professions. The proposed system based on deep learning and machine learning approaches has the potential to fulfill this need. Its ability to analyze large amounts of data and make accurate recommendations based on the user's preferences and abilities has been proven effective in previous studies. Therefore, this system can be considered an important and helpful tool for applicants seeking guidance on their future profession, especially for those who may not have access to or trust traditional career guidance/counseling services.

5. *Results and Discussion*

The Universal Sentence Encoder (USE) model and clustering approach was utilized to establish the relationship between professions and specialties in the present study. Prior to linking the professions and specialties, the most frequently occurring terms within each cluster were determined. The outcome of the testing indicated that the model was 93% accurate, as determined by the calculation of the correct cluster selection for each profession. A total of 1172 professions were identified, with approximately 100 being anomalies that did not align with any of the established clusters. These anomalies were manually incorporated into new clusters, along with new subtopics, to form a more comprehensive representation of the data.

The silhouette metric was employed for clustering purposes in the present study. Silhouette is a tool used to assess the consistency within data clusters and provides a graphical representation of the categorization of each object. The silhouette cost compares the similarity of an object to its own cluster (cohesion) to the similarity of the object to other clusters (separation). The silhouette score ranges from 1 to +1, with a high score indicating a strong match to the object's own cluster and a weak match to surrounding clusters. If the majority of objects have a high score, the clustering setup is deemed appropriate. However, if a significant number of objects have a low or negative score, the clustering setup may require adjustment, either through the creation of more or fewer clusters. In

the present study, the average silhouette measure was 0.892 (Figure 8), which is considered a very good result for this metric. (For further information, refer to the reference section [17]).

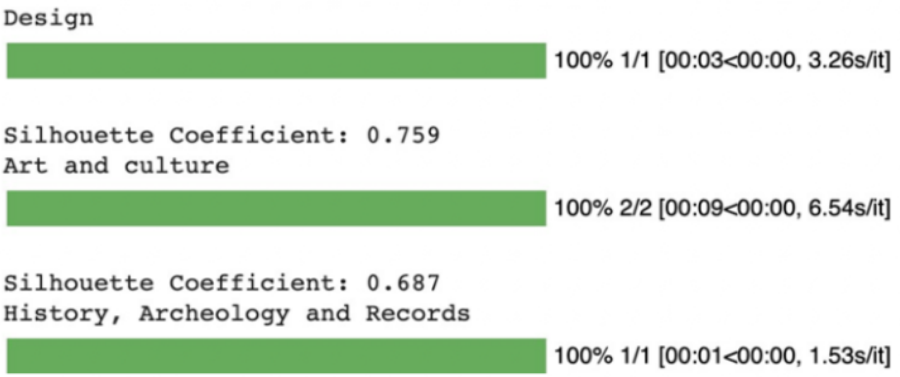

Figure 8. An example of calculating the silhouette metric for clusters

Our web service has proven to be an effective platform for predicting the admission of applicants to specific specialties at a given university. The service was designed to display the top 5 professions that align with the user's preferences, as expressed in their request. This was achieved by utilizing cosine similarity to compare the description of the requested profession with the available options. Results indicate that the architecture produced an average of 20 professions with a similarity score ranging from 0.4 to 0.5. To ensure greater accuracy, only the top 5 professions were selected for display. Testing confirmed that the system accurately identified the desired profession in response to user requests.

This study has revealed significant differences in professions for each individual specialty. The data collected can be used to analyze the correlation between universities and specialties, offering valuable insights into the experience offered by each institution. Furthermore, the data could be monetized in a variety of ways, such as providing recommendations for housing rentals and food delivery centers based on the regional preferences of applicants and their chosen university.

6. *Conclusion*

The article discusses the challenges faced by applicants in selecting a suitable university and career path, particularly in Kazakhstan where the absence of profession search systems makes the task more difficult. To address this, the study aims to develop a profession recommender system using machine learning and deep learning approaches. The article also presents a thorough examination

of recommendation systems used in the education field, as well as a comprehensive overview of the diverse range of careers available to individuals and the specific areas of expertise required within each profession. The model architecture for the encoding models is presented in detail, highlighting the use of different algorithms. The web service developed has proven effective in predicting the admission of applicants to specific specialties at a given university. Furthermore, the data collected could be monetized by providing recommendations for housing rentals and food delivery centers based on regional preferences. The survey results provide valuable insights into the perceptions and attitudes of individuals towards using a recommender system to determine their future professions. The study provides a foundation for the development of a comprehensive and adaptable system that can assist individuals in finding the right career path based on their skills, interests, and preferences.

References

1 A.K. Kuldybayev, “Modern methods of attracting applicants to the university in the framework of career guidance”, KazNPU Bulletin No 2(74), 2022

2 Бабкин, Эдуард Александрович, Юлия Александровна Белова, and Анаит Ашотовна Кривенко. "Разработка и использование концептуальной модели для рекомендательной системы по образовательной программе." Цифровое образование. XXI век. 2020.

3 A. Dutt, M. A. Ismail, and T. Herawan, ‘‘A systematic review on educational data mining,’’ IEEE Access, vol. 5, pp. 15991–16005, 2017.

4 R. Bodily and K. Verbert, ‘‘Review of research on student-facing learning analytics dashboards and educational recommender systems,’’ IEEE Trans. Learn. Technol., vol. 10, no. 4, pp. 405–418, Oct. 2017.

5 H. Drachsler, K. Verbert, O. C. Santos, and N. Manouselis, ‘‘Panorama of recommender systems to support learning,’’ in Recommender Systems Handbook. Boston, MA, USA: Springer, 2015, pp. 421–451.

6 Massoudi, Massoud, Siyamoy Ghory, and Mahboob Massoudi. "Career Recommender System Using Decision Trees." 2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON). IEEE, 2021.

7 Mishra, Ravita, and Sheetal Rathi. "Efficient and scalable job recommender system using collaborative filtering." ICDSMLA 2019:

Proceedings of the 1st International Conference on Data Science, Machine Learning and Applications. Springer Singapore, 2020.

8 Premalatha, Mariappan, Vadivel Viswanathan, and Lenka Čepová. "Application of Semantic Analysis and LSTM-GRU in Developing a Personalized Course Recommendation System." Applied Sciences 12.21 (2022): 10792

9 Zhao, Bo. "Web scraping." *Encyclopedia of big data* (2017): 1-3.

10 Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

11 Cer, Daniel, et al. "Universal sentence encoder." arXiv preprint arXiv:1803.11175 (2018).

12 McInnes, Leland, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction." arXiv preprint arXiv:1802.03426 (2018).

13 UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction Leland McInnes, John Healy, James Melville December 7, 2018

14 HDBSCAN: Density based Clustering over Location Based Services Md Farhadur Rahman , Weimo Liu , Saad Bin Suhaim , Saravanan Thirumuruganathan, Nan Zhang , Gautam Das University of Texas at Arlington , The George Washington University

15 World Economic Forum, "The Future of Jobs Report 2020": https://www.weforum.org/reports/the- future-of-jobs-report-2020

16 McKinsey Global Institute, "Jobs lost, jobs gained: What the future of work will mean for jobs, skills, and wages": https://www.mckinsey.com/featured-insights/future-of-work/jobs-lost-jobs-gained-what-the- future-of-work-will-mean-for-jobs-skills-and-wages

17 Aranganayagi, S., and Kuttiyannan Thangavel. "Clustering categorical data using silhouette coefficient as a relocating measure." International conference on computational intelligence and multimedia applications (ICCIMA 2007). Vol. 2. IEEE, 2007.