DEU FMD 28(83) (2026) 276-283

Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi
Dokuz Eylul University Faculty of Engineering Journal of Science and Engineering
Elektronik/Online ISSN: 2547-958X

RESEARCH ARTICLE / ARAŞTIRMA MAKALESI

Automatic Feature Recommendation System for Job Postings Using Unsupervised Clustering and Association Rule Mining

Denetimsiz Kümeleme ve İlişkisel Kural Analizi Kullanılarak İş İlanları için Otomatik Özellik Öneri Sistemi

Feriştah Dalkılıç [1*], Yunus Doğan [1], Alp Kut [1], Kemal Can Kara [2], Uygar Takazoğlu [2]

[1] Dokuz Eylül Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, İzmir, TÜRKİYE
[2] Kariyer.net Araştırma ve Geliştirme Merkezi, İstanbul, TÜRKİYE

Corresponding Author / Sorumlu Yazar *: feristah.orucu@deu.edu.tr

Abstract

All sectors of business use online employment platforms to publish job postings to find the qualified personnel they need. A job posting text should be neither long enough to make it difficult to read, nor short enough to create a feeling of carelessness, and should contain clear information about the position. However, the preparation of a proper job posting text that complies with the qualifications needed by the job requires expertise and experience. Here, we propose a feature recommendation system to assist recruiters who prepare job posting text. The proposed system mines the relations between the features on a job posting dataset obtained from Kariyer.Net. We used text-mining and natural language processing techniques and employed the DBSCAN clustering algorithm to obtain feature clusters. In addition, we used the Apriori algorithm for finding association rules and frequent item sets to discover relations between feature clusters. Furthermore, we developed a user interface to experiment with feature suggestion results on selected job titles. Our results demonstrate that using such a recommender system makes the job posting preparation process fast and easy and increases the quality of the job postings created.

Keywords: *Artificial Intelligence, Association Rule Mining, Clustering Algorithms, Natural Language Processing, Text Mining*

Öz

Tüm sektörlerdeki işletmeler, ihtiyaç duydukları nitelikli personeli bulmak için iş ilanları yayınlamak üzere çevrimiçi istihdam platformlarını kullanmaktadır. Bir iş ilanı metni, okunmasını zorlaştıracak kadar uzun veya özensizlik hissi uyandıracak kadar kısa olmamalı ve pozisyon hakkında net bilgiler içermelidir. Ancak, işin gerektirdiği niteliklere uygun, doğru bir iş ilanı metni hazırlamak uzmanlık ve deneyim gerektirir. Bu çalışmada, iş ilanı metni hazırlayan insan kaynakları uzmanlarına yardımcı olmak için bir özellik önerme sistemi geliştirilmesi hedeflenmiştir. Geliştirilen sistem ile, Kariyer.Net'ten alınan iş ilanı verisetindeki özellikler kümelenmiş ve aralarındaki ilişkiler incelenmiştir. Özellik kümelerinin oluşturulması için metin madenciliği ve doğal dil işleme tekniklerinden yararlanılmış ve DBSCAN kümeleme algoritması kullanılmıştır. Ayrıca, özellik kümeleri arasındaki ilişkileri keşfetmek, ilişki kurallarını ve birlikte sık kullanılan özellikleri belirlemek amacıyla Apriori algoritması tercih edilmiştir. Belirtilen iş alanına uygun özellik önerilerinin sunulması amacıyla bir kullanıcı arayüzü geliştirilmiştir. Sonuçlarımız, yapay zeka destekli bir öneri sisteminin kullanılmasının, iş ilanı hazırlama sürecini hızlı ve kolay hale getirdiğini ve oluşturulan iş ilanlarının kalitesini artırdığını göstermektedir.

Anahtar Kelimeler: *Yapay Zeka, İlişkisel Kural Analizi; Kümeleme Algoritmaları, Doğal Dil İşleme, Metin Madenciliği*

1. Introduction

Every day, millions of people look for a job, and thousands of businesses look for the qualified personnel they need, applying to employment platforms that broadcast over the Internet. Kariyer.net has been one of the leading employment platforms in Turkey, preferred by recruiters and job seekers since 1999. Their database contains over 25 million resumes and 1.5 million job postings for more than 12,000 different job titles.

Today, many companies of different scales, with or without a human resources department, share the job advertisement texts they have prepared through online employment platforms to reach the candidates they will employ. However, the preparation of an advertisement text that complies with the qualifications needed for the job requires expertise and experience. In some cases, spelling or semantic errors can be made, and expressions with typos can exist. Some of the required features may even be forgotten to be added. Assisting an employer without human resources expertise through a supportive recommendation system in the process of preparing the advertisement text will both facilitate the employer's job and ensure the production of a higher-quality and accurate advertisement text.

Conversion of the unstructured text, which is difficult to understand and interpret by the computer, into a structured format has great importance in computational linguistics. In this way, it will be possible for the computer to interpret the information contained in the advertisement text and use it for different purposes. By using job postings in a structured format, many different studies can be carried out, including filtering the most suitable postings for job seekers, listing the most suitable resumes for job postings, and making various suggestions that will facilitate the work of the users and increase the quality while creating job postings and resumes.

A domain-specific natural language understanding model is necessary for developing an accurate job understanding model

DOI: https://doi.org/10.21205/deufmd.2026288313

Geliş Tarihi / Received: 11.09.2025
Kabul Tarihi / Accepted: 17.10.2025

thanks to job-specific writing styles and entities. In the scope of this study, features mentioned in advertisement texts of seven different job titles have been automatically extracted by applying various natural language processing techniques including text cleaning, punctuation, and number removal, stop word elimination, stemming, etc. Then, the extracted features have been clustered by using the density-based spatial clustering of applications with noise (DBSCAN) algorithm to obtain feature clusters. Each cluster has been represented by a cluster id, feature text, and category id. Categories consist of Graduation, Military Status, Foreign Language, Hard Skills, and Soft Skills. Apriori association rule mining algorithm has been employed to discover the domain-specific associations between feature clusters according to their coexistence in texts. Then discovered associations have been used to suggest features during the preparation of advertisements for the relevant job title.

The main contributions of this study are the presentation of a novel, efficient and automated approach for extraction of domain-specific features from unstructured job posting texts in Turkish and providing support and guidance with the feature recommendation system that will facilitate the recruiters' job, ensuring the production of a higher quality and accurate job posting texts in a shorter time and with less effort, and standardizing job information in postings.

The rest of the paper is organized as follows: Section 2 reviews the relevant literature. Section 3 outlines the methods we used. Section 4 presents the results obtained with the experimental studies. Finally, Section 5 gives the concluding remarks with ideas about future work regarding this study.

2. Related Works

There has been a considerable amount of research on unstructured resumes and job posting texts in recent years. The main task is Information Extraction, which is extracting the desired pieces of information from natural language texts [1-3]. Some of the studies deal with matching applicants with positions by using extracted information from resumes and job postings [4-8]. They first extracted information from resumes and job postings, performed matching and scoring tasks, and finally ranked the results to present to recruiters or job seekers. Another research topic is the online recruitment fraud detection to distinguish fraudulent job advertisements that collect sensitive information (e.g. contact information) of job seekers [9-11]. Also, there exists some other research, including analyzing the job industries to get an insight into the demand in each industry [12] and performing a job skill demand analysis in the recruitment market to measure the popularity of job skills [13, 14].

Several machine learning and deep learning techniques have been applied to resumes and job posting texts in addition to text-mining and natural language processing techniques. In a recent study by LinkedIn researchers, a job understanding model has been developed to convert job postings into standardized job entities (job title, hiring company, skills, qualifications, etc.) by considering user feedback [15]. They used deep transfer learning to adapt open-domain natural language processing (NLP) models and trained professional entity embeddings to achieve better matches between jobs and job seekers. Researchers in [16] conducted an occupational data analysis by using job advertisements from an online recruitment platform in China. The similarity calculation and classification algorithm TextCNN was applied to find the job advertisements related to the target occupation. They constructed a bidirectional long short-term memory conditional random field (BiLSTM-CRF) model based on bidirectional encoder representations from transformers (BERT) pre-training Chinese embedding vectors for the Named Entity Recognition task. In another Chinese language study, the word2vec vector model with a recurrent convolutional neural network-based model was proposed to process the resume texts and BiLSTM was used to extract features within the text [17]. In [18], skills in a job post were extracted by classifying the relevant parts of the description with a multinomial Naïve Bayes model and a named entity recognition (NER) model extracted the required skills from the classified text. RésuMatcher was another study that extracted the qualifications and experience from resumes, and requirements from job postings [19]. It returns relevant jobs for job seekers by using similarity index thanks to the developed finite state transducer (FST) framework for processing regular expression matching.

There are limited number of studies on resumes and job postings in Turkish language. Çelik [20] developed an ontology-driven information extraction system to convert resume content into a structured format. Firstly, resumes were converted into plain text and were partitioned into segments including personal information, education, work experience, personal experiences, etc. Then, proper names/concepts, abbreviations, suffixes, and prefixes of well-known patterns in sentences were parsed, and the abbreviations were converted into standard usage. Classification and clustering module detected the meaningful sentences involved for the segments of the resume, and finally, the Web Ontology Language (OWL) form of an input resume was generated. The authors of [21] aimed to develop a job search engine that retrieves semantically related results. A Word2Vec model was used to represent the documents, and then they were clustered based on their similarity values by integrating into a Lucene-based NoSQL solution. In [22], a job posting recommendation system was developed based on Item-Based Collaborative Filtering.

Existing methods have some shortcomings, such as not handling the feature extraction task in a domain-specific way, working only by adhering to a specific domain, or not using text mining and machine learning techniques. In addition, previous studies are not intended to recommend features for job postings. In this work, we proposed an automatic feature recommendation system that recommends domain-specific features during the preparation of job postings in Turkish. The DBSCAN unsupervised clustering technique and the Apriori association rule mining algorithm were employed for efficient and automatic feature extraction and recommendation operations.

3. Proposed Method

The operation flow of the proposed system is illustrated in Figure 1. Firstly, we performed text cleaning tasks on job posting texts and extracted the features contained in each posting. Some text preprocessing and normalization tasks were applied to the extracted features. Then, for each job title, all extracted features have been clustered with the DBSCAN algorithm. Detected clusters have been assigned to one of the predefined categories. Feature associations have been discovered by using the Apriori association rule mining algorithm. Finally, we suggested features to the user while composing a new job posting by using discovered association. The details of this pipeline will be explained in the following subsections.

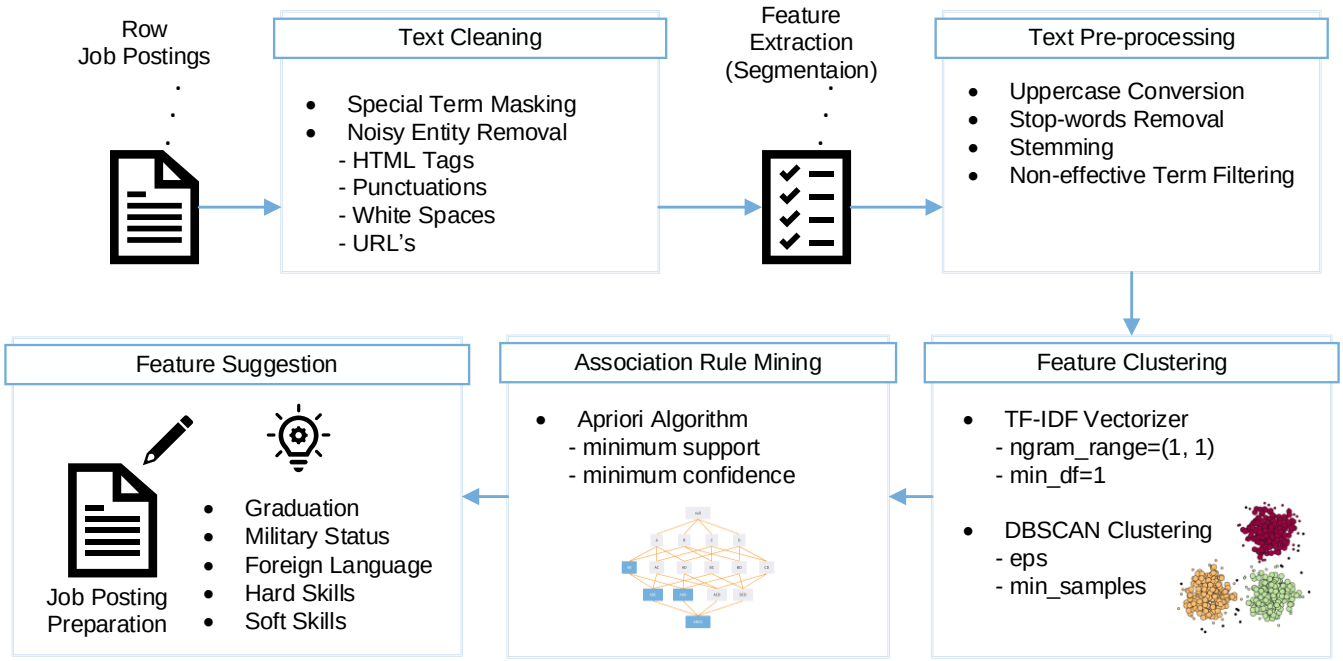


Figure 1. Processing pipeline of the proposed system.

3.1. Text Preprocessing and Feature Extraction

The original job posting texts were dirty and unstructured. At this stage of the study, we cleaned these job posting texts and extracted the features they have. Then, we employed natural language processing techniques to prepare extracted features for clustering. Text preprocessing tasks were conducted in the Visual Studio environment using the C# programming language. Text cleaning operations include removing HTML tags, unnecessary punctuation (except dots and commas), accent marks, and other diacritics from texts and eliminating duplicate white spaces and line feeds. A special term dictionary has been prepared to keep terms containing some special characters, like C++, ASP.NET, etc. These terms are replaced with temporal mask terms.

Feature extraction has been performed by using dots, commas, and line feeds. All letters have been converted to upper case. Then we removed the stop words and predetermined non-effective terms from the extracted features. We stemmed the words except for proper nouns. Zemberek-NLP natural language processing tools for Turkish were used for stemming [23]. Zemberek is officially used as a spell checker in the Turkish version of Open Office and the Turkish national Linux distribution, Pardus. Finally, a non-effective term dictionary has been prepared and employed to filter terms commonly used in many of the features in the dataset, and do not contain information specific to a particular feature. Non-effective terms can be exemplified with the *... konusunda bilgi sahibi (have knowledge of ...)*, and *... becerisi yüksek (good ... skills)* etc.

3.2. Feature Clustering

We developed the feature clustering part of the study by using the Python language and the Google Colab environment. We used pandas, NumPy, and scikit-learn libraries for efficient and rapid development of clustering models. We compared DBSCAN, Ordering Points To Identify the Clustering Structure (OPTIC), and K-means clustering algorithms and decided on the DBSCAN algorithm thanks to its promising clustering results.

Before applying clustering algorithms, we prepared a term frequency-inverse document frequency (TF-IDF) vector for the features by using the TfidfVectorizer package of the scikit-learn library. When building the vocabulary, we used unigrams and bigrams and ignored the terms having document frequency strictly lower than the threshold of 10. Then, we applied the DBSCAN algorithm to obtain feature clusters [24]. DBSCAN is a density-based clustering algorithm that requires eps and min_points as parameters to form clusters. The algorithm classifies the points as core points, density-reachable points, and outliers. A point is labeled as a core point if at least min_points points are within distance eps of it. A point density-reachable from a core point is labeled as a border point and is included in its neighbor core point's cluster. The remaining points not reachable from any core points are outliers.

3.3. Associations Rule Mining on Feature Clusters

After the clustering task has been performed by the DBSCAN algorithm, similar features have been grouped into the same clusters. This part of the study aimed to discover the relations between feature clusters. We employed the Apriori algorithm [25, 26] to find association rules and frequent item sets to discover these relations. Apriori algorithm is frequently used in market basket analysis, which aims at finding regularities in the shopping behavior of customers of supermarkets (or online shops etc.). The discovered associations are used to arrange the products on the shelves of a supermarket or the pages of an online catalog. In addition, these findings facilitate suggesting items to a customer, which may be of interest to him/her. Except for basket market studies, Apriori algorithm has been employed in many fields to find relations between various entities including mining of medical data to identify frequent diseases [27], analysis of the factors influencing highway crash risk [28], detection of the factors affecting non-attendance behaviors of students and determining the factors affecting students' absenteeism behaviors [29], etc.

3.4. Cluster Categories

Clusters were categorized under five different concepts according to the features they represent. These categories were determined as *Graduation, Military Status, Foreign Language, Hard Skills,* and *Soft Skills* by the experts in the field. For example, features like "Mastering C#, Asp.net, and MVC" are in the Hard

Skills category, while features like "Having analytical thinking skills and prone to teamwork" are in the Soft Skills category. The proposed system generates feature suggestions for each of these predefined categories. The explanation of the cluster categories is given in Table 1.

Table 1. Cluster Categories (Labels).

Category	Explanation
Graduation	for clusters that contain features regarding completion of a degree or diploma program
Military Status	for clusters that contain job requirements about mandatory or voluntary military service (completed, deferred/postponed, etc)
Foreign Language	for clusters that indicate language knowledge that is not the native (first) language
Hard Skills	for clusters that contain features indicate a specific, teachable ability that is usually learned through education, training, or practice
Soft Skills	for clusters that contain features indicate a personal or interpersonal ability that helps work well with others, adapt to situations, and succeed in different environments

4. Experimental Studies

We conducted this study on the job posting datasets supplied by Kariyer.net. We selected 7 different job posting datasets belonging to job categories from the most popular ones in demand, including Sales Consultant (SC), Store Manager (SM), Human Resources Specialist (HRS), Secretary (Sec.), Accounting Specialist (AS), Software Specialist (SS), and Industrial Engineering (EE) as a pilot study. The job posting counts for each dataset are given in Table 2.

Table 2. Job posting counts for the datasets.

Dataset	Job Title	Job Posting Count
SC	Sales consultant	3261
SM	Store manager	1758
HRS	Human resources specialist	1594
Sec.	Secretary	1254
AS	Accounting specialist	1037
SS	Software specialist	582
IE	Industrial engineering	257

A total of 9743 job postings were cleaned using natural language processing techniques, and the features they included were extracted. Table 3 exemplifies an unprocessed job posting text from the SS dataset and lists the features extracted after the text preprocessing and feature extraction phases.

Table 3. An example of job posting text before preprocessing and the extracted features from this posting

Unprocessed Job Posting Text	İş Tanımı: - Geliştirmekte olduğumuz oyunun, client/server taraflı iletişimini geliştirmede, daha hızlı ve daha performanslı çalışması konusunda çalışmalar yapmak, - İhtiyaca göre var olan veya yeni C# tabanlı uygulamalar geliştirmek Aranan Nitelikler: - En az 2 yıl C# ile yazılım geliştirme deneyimine sahip, - Nesneye yönelik analiz ve tasarım bilgisine sahip, - TCP/IP, Socket, Buffer, Stream, Network, Thread konularında deneyimli, - Console uygulamalarında tecrübe sahibi, - Byte-Bit, Ram, Read-Write konularında tasarruf yapabilen, - Array/Dizinleme ile arası iyi, - SQL ile uygulamaları iyi harmanlayabilen, - SQL veritabanı optimizasyonu konusunda deneyimli, - Uygulamaların hızlı çalışma ve optimizasyonu konusunda meraklı, - Yaptığı çalışmalarda güvenliğe önem veren, - Web tabanlı sistemler ile entegrasyonlar yapmış (API, Webservis), - Tercihen daha önce oyun projesinde çalışmış, - Tercihen İngilizce bilen ve kaynak araştırması yapabilen, - Planlı ve dökümantasyonlu çalışma yapabilen takım arkadaşları arıyoruz.
Extracted Features (in Turkish)	-Geliştirmekte olduğumuz oyunun client/server taraflı iletişimini geliştirmede daha hızlı ve daha performanslı çalışması konusunda çalışmalar yapmak -İhtiyaca göre var olan veya yeni c# tabanlı uygulamalar geliştirmek -En az 2 yıl c# ile yazılım geliştirme deneyimine sahip -Nesneye yönelik analiz ve tasarım bilgisine sahip -TCP/IP, Socket, Buffer, Stream, Network, Thread konularında deneyimli -Console uygulamalarında tecrübe sahibi -Byte-Bit, Ram, Read-Write konularında tasarruf yapabilen -Array/dizinleme ile arası iyi -SQL ile uygulamaları iyi harmanlayabilen -SQL veritabanı optimizasyonu konusunda deneyimli -Uygulamaların hızlı çalışma ve optimizasyonu konusunda meraklı -Yaptığı çalışmalarda güvenliğe önem veren -Web tabanlı sistemler ile entegrasyonlar yapmış (API, Webservis) -Tercihen daha önce oyun projesinde çalışmış -Tercihen İngilizce bilen ve kaynak araştırması yapabilen -Planlı ve dökümantasyonlu çalışma yapabilen takım arkadaşları arıyoruz
Extracted Features (in English)	-To work on the client/server side communication of the game we are developing to work faster and more efficiently. -Developing existing or new c#-based applications as needed -At least 2 years of software development experience with C# -Has knowledge of object-oriented analysis and design -Experienced in TCP/IP, Socket, Buffer, Stream, Network, Thread -Experience in console applications -Able to save on Byte-Bit, Ram, Read-Write -Good with Array/indexing -Able to blend SQL and applications well -Experienced in SQL database optimization -Curious about fast running and optimization of applications -Attaches importance to safety in his work -Integration with web-based systems (API, Webservice) -Preferably worked on a game project before -Preferably fluent in English and able to do resource research -We are looking for teammates who can do planned and documented work.

Several NLP tasks were applied to the extracted features to achieve text normalization and dimension reduction. Table 4 handles an example extracted feature and shows its changes after each preprocessing task. The final version of the features was used in the clustering process.

A clustering study was applied to whole extracted features of each dataset respectively. The DBSCAN algorithm was preferred for feature clustering thanks to its clustering performance. We conducted a parameter tuning study to optimize the parameters of the DBSCAN algorithm. The Scikit-learn Silhouette score metric, which computes the mean Silhouette coefficient of all samples, has been used to evaluate the quality of the clusters created to understand how well samples are clustered with other samples that are similar to each other. We executed the DBSCAN algorithm for eps values ranging from 0.1 to 1 (in increments of 0.1) and min_points values from 5 to 50 (in increments of 5), resulting in a total of 100 runs for each dataset. The silhouette score ranges from −1 to +1. Values near +1 indicate that a data point is well matched to its assigned cluster and distinctly separated from neighboring clusters, whereas values near −1 suggest that the data point may have been incorrectly assigned to

its cluster. The best parameter combination, offering the highest silhouette score, was determined for each dataset individually, as shown in Table 5, and the final clusters were formed accordingly. The min_samples parameter was set to 5 for all datasets. The Eps parameter was generally set to 0.6. However, it was 0.5 for the SC dataset and 0.7 for the SS dataset.

Table 4. An example feature and NLP tasks applied to it. Suffixes and words affected by the operations are marked in bold.

Feature	Task
NESNEYE YÖNELİK ANALİZ VE TASARIM BİLGİSİNE SAHİP (HAVING OBJECT-ORIENTED ANALYSIS AND DESIGN KNOWLEDGE)	Feature Extraction
NESNEYE YÖNELİK ANALİZ **VE** TASARIM BİLGİSİNE SAHİP	Stop Word Elimination
NESNE**YE** YÖNELİK ANALİZ TASARIM BİLGİ**SİNE** SAHİP	Stemming
NESNE YÖNELİK ANALİZ TASARIM **BİLGİ SAHİP**	Non-effective term filtering
NESNE YÖNELİK ANALİZ TASARIM	(Final version)

When the experimental results were analyzed, it was observed that a large eps value caused features from different clusters to be included within the same neighborhood, resulting in distinct clusters being incorrectly merged into a single cluster. This also resulted in a reduced ability to detect outliers and a lower overall clustering quality. For instance, when eps was set to 1.0 with a min_samples of 50, the DBSCAN detected 192 outliers; however, it generated just two clusters. In contrast, small eps values resulted in too many small clusters and an excessive number of points being mislabeled as noise. Accordingly, the Silhouette Score facilitated the selection of the most optimal and balanced eps and min_samples values. These parameter values are determined automatically according to the best Silhouette score they achieved. Figure 2 shows the silhouette coefficient plot generated in the course of clustering the SS Dataset features.

Table 6 gives an example feature cluster obtained for the SS Dataset with the best parameter settings. This cluster contains similar features from different job postings about object-oriented programming skills.

After applying the DBSCAN algorithm to all datasets, the number of clusters obtained is summarized in Table 7. The total feature counts are directly proportional to the job posting counts given in Table 1. At the same time, cluster counts are also directly proportional to the total feature counts. The features that could not form a cluster were labeled as outliers and not included in further association rule mining studies. In some datasets, more than half of the features were labeled as outliers. An examination of datasets that yield a large number of outliers after the clustering process reveals that they contain more features with detailed information, including ad-specific descriptions, brand identifiers, and corporate details, which are inherently difficult to cluster.

Table 5. Best parameters for the DBSCAN Algorithm

Dataset	eps	min_samples
SC	0.5	5
SM	0.6	5
HRS	0.6	5
Sec.	0.6	5
AS	0.6	5
SS	0.7	5
IE	0.6	5

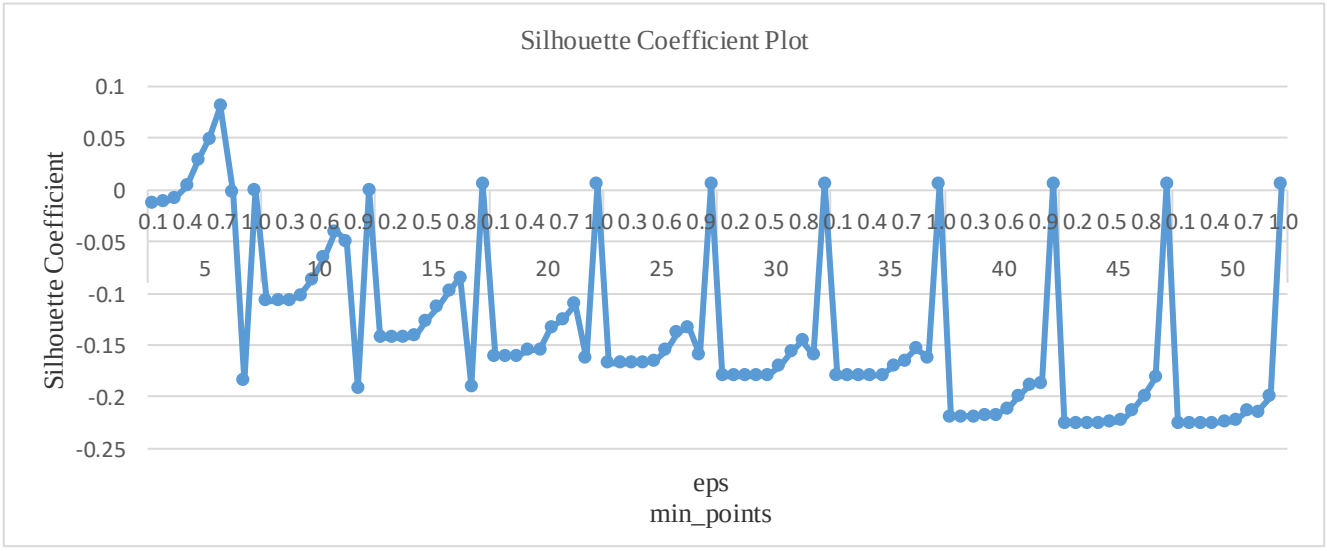


Figure 2. Silhouette coefficient plot for SS Dataset with eps 0.1 to 1 and min_points 5 to 50.

Table 6. A sample cluster for the SS dataset.

Job Posting ID	Feature (TR/EN)
2451428	NESNE YÖNELİMLİ PROGRAMLAMA KONUSUNDA TECRÜBELİ (EXPERIENCED IN OBJECT ORIENTED PROGRAMMING)
2443985	OOP NESNE YÖNELİMLİ PROGRAMLAMA KONUSUNA HAKİM (MASTERING OBJECT ORIENTED PROGRAMMING)
2350837	TERCİHEN NESNE YÖNELİMLİ PROGRAMLAMA DENEYİMİ OLAN (PREFERRED WITH OBJECT ORIENTED PROGRAMMING EXPERIENCE)
2432910	NESNE YÖNELİMLİ PROGRAMLAMA KONUSUNDA DENEYİMLİ (EXPERIENCED IN OBJECT ORIENTED PROGRAMMING)
2366737	NESNE YÖNELİMLİ PROGRAMLAMA MANTIĞINI ÇÖZMÜŞ (SOLVED OBJECT-ORIENTED PROGRAMMING LOGIC)
2426607	NESNE YÖNELİMLİ PROGRAMLAMA BİLEN (KNOWS OBJECT ORIENTED PROGRAMMING)
2414824	NESNE YÖNELİMLİ PROGRAMLAMA YAPABİLEN (CAPABLE OF OBJECT-ORIENTED PROGRAMMING)

Table 7. Cluster counts obtained for each dataset

Dataset	Total Feature Count	Cluster Count	Outlier Count
SC	44748	1326	8115 (18%)
SM	31063	1215	4960 (16%)
HRS	26483	831	15219 (57%)
Sec.	13677	509	7331 (54%)
AS	17672	513	10265 (58%)
SS	8691	289	4651 (54%)
IE	3419	93	2197 (64%)

Table 8 presents the extracted features from a sample job posting, the cluster IDs assigned by the DBSCAN Algorithm, and the automatically assigned cluster categories based on their content. The features that could not be assigned to any cluster have been labeled as -1. The cluster IDs except -1 indicate the cluster associated with that ID. Cluster IDs have been used in Apriori Algorithm to discover cluster associations and to suggest features for new job postings.

We reduced our cluster analysis problem to the market basket analysis and represented each job posting as a market transaction and each feature cluster in it as an item. We generated input data that contains an entity for each job posting for the Apriori algorithm. For example, we denoted the job posting exemplified in Table 8 as "5 32 33 272 2 183 11 3" to indicate its feature clusters.

We have computed both association rules and frequent item sets to discover relations between feature clusters. The minimum support and minimum confidence threshold values have been determined for each dataset individually to gather the necessary number of associations. For example, a minimum support threshold value of 1 and a minimum confidence threshold value of 50 were assigned for the SS Dataset. The minimum number of items in an item set was set to 2. In these settings, we obtained 32,475 lines of frequent item sets and strong association rules for the SS Dataset. The item sets and association rules that have the top 3 support and confidence values are exemplified in Table 9.

After the execution of the Apriori algorithm for all datasets, we obtained a list of associated feature clusters for each. Then, these association lists were used in feature suggestion studies.

Table 8. Extracted features from a sample job posting from the SS Dataset and their cluster IDs.

Feature (TR/EN)	Cluster ID	Cluster Category
TERCİHEN BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜNDEN MEZUN (PREFERRED TO GRADUATE FROM COMPUTER ENGINEERING DEPARTMENT)	5	Graduation
C# VİSUAL BASİC VE MACRO DİLLERİ İLE EN AZ 2 YIL KOD GELİŞTİRMİŞ (AT LEAST 2 YEARS OF DEVELOPMENT EXPERIENCE WITH C# VISUAL BASIC AND MACRO LANGUAGES)	-1	Hard Skills
ERP PROJELERİNDE YER ALMIŞ (ATTENDED IN ERP PROJECTS)	32	Hard Skills
MS SQL SERVER PROGRAMLAMA BİLGİSİNE SAHİP (KNOWS MS SQL SERVER PROGRAMMING)	33	Hard Skills
VERİ TABANI VE NESNEYE DAYALI PROGRAMLAMA ÜZERİNE BİLGİ SAHİBİ (KNOWS DATABASE AND OBJECT-ORIENTED PROGRAMMING)	272	Hard Skills
ANALİTİK DÜŞÜNEN VE ÇÖZÜM ODAKLI ÇALIŞAN (ANALYTICAL THINKING AND SOLUTION-FOCUSED)	2	Soft Skills
NETSİS CRM VE PDM PROGRAMLARINDA TECRÜBELİ (EXPERIENCED IN NETSIS CRM AND PDM PROGRAMS)	183	Hard Skills
ERKEK ADAYLAR İÇİN ASKERLİĞİNİ YAPMIŞ VEYA EN AZ 2 YIL TECİLLİ OLAN (FOR MALE CANDIDATES, TO HAVE MADE THE MILITARY OR DELAYED FOR 2 YEARS)	11	Military Status
İYİ DÜZEYDE İNGİLİZCE BİLEN (SPEAKING ENGLISH GOOD)	3	Foreign Language

Table 9. Examples of the resulting frequent item sets and association rules for SS Dataset.

Top 3 Item sets

Item set	Support
{5, 11}	20.65
{44, 11}	12.50
{2, 11}	8.88

Top 3 Association Rules

Association rule	Support	Confidence
11 <- 246	1.27	100
5 <- 183	1.09	100
44 <- 259 5 11	1.09	100

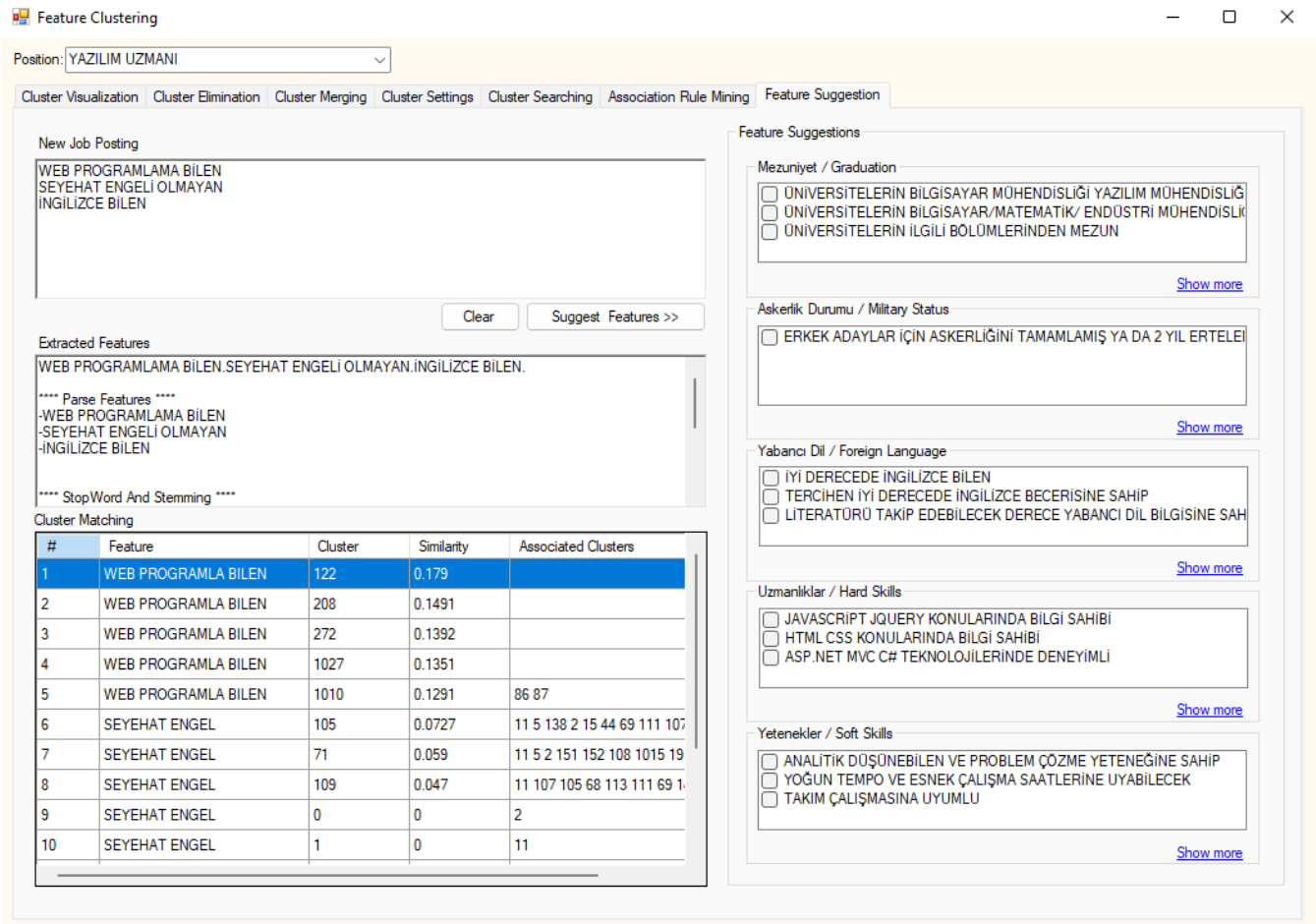


Figure 3. The user interface for the feature recommendation system and suggestions for a new Software Specialist job posting text.

The proposed system has been designed for recommendation features while a user is creating a new job posting. A Windows Form application shown in Figure 3 has been developed to conduct user experiments for pilot studies. When a user specifies the job title and enters some text including one or more features, firstly, the system applies the same preprocessing tasks explained in Section 3.1 on the given text and extracts its features. Then, the extracted features have been matched with the top five clusters they are most similar to. Similarity scores have been calculated by using the cosine similarity metric, which is often used to measure document similarity in text analysis [30]. Finally, clusters to suggest have been determined according to the pre-calculated associated cluster lists. Feature suggestions are presented to the users on the right side of the window under the categories mentioned in Section 3.4. By default, at most three features are suggested under each category. However, users can monitor up to ten feature recommendations by clicking on the "Show more" choice. They can append the features they select from the suggested features to the job posting text by clicking on them.

5. Conclusion and Future Works

Creating an effective job posting to attract top talents requires expertise and knowledge in the field. This paper presents an automatic feature suggestion system to assist recruiters who prepare job posting text. The dataset was supplied by Kariyer.net, and 7 different job titles, including software specialist, industrial engineering, accounting specialist, sales consultant, human resources specialist, store manager, and secretary, were selected as a case study. Text cleaning, feature extraction, and text preprocessing operations were applied to the job postings belonging to selected job titles. A feature clustering task was performed by using the DBSCAN clustering algorithm for each job title separately to obtain domain-specific clusters.

Each cluster has been represented by a cluster ID, feature text, and category ID (for Graduation, Military Status, Foreign Language, Hard Skills, and Soft Skills). We employed the Apriori algorithm to discover the frequently seen feature clusters together in previous job postings. We assumed each job posting was a market transaction, and each cluster in it was an item. The discovered relations were used in the feature recommendation stage of the study.

While a user is typing the job posting text, the text is processed instantly, and the existing feature clusters in the text are detected. The cosine similarity score is used to determine which clusters currently represent the specified features in the text. Frequent itemsets and association rules generated by the Apriori algorithm are used to recommend the features most related to the existing features.

The system was tested across all datasets, and thanks to the domain-specific rules it generated, it produced accurate recommendations for each job title. In this way, the proposed system streamlines recruiters' work and produces higher-quality, standardized job postings. To the best of our knowledge, this is the first study to apply the association rule mining and text clustering techniques together on job posting texts.

While the findings are encouraging, one limitation of this study is that the system operates offline. New job postings appended to

the dataset require the generation of clusters and rules from scratch. Currently, these procedures is triggered automatically every night. Therefore, future research will focus on improving the system to operate online and maintaining the cluster associations with each new job posting. Future improvement scopes for our research also include extending the domain space for all job titles, integrating with Kariyer.Net's online recruitment platform, analyzing the model's usage statistics, and collecting user feedback.

References

[1] Califf ME, Mooney RJ. Bottom-up relational learning of pattern matching rules for information extraction. J Mach Learn Res 2003;4:177-210. doi:10.1162/153244304322972685.

[2] Muludi K, Widyantoro DH, Santoso OS. Multi-inductive learning approach for information extraction. In: Proceedings of the International Conference on Electrical Engineering and Informatics, IEEE, Bandung, Indonesia; 2011, p. 1-6. doi:10.1109/ICEEI.2011.6021680.

[3] Lavelli A, Califf ME, Ciravegna F, Freitag D, Giuliano C, Kushmerick N, et al. Evaluation of machine learning-based information extraction algorithms: criticisms and recommendations. Lang Resour Eval 2008;42(4):361-393. doi:10.1007/s10579-008-9079-3.

[4] Elgammal Z, Barmu A, Hassan H, Elgammal K, Özyer T, Alhajj R. Matching applicants with positions for better allocation of employees in the job market. In: Proceedings of the 22nd International Arab Conference on Information Technology (ACIT), IEEE, Muscat, Oman; 2021, p. 1-5. doi:10.1109/ACIT53391.2021.9677374.

[5] Sandanayake TC, Limesha GAI, Madhumali TSS, Mihirani WPI, Peiris MSA. Automated CV analyzing and ranking tool to select candidates for job positions. In: Proceedings of the 6th International Conference on Information Technology: IOT and Smart City, Osaka, Japan; 2018, p. 13-18. doi:10.1145/3301551.3301579.

[6] Mhamdi D, Azzouazi M, El Ghoumari MY, Moulouki R, Rachik Z. Enhancing recruitment process using semantic matching. In: Proceedings of the International Conference on Advanced Intelligent Systems for Sustainable Development, Springer, Tangier, Morocco; 2020, p. 370-378. doi:10.1007/978-3-030-90639-9_29.

[7] Menacer MA, Hamda FB, Mighri G, Hamidene SB, Cariou M. An interpretable person-job fitting approach based on classification and ranking. In: Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP), Trento, Italy; 2021, p. 130-138.

[8] Ahmed Awan MN, Khan S, Latif K, Khattak AM. A new approach to information extraction in user-centric e-recruitment systems. Appl Sci 2019;9(14):2852. doi:10.3390/app9142852.

[9] Mahbub S, Pardede E. Using contextual features for online recruitment fraud detection. In: Proceedings of the 27th International Conference on Information Systems Development, Lund, Sweden; 2018.

[10] Nasser IM, Alzaanin AH, Maghari AY. Online recruitment fraud detection using ANN. In: Proceedings of the Palestinian International Conference on Information and Communication Technology (PICICT), IEEE, Gaza, Palestine; 2021, p. 13-17. doi:10.1109/PICICT53635.2021.00015.

[11] Mehboob A, Malik MSI. Smart fraud detection framework for job recruitments. Arab J Sci Eng 2021;46(4):3067-3078. doi:10.1007/s13369-020-04998-2.

[12] Chern A, Liu Q, Chao J, Goindani M, Javed F. Automatically detecting errors in employer industry classification using job postings. Data Science and Eng 2018;3(3):221-231. doi:10.1007/s41019-018-0071-7.

[13] Wu X, Xu T, Zhu H, Zhang L, Chen E, Xiong H. Trend-aware tensor factorization for job skill demand analysis. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI), Macao, China; 2019, p. 3891-3897. doi:10.24963/ijcai.2019/540.

[14] Xu T, Zhu H, Zhu C, Li P, Xiong H. Measuring the popularity of job skills in recruitment market: A multi-criteria approach. In: Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA; 2018, p. 2572-2579. doi:10.1609/aaai.v32i1.11847.

[15] Li S, Shi B, Yang J, Yan J, Wang S, Chen F, et al. Deep job understanding at LinkedIn. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Xian, China; 2020, p. 2145-2148. doi:10.1145/3397271.3401403.

[16] Cao L, Zhang J, Ge X, Chen J. Occupational profiling driven by online job advertisements: Taking the data analysis and processing engineering technicians as an example. Plos One 2021;16(6):e0253308. doi:10.1371/journal.pone.0253308.

[17] Xu G. Human resource recommendation based on recurrent convolutional neural network. In: Proceedings of the 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE), IEEE, Guangzhou, China; 2022, p. 54-58. doi:10.1109/ICCECE54139.2022.9712799.

[18] Wild S, Parlar S, Hanne T, Dornberger R. Naïve Bayes and named entity recognition for requirements mining in job postings. In: Proceedings of the 3rd International Conference on Natural Language Processing (ICNLP), IEEE, Beijing, China; 2021, p. 155-161. doi:10.1109/ICNLP52887.2021.00032.

[19] Guo S, Alamudun F, Hammond T. RésuMatcher: A personalized résumé-job matching system. Expert Syst Appl 2016;60:169-182. doi:10.1016/j.eswa.2016.04.013.

[20] Çelik D. Towards a semantic-based information extraction system for matching résumés to job openings. Turk J Electr Eng Comput Sci 2016;24(1):141-159. doi:10.3906/elk-1304-130.

[21] Tuncer I, Kara KC, Karakaş A. Improving search relevance with word embedding based clusters. In: Proceedings of the International Conference on Artificial Intelligence and Applied Mathematics in Engineering, Springer, Antalya, Turkey; 2020, p. 15-24. doi:10.1007/978-3-030-79357-9_3.

[22] Kara KC, Esen S, Kahyalar N, Karakaş AA, Aytekin T. Design and implementation of a job recommender system. In: Proceedings of the International Conference on Computer Science and Engineering (UBMK), IEEE, Antalya, Turkey; 2017, p. 729-733. doi:10.1109/UBMK.2017.8093515.

[23] Akın AA, Akın MD. Zemberek, An open source NLP framework for Turkic languages. Structure 2007;10:1-5.

[24] Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, AAAI Press, Portland, Oregon, USA; 1996, p. 226-231.

[25] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. ACM SIGMOD Record 1993;22(2):207-216. doi:10.1145/170035.170072.

[26] Agrawal R, Srikant R. Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases, Santiago de Chile, Chile; 1994, p. 487-499.

[27] Ilayaraja M, Meyyappan T. Mining medical data to identify frequent diseases using Apriori algorithm. In: Proceedings of the International Conference on Pattern Recognition, Informatics and Mobile Engineering, Salem, India; 2013, p. 194-199. doi:10.1109/ICPRIME.2013.6496471.

[28] Yang Y, Yuan ZZ, Sun DY, Wen XL. Analysis of the factors influencing highway crash risk in different regional types based on improved Apriori algorithm. Adv in Transportation Studies 2019;49:165-178.

[29] Dalkılıç F, Aydın Ö. The factors affecting the non-attendance behaviors of the students in the Faculty of Economics and Administrative Sciences in Dokuz Eylül University. Journal of Higher Education and Science 2017;7(3):546-553. doi:10.5961/jhes.2017.231.

[30] Manning CD, Raghavan P, Schutze H. Introduction to information retrieval. New York: Cambridge; 2008, p. 121-126.