INTERNATIONAL JOURNAL OF APPLIED MATHEMATICS ELECTRONICS AND COMPUTERS

www.dergipark.org.tr/ijamec

e-ISSN: 2147-8228

International Open Access

Volume 08
Issue 03

September, 2020

Research Article

Discovering the same job ads expressed with the different sentences by using hybrid clustering algorithms

Yunus Doğan [a,*], ***Feriştah Dalkılıç*** [a], ***Alp Kut*** [a], ***Kemal Can Kara*** [b], ***Uygar Takazoğlu*** [b]

[a]*Computer Engineering Department, Faculty of Engineering, Dokuz Eylül University, Izmir, Turkey*
[b]*Kariyer.net R&D Centre, Istanbul, Turkey*

ARTICLE INFO

Article history:
Received 18 September 2020
Accepted 30 September 2020

Keywords:
Hybrid clustering algorithms
Job ads
Machine learning
Natural language processing

ABSTRACT

Text mining studies on job ads have become widespread in recent years to determine the qualifications required for each position. It can be said that the researches made for Turkish are limited while a large resource pool is encountered for the English language. Kariyer.Net is the biggest company for the job ads in Turkey and 99% of the ads are Turkish. Therefore, there is a necessity to develop novel Natural Language Processing (NLP) models in Turkish for analysis of this big database. In this study, the job ads of Kariyer.Net have been analyzed, and by using a hybrid clustering algorithm, the hidden associations in this dataset as the big data have been discovered. Firstly, all ads in the form of HTML codes have been transformed into regular sentences by the means of extracting HTML codes to inner texts. Then, these inner texts containing the core ads have been converted into the sub ads by traditional methods. After these NLP steps, hybrid clustering algorithms have been used and the same ads expressed with the different sentences could be managed to be detected. For the analysis, 57 positions about Information Technology sectors with 6,897 ad texts have been focused on. As a result, it can be claimed that the clusters obtained contain useful outcomes and the model proposed can be used to discover common and unique ads for each position.

1. Introduction

Today, all sectors have witnessed a continuous transformation and, like many other fields, could not escape the digitization period. Although the digitization of the job market has led to positive improvements in the interaction of recruiters and candidates, the amount of data (job ads) produced every day has become so large that it has become impossible to manually examine them. With the increasing number of electronic documents and the rapid growth of the internet, the task of automatically categorizing documents has become the critical method for detecting and preparing information for usages. Machine learning algorithms can offer a sustainable solution. In fact, the most important benefits of this technology are speed and efficiency. If we go back to the studies for English ads, we see that in the study conducted in 2010, the extraction of the qualities started to be realized with a linguistic approach, namely natural language processing methods (NLP) [1]. Then we can see that machine learning algorithms are used. For example, in a doctoral dissertation in 2018, skill mining was carried out in job ads using Regression and Artificial Neural Networks besides NLP, and suitable candidates were matched [2]. In some studies, it has been observed that it is made dependent on the field. For example, in the study in 2019, "skill recognition" was carried out on the German language and only in Computer Science with machine learning algorithms such as supervised learning algorithms like Random Trees [3].

It can be said that the researches made for Turkish are limited while a large resource pool is encountered for the English language. In this study, up-to-date Turkish job ads having a large size have been analyzed by machine learning algorithms and the same ads expressed with the different

[*] Corresponding author. E-mail address: *yunus@cs.deu.edu.tr*
DOI: 10.18100/ijamec.797572

sentences could be managed to be detected.

In this study, the job ads of Kariyer.Net have been analyzed, and by using hybrid clustering algorithm, the hidden associations in this dataset as the big data have been discovered. Firstly, all ads in the form of HTML codes have been transformed into regular sentences by the means of extracting HTML codes to inner texts. Then, these inner texts containing the core ads have been converted into the sub ads by traditional Natural Language Processing (NLP) methods. After these NLP steps, hybrid clustering algorithms have been used and the same ads expressed with the different sentences could be managed to be detected.

The next sections are about related works, methodologies, and experimental studies with the results separately.

2. Related Works

Recently, rapid changes in business life have forced employees to keep up with these new conditions. Adapting to these changes requires not only technical skills but also many skills and competencies. Analysis studies on job postings were generally used to determine the qualifications needed for many positions. For example, in a study by Kennan et al., They focused on the information systems jobs of Australian employers and IT technology graduates. In the results, it was observed that communication skills and personal characteristics were also very important apart from IT knowledge, skills, and competencies [4]. In his study, Choi and Rasmussen determined the priority qualifications for digital library positions by focusing on academic libraries by examining job postings. In the results obtained, management, communication skills and digital technology competencies were put forward as required qualifications [5]. In another study, Pember concluded that record keeping experts should have experience other than the knowledge and skills they need in their articles. Apart from that, they have achieved that record keeping positions should have proficiency in various areas of information management [6]. For example, record keeping professionals must have skills such as good computer use, well-developed communication, and leadership skills, acquiring personnel management skills and experience, a good level of teamwork and strong customer focus. Abstract concepts such as motivation and enthusiasm for work and personal characteristics such as analytical problem-solving have also become prominent talents. These talents have been revealed in the study conducted in many other positions such as civil engineering [7]. On the other hand, some studies have focused on the competence criteria of the training of employees in any sector, except for the consequences of what qualifications are needed. For example, Kwon Lee and Han have been observed that for the United States labor market, most universities place great emphasis on education with courses that have the most operating systems and hardware content. However, according to the results obtained from the study, it has been concluded that employers do not care so much about these qualities [8]. Yongbeom et al. It also addresses the gap between employers 'needs and universities' IT curricula [9].

In addition, there are many studies analyzing job ads to obtain useful information in decision making with text mining. For instance, data mining can be preferred to identify competencies arising in companies and to promote employment opportunities or career development [10]. The system of classification of business areas has been used and it has been able to be placed in a system that varies from traditional to a flexible structure, e.g., market changes [11] or focused on studies that can make quick decision makings based on the changes observed in the job market [12]. Text mining or document-based analysis requires algorithms such as hidden semantic analysis or average link to obtain meaningful information from large amounts of text data. Compared to manual content analysis, document-based analysis takes less time and is cheaper [13].

Document-based analysis involves several pre-processing steps used to clear text using techniques such as removing unnecessary words and roots of corpus. After the pre-processing, document-based analysis algorithms extract words from the cleared text of job ads. The extracted words are processed by cluster analysis [14], hidden semantic analysis [15-16], classification algorithms such as support vector machines and random trees [17-18], open rules, and hidden Dirichlet allocation algorithms [17]. Document-based analysis has also been used for various purposes, such as researching consumer perceptions of hotels based on online customer textual texts [19] and using social media such as the texts on Facebook and Twitter to conduct competition [20]. Examples of using social media broadcasts for product planning [21] and big data analysis in the financial sector [22] are also found.

Document-based analysis researches that analyze job ads can also be combined in 3 clusters. The cluster 1 is to use document-based analysis for the analysis of job ads to implement a novel scheme clustered of job ads and compare them with the traditional occupational system clustered [15-17]. For instance, Mezzanzanica [12] used document-based analysis to investigate job ads in marketing. The corpus contained the job ads in Italy and the texts were in Italian. The researchers evaluated the ESCO taxonomy about job classification and data mining algorithm on over 1.9 million job ads because of trends and dynamics observed in the evolution of the labor market and identified several potential professions that emerged. The cluster 2 of authors used document-based analysis to increase the caliber of work compliance with potential candidates based on 'time to commute workplace, type of work, hourly wages, and the candidate's skill set [14]. These studies consider that inadequate matching of candidates to job positions can cost organizations significantly; therefore, document-based analysis is needed. The researchers in the cluster 3 implemented document-based algorithms to obtain the work profiles for certain vertical fields [16], information management [23] and big data [24], or horizontal fields such [25-26].

There are two methods for analyzing job ads. One of them is manual content analysis [27] and the other one is automatic text analysis, often called text mining or document-based analysis [28]. Document-based analysis has significant advantages over manual content analysis, such as less time needed for analysis and human resource [29]. In addition, there are examples in the literature that use social media about document-based analysis to gain a competitive advantage in various organizations [30-31]. There are also studies that have been observed to increase marketing efficiency with document-based analysis as well [32-33]. Document-based analysis methods are widely used to analyze information stored on social media websites such as Twitter tweets [34]. Therefore, this study focused on document-based analysis techniques for analyzing job ads on Kariyer.Net platform, which can be interpreted as a social media platform.

Job postings contain unstructured texts that make analysis difficult. To successfully cluster these postings, grammatical and syntax errors must also be addressed, and this is where machine learning algorithms can make the difference. Machine learning algorithms need a lot of appropriate text data to learn before models are created for clustering job postings. One of the first phases to using the corpus is to analysis the data in advance. In other words, having data ready for analysis is a very important step. Most of the available corpus is highly unstructured and contains an incomplete and noisy content. It is necessary to clear the data to obtain healthy inferences and create better algorithms. For example, job posting data, although careful, is not yet standard and can be interpreted as informal. In this case too, spelling errors, bad grammar, URLs, words to be blocked, expressions, etc. The presence of undesirable content, such as, are the usual suspects.

In addition to these, there is an inevitable presence of other HTML commands in the text because it contains HTML for the data obtained from the web. Besides, advertisement data, "Latin", "UTF8" etc. It may be subject to various decoding formats such as. Therefore, it needs to be decoded. By converting information from these complex symbols to simple and easy-to-understand characters, all data are kept in standard coding for better analysis. UTF-8 encoding is widely accepted and recommended for use. Finally, another thing that must be operated is to delete ineffective words and stop words. Commonly found words, ineffective words should be ignored when data analysis needs to be directed to data at the word level. Also, punctuation marks need to be removed. All punctuation marks should be handled according to priorities. For example: ".", ",","?" important punctuation marks, while others are what should be removed [35].

3. Methodologies

Firstly, all ads in the form of HTML codes transformed into regular sentences by the means of extracting HTML codes to inner texts. This implementation has been coded by the following algorithm.

```
Function string[] getSubFeatures(string _ad)
string ilanlar = "";
_ad = _ad.Replace("&nbsp;", " ").Replace("    ", " ").Replace("   ", "
").Trim().ToLower().Replace("<      br", "<br").Replace("<      li",
"<li").Replace("</ li", "</li").Replace("< /li", "</li").Replace("< p",
"<p").Replace("</ p", "</p").Replace("< /p", "</p");

string ayrac = ".";  string bitis = "";
if (_ad.Contains("<li")) {ayrac = "<li";bitis = "</li";}
else if (_ad.Contains("<p")){ayrac = "<p";bitis = "</p";}
else if (_ad.Contains("<br")){ayrac = "<br";}
else if (_ad.Split('+').Length > 1){ayrac = "+";}
else if (_ad.Split('*').Length > 1){ayrac = "*";}
else if (_ad.Split('-').Length > 1){ayrac = "-";}
else if (_ad.Split(',').Length > 1){ayrac = ",";}
else if (_ad.Split(';').Length > 1){ayrac = ";";}
else if (_ad.Split('·').Length > 1){ayrac = "·";}

while (_ad.Length > 1){
if (ayrac.Contains("<")){_ad = _ad.Remove(0, _ad.IndexOf(ayrac) +
ayrac.Length); _ad = _ad.Remove(0, _ad.IndexOf(">") + 1);}
else if (_ad.IndexOf(">") <_ad.IndexOf(ayrac) &&
_ad.IndexOf(">") >-1)
{_ad = _ad.Remove(0, _ad.IndexOf(">") + 1);
if (bitis.Length > 0 && _ad.Contains(bitis))
{string inp = ExtractHtmlInnerText(_ad.Substring(0,
_ad.IndexOf(bitis))).Trim();

if (inp.Length > 0 && (inp.EndsWith("•") || inp.EndsWith("·") ||
inp.EndsWith(",") || inp.EndsWith(";") || inp.EndsWith(".") ||
inp.EndsWith(":") || inp.EndsWith("-") || inp.EndsWith("+") ||
inp.EndsWith("*")))    {inp = inp.Substring(0, inp.Length - 1);}

if (inp.Length > 0 && (inp.StartsWith("•") || inp.StartsWith("·") ||
inp.StartsWith(",") || inp.StartsWith(";") || inp.StartsWith(".") ||
inp.StartsWith(":") || inp.StartsWith("-") || inp.StartsWith("+") ||
inp.StartsWith("*")))    {inp = inp.Substring(1);}

if (inp.Length > 1)ilanlar += inp.Trim() + "|";
if (bitis == "</li") _ad = _ad.Remove(0, _ad.IndexOf("/li>") + 4);
else if (bitis == "</p") _ad = _ad.Remove(0, _ad.IndexOf("p>") + 2);
else if (bitis.Length == 0 && _ad.Contains(ayrac))
{string inp = ExtractHtmlInnerText(_ad.Substring(0,
_ad.IndexOf(ayrac))).Trim();

if (inp.Length > 0 && (inp.EndsWith("•") || inp.EndsWith("·") ||
inp.EndsWith(",") || inp.EndsWith(";") || inp.EndsWith(".") ||
inp.EndsWith(":") || inp.EndsWith("-") || inp.EndsWith("+") ||
inp.EndsWith("*"))){inp = inp.Substring(0, inp.Length - 1);}

if (inp.Length > 0 && (inp.StartsWith("•") || inp.StartsWith("·") ||
inp.StartsWith(",") || inp.StartsWith(";") || inp.StartsWith(".") ||
inp.StartsWith(":") || inp.StartsWith("-") || inp.StartsWith("+") ||
inp.StartsWith("*"))){inp = inp.Substring(1);}

if (inp.Length > 1){ilanlar += inp.Trim() + "|";
 _ad = _ad.Remove(0, _ad.IndexOf(ayrac) + ayrac.Length);}
else {_ad = ExtractHtmlInnerText(_ad).Trim();}
if (_ad.Length > 1){ilanlar += _ad + "|";break;}
if (ilanlar.Length > 0){ilanlar = ilanlar.Substring(0,ilanlar.Length -1);}

return ilanlar.Split('|');
```

Then, these inner texts containing the core ads have been converted into the sub ads by traditional National Language Processing (NLP) methods as given in the following algorithm. In this function, Zemberek library has been used and the roots could be obtained for each word.

```
Function string[] getRoots(string ads)
    List<string> lstSW = new List<string>();

    Zemberek zemberek =new Zemberek(new TurkiyeTurkcesi());
    List<string> _stopWords = stopWords.ToList<string>();
    string[] adx = ads.Split(' ');

    foreach (string ad in adx)
      if (!lstSW.Contains(ad) && zemberek.kelimeDenetle(ad))
      {string kok = zemberek.kelimeCozumle(ad)[0].kok().icerik();
      if (!lstSW.Contains(kok) && !_stopWords.Contains(kok))
      {lstSW.Add(kok);}}

return lstSW.ToArray();
```

In the next steps, Density-based spatial clustering of applications with noise (DBSCAN) for obtaining the noises; X-means, which contains K-means++ algorithm without any the cluster number as a parameter, for obtaining the initial centroids; Self Organizing Map (SOM) as a novel usage of SOM named Improved Parallel SOM (iPSOM) for the eventual results have been used.

The K-means++ algorithm is a method based on the main idea that the centre point represents the set. It tends to find global clusters of equal size. According to the mechanism of operation of the K-means++ algorithm, first, *k* objects are selected to represent the centre or mean of a set.

The remaining objects are included in the clusters to which they are most similar, considering the distance from the clusters' mean values. Then, by calculating the average value of each cluster, new cluster centres are determined, and object centre distances are examined again. Total square error criterion SSE (Summed Squared Error) is most used in the evaluation of the K-means++ clustering method. The clustering result with the lowest SSE value gives the best result. The sum of squares of the distance of objects from the centre points of the cluster in which they are found is calculated by Eq. 1.

$$\mathrm{SSE} = \sum_{k=1}^{K} \sum_{i=1}^{I} (xi - ck)^2 \qquad (1)$$

Here, the standard Euclidean Distance (*ED*) between two objects is an object whose x value is in the Ci set, the mi value is the centre point of the Ci set.

The K-means++ algorithm described above works according to the *ED* criterion on two-dimensional data and is shifted until no object set leaves. However, the structure of this K-means++ algorithm is not suitable for web applications. Since comparing whether there is an object leaving the cluster in every translation will cause time negativity in large data sets, a K-means++ version based on objective function has been preferred and this algorithm has been made to work on multidimensional data to cluster web pages. First, it is ensured that the vector representing each document is called in order since it is not possible to memorize and process all the data. The Cosine Similarity criterion was added to calculate the distance of these vectors to the cluster centres by different methods.

The DBSCAN algorithm is based on revealing the neighbours of data points in two or multi-dimensional space. The database is mostly used in the analysis of spatial data since it deals with a spatial perspective. For the DBSCAN algorithm, the terms core object, Eps, MinPts, direct density accessible point, density accessible point, density bound point are basic concepts. It takes the algorithm, Eps and MinPts values as input parameters. Starting from any object in the database, it checks all objects. If the checked object has already been included in one set, it moves to the other object without any action. If the object has not been previously clustered, it performs a Region Query and finds its neighbours in the Eps neighbourhood. If the number of neighbours is more than MinPts, it will call this object and its neighbours a new cluster. It then finds new neighbours by making a new zone query for each neighbour that is not already clustered. If the neighbour numbers of the points where the region is questioned are more than MinPts, they are included in the cluster. Neighbourhood discovery is the most demanding part of the DBSCAN algorithm.

Performance improvements in this section significantly increase the performance of the algorithm. In the neighbourhood analysis, instead of examining every point, various indexing algorithms such as R * tree or spatial query have been introduced. With these algorithms, the complexity of the DBSCAN algorithm from O (n * log n) to O (log n) can provide significant performance increases. Since the DBSCAN algorithm takes two parameters, Eps and MinPts, it has been applied 7 times with different parameters to see the effect of both parameters on the cluster result. Unlike K-means++, DBSCAN algorithm does not include every element of the database in a cluster, it could filter the exception data. Values determined by the algorithm as noise (exception) are not shown in the result graphs. When very small value is given to Eps neighbourhood distance, only very dense cluster areas, in other words, cluster cores were found. When the EPS value is applied as 0.2, an unwanted small cluster has occurred near the 3rd cluster although very close to the ideal cluster has occurred.

SOM consists of two layers of artificial neurons: an input layer and an output layer. The input layer is fed into feature vectors, so it is the same as the number of dimensions of the input feature vector. Output layer, also called the output map, is usually arranged in a regular two-dimensional structure such that there are neighbourhood relations among the neurons. Every neuron in input layer is fully connected to every output neuron, and each connection has a weighting value attached to it.

Algorithm:

1. We randomly start the weight values of neurons in our network
2. We get the input vectors. (Our target vectors in the system)
3. All values on the map are roaming and:
4. The distance between the input vector and the current map value is calculated as Euclidean distance.
5. The node with the shortest distance is taken (this method is called the best matching unit (BMU))
6. All nodes adjacent to this best-fit node we selected are updated and brought closer to the input vector. (The following formula is used):
7. $Wv(t + 1) = Wv(t) + \Theta\ (t)\ \alpha(t)\ (D(t) - Wv(t))$
8. If $t < \lambda$, the operations are repeated by going to step 2.

where t is current step, λ is time limit on step, Wv is current weight vector, D is targeted input value, *(T)* is neighborhood function (how far to go from the most suitable neighbor) and $\alpha(t)$ is time dependent learning limit

The random choosing the data in the training phase of SOM is abandoned and the data are chosen in the same order for each time. The accuracy problem is passed over at connecting the subparts of maps.

This new approach first divides the area into four, and this standard SOM processes it by parallel processing for all small areas. Thus, datasets are split for all processes and complexity is reduced.

iPSOM consists of standard SOM (SSOM). This algorithm is used for 2x2 neurons stably in each phase. iPSOM starts training with SSOM with 2x2 neurons and usage of all dataset. After that, the recursive structure of iPSOM is activated and recursively, SSOM for 2x2 neurons is processed with usage of divided datasets. The following Figure 1 shows the process-flow of iPSOM for 4x4 neurons. There are four parallel maps and they are trained for 2x2 maps. At the end of the iPSOM, a 4x4 map is obtained after combining operation in the same order before splitting.

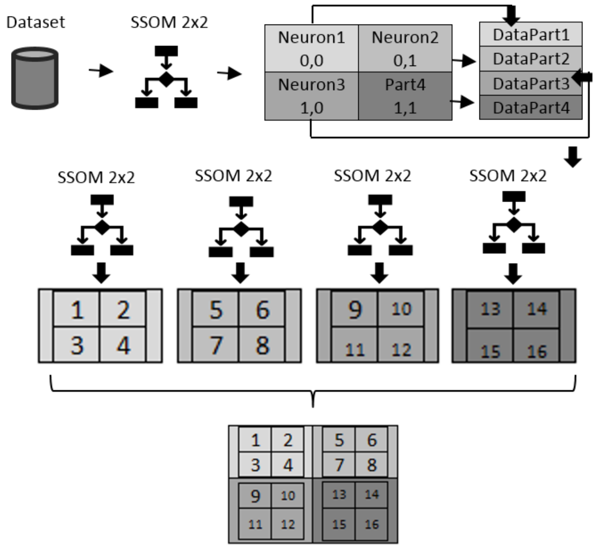

Figure 1. The process-flow of iPSOM for 4x4 neurons.

The complexity of SOM is $O(N^2)$. However, this formula is obtained by the assumption of the multiply of the map size and weight numbers equal to the multiply of the tuple number and attribute number in the dataset. Therefore, if N^2 is split to N.C as N is the total size of map and C is the total size of dataset, the changes of SOM speed according to the different datasets and its effects appears in more detail. When this formula is split sub-components, these following formulas in Eq. 2 and Eq. 3 are obtained.

$$F\ .\ A = C \tag{2}$$

$$M\ .\ W = N \tag{3}$$

where F is the tuple number in the dataset, A is the attribute number for a tuple in the dataset, M is the total neuron number in the map and W is the total number of weight variables for a neuron in the map.

$$F\ .\ A\ .\ M\ .\ W = \alpha. \tag{4}$$

where α is the total time for SSOM algorithm.

When iPSOM is processed for 4x4 neurons (M = 16), SSOM algorithm is processed for 2x2 neurons (M = 4) and all dataset in the first phase. Secondly, the datasets are divided into four pieces for each neuron according to the proximities to the weight values of neurons. All proximities calculations in the algorithm are done by ED formula in Eq. 5.

$$ED = \sqrt{\sum_{k=0}^{d} (X_k - Y_k)^2} \tag{5}$$

where X and Y are tuples in the dataset or neurons in the map and d is the attribute number if X and Y are tuples or d is the total number of weight variables if X and Y are neurons.

Four pieces of dataset are used for training by SSOM algorithm in parallel. SSOM is processed for 2x2 neurons (M = 4) for each piece again. After these pieces are trained, these pieces are joined and the map with 4x4 neurons is obtained. The process time calculation is done the following formulas:

$$F.A.\frac{M}{2x2}.W = \frac{F.A.M.W}{4} = \beta \tag{6}$$

$$\frac{F}{4}.\frac{A.M.W}{4} = \frac{F.A.M.W}{16} = ɵ \tag{7}$$

$$\frac{5.F.A.M.W}{16} = ɵ + \beta \tag{8}$$

If it is assumed that SSOM is processed for all dataset and 4x4 neuron, process time of SSOM is found by Eq. 4. For 4x4 neurons, iPSOM is processed for 2x2 neurons firstly and the process time of this phase is calculated by Eq. 6. The result shows that this phase is equals to quarter of SSOM. However, iPSOM continues to be trained for 4x4 neurons and in parallel and for four pieces of datasets, iPSOM is processed for 2x2 neurons again. The process time of this phase is calculated by Eq. 7. Totally, the process time is calculated by Eq. 8 and it shows that iPSOM takes less time than SSOM as 5/16 times or nearly 1/3 times. This difference is kept even if the size of the dataset gets larger. Also, if the map size gets larger, new components add to Eq. 8 and because the dataset is divided again and again, these process times are approximate to zero a lot like. Thus, this difference is kept. Theoretically, this is possible; however, the machine, where iPSOM is processed, must have enough numbers of cores to supply the parallel processing.

4. Experimental Studies and Results

For the analysis, 57 positions about Information Technology sectors with 6,897 ad texts have been focused on. Firstly, by the means of DBSCAN algorithm, 94,246 instances have been grouped to detect the noises. Principal component analysis (PCA) has been used to reduce the 2,682 attributes to 2. The result figure has been given in Figure 2.

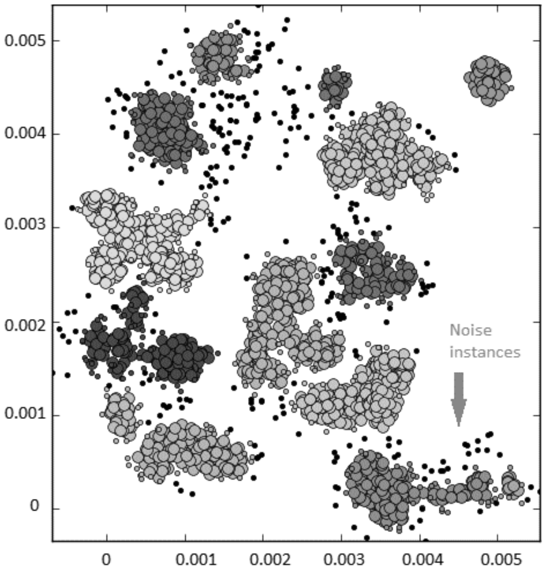

Figure 2. The detection of noises by using DBSCAN

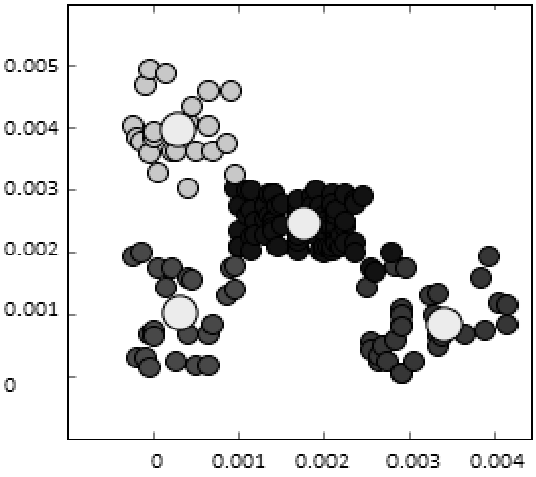

Figure 3. A sample clusters obtained by X-means

Then, X-means algorithm has been used for clustering to obtain the initial centroids for iPSOM with 73,372 instances reduced. A sample of PCA output is given in Figure 3. In this step, by iPSOM algorithm, 5,996 centroids have been obtained. The iPSOM result has 24x24 neurons and it has been given in Figure 4.

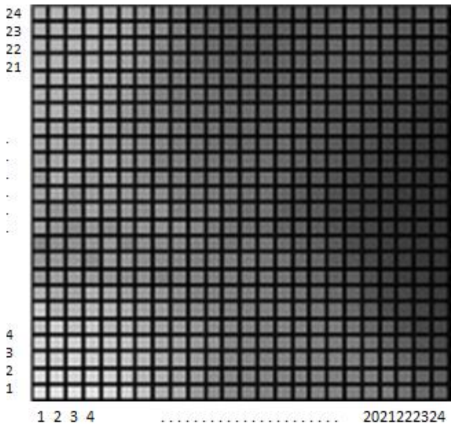

Figure 4. iPSOM output

The following clusters are sample instances in random 4 clusters in the result pattern.

Cluster1;

- good knowledge of optimization methods
- to have a good command of the standards and regulations related to the sectors
- good command of sql and database applications
- good command of standard time measurement (work and time study)
- good command of efficiency and oee measurement
- good command of erp systems methodology
- good command of database connection and functions on excel
- master the relational database logic
- good command of the workflow
- nebim mastered the erp system
- good command of nebim pos and campaign modules
- good command of optimization methods
- good command of sql server setup and sql query
- good command of windows server systems
- …

Cluster2;

- completed military service (for male candidates)
- male candidates must have completed military service
- completed military service in male candidates
- military service completed or exempt from military service for male candidates
- male candidates who have completed their military service or are waiting for payment
- completed military service for male candidates
- completed military service in male candidates
- candidates who have completed military service are sought for male candidates.
- you have completed military service (for male candidates)

- …

Cluster3;

- university graduate (preferably industrial engineering, computer engineering, computer programming, management information systems and equivalent)
- graduated from 4-year departments of universities (computer mathematics industrial business engineering and management information systems (müh or statistics)
- graduated from universities in industrial engineering, computer engineering, business engineering, management information systems or similar related departments
- graduated from related departments of universities preferably (computer engineering, industrial engineering or information systems management)
- graduated from business mathematics industrial engineering computer engineering or management information systems departments of universities
- …

Cluster4;

- has developed an application for android developer with at least 4 years of experience in Kotlin language.
- at least 3 years of experience has developed an application with swift language for ios developer
- at least 4 years of experience in application development with java
- at least 5 years of experience in developing applications using microsoft .net and c #
- at least 5 years of experience in developing applications using microsoft .net and c #
- large-scale web application development experience using net technologies
- at least 3 years of experience in web-based application development with phyton 2 and phyton 3
- 3 years or more of experience in developing ios mobile applications
- at least 3 years of experience in ios application development
- …

In the first set example, the technical skills needed in various positions could be grouped. In cluster 2, the same requirements related to military service obligation could be grouped with different advertisements. In the example of the third cluster, it was observed that the advertisements indicating the university graduation departments required for the positions were grouped. In cluster 4, it was observed that the postings containing the experience year requests for the positions were grouped. According to Table 1, the positions that require more than 5 and 5 years of experience are Business Development Manager, Hardware Design Team Leader, Business Development Manager, Senior Software Engineer and Hardware Specialist. Positions with the least expectation of experience were observed as Hardware Engineer, Hardware Support Specialist and Hardware Specialist, which were requested for 2.5 years or less.

Table 1. The Experience Year Requests for The Positions

The Positions	Average Year
Business Development Manager	6.19
Hardware Design Team Leader	5.50
Business Development Manager	5.32
Senior Software Engineer	5.02
Hardware Specialist	5.00
Hardware Design Engineer	4.78
Electronics engineer	4.69
Network Expert	4.62
Cyber Security Specialist	4.58
Senior Software Specialist	4.53
System Engineer	4.48
Responsible for information processing	4.24
SAP Consultant	4.16
Senior Software Development Specialist	4.14
Information Technology Specialist	4.11
Web Development Specialist	4.09
System Specialist	4.05
Electrical Electronics Engineer	3.99
Information Technology IT Specialist	3.97
Job Security Specialist	3.95
System Support Expert	3.93
Computing Expert	3.81
Information Security Specialist	3.78
Software Development Engineer	3.76
Web Software Specialist	3.50
Research and Development Engineer	3.44
Interface Software Specialist	3.42
Java Software Specialist	3.40
Business Development Specialist	3.36
Computer Engineer	3.33
Information Processing Staff	3.26
Industrial Engineer	3.22
Hardware Development Engineer	3.20
Graphic designer	3.15
Graphic Design Specialist	3.13
Research Development R&D Specialist	3.06
Hardware Support Staff	3.00
ERP Specialist	3.00
Software Support Specialist	2.98
Android Developer	2.93
Graphic Artist	2.91
Mobile Application Developer	2.91
Web Design Specialist	2.88
Business Intelligence Specialist	2.85
SAP Expert	2.82
Digital Marketing and Social Media Specialist	2.81
Mobile Software Specialist	2.77
Digital Marketing Specialist	2.66
Social Media Expert	2.63
Hardware Engineer	2.50
Hardware Support Specialist	2.40
Hardware Specialist	2.00

5. Conclusions

Job advertisement analysis studies have become widespread in recent years to determine the necessary qualifications for various positions. It can be said that the researches made for Turkish are limited while a large resource pool is encountered for the English language. Kariyer.Net is the biggest company for the job ads in Turkey and 99% of the ads are Turkish. Therefore, there is a necessity to develop novel Natural Language Processing (NLP) models in Turkish for analysis this big database. In this study, the job ads of Kariyer.Net have been analyzed, and by using a hybrid clustering algorithm, the hidden associations in this dataset as the big data have been discovered. Firstly, all ads in the form of HTML codes have been transformed into regular sentences by the means of extracting HTML codes to inner texts. Then, these inner texts containing the core ads have been converted into the sub ads by traditional methods. After these NLP steps, hybrid clustering algorithms have been used and the same ads expressed with the different sentences could be managed to be detected. For the analysis, 57 positions about Information Technology sectors with 6,897 ad texts have been focused on. As a result, it can be claimed that the clusters obtained contain useful outcomes and the model proposed can be used to discover common and unique ads for each position. The results obtained in Section IV shows that Cluster 1 contains the technical skills, Cluster 2 contains the military competences and Cluster 3 contains the graduation department ads. Lastly, in Cluster 4, it was observed that the postings containing the experience year requests for the positions were grouped. As a result, it can be claimed that the clusters obtained contain useful outcomes and the model proposed can be used to discover common and unique ads for each position.

Author's Note

Abstract version of this paper was presented at 9th International Conference on Advanced Technologies (ICAT'20), 10-12 August 2020, Istanbul, Turkey with the title of "Discovering The Same Job Ads Expressed with The Different Sentences by Using Hybrid Clustering Algorithms".

Acknowledgment

This study is supported by Kariyer.Net R&D Center. In addition, we thank the research team for the dataset.

References

[1] R. Loth, D. Battistelli, F. R. Chaumartin, H. De Mazancourt, J. L. Minel, and A. Vinckx, "Linguistic information extraction for job ads (SIRE project)," In 9th *RIAO: Adaptivity, Personalization and Fusion of Heterogeneous Information*, 2010, pp. 222-224.

[2] J. L. F. D. M. Pombo, "Landing on the right job: a machine learning approach to match candidates with jobs applying semantic embeddings," *Doctoral dissertation*, 2019.

[3] J. Grüger, and G. J. Schneider, "Automated analysis of job requirements for computer scientists in online job advertisements," in *15th International Conference on Web Information Systems and Technologies*, 2019, pp 226-233.

[4] M. A. Kennan, P. Willard, P., D. C. Kecmanovic, and C. S. Wilson, "25. IS early career job advertisements: A content analysis," in *11th Pacific-Asia Conference on Information Systems, New Zealand*, 2007, pp. 340-353.

[5] Y. Choi, and E. Rasmussen, "What qualifications and skills are important for digital librarian positions in academic libraries? A job advertisement analysis," *The Journal of Academic Librarianship*, vol. 35, no. 5, pp. 457–467, 2009.

[6] M. Pember, "Content analysis of recordkeeping job advertisements in Western Australia: Knowledge and skills required by employers," *Australian Academic & Research Libraries*, vol. 34, no 3, pp. 194-210, 2003.

[7] D. C. Angelides. "From the present to the future of civil engineering education in Europe: A strategic approach," in *Proceedings of the International Meeting in Civil Engineering Education, Ciudad Real, Spain*, 2003, pp. 1-21.

[8] C. Kwon Lee, and H. Han, "Analysis of skills requirement for entry-level programmer/analysts in fortune 500 corporations," *Journal of Information Systems Education*, vol. 19, no. 1, pp. 17-27, 2008.

[9] K. Yongbeom, H. Jeffrey, and S. Mel, "An update on the is/it skills gap," *Journal of Information Systems Education*, vol. 17, no. 4, pp. 395–402, 2008.

[10] T. Chamorro-Premuzic, D. Winsborough, R. A. Sherman, and R. Hogan, "New talent signals: Shiny new objects or a brave new world," *Industrial and Organizational Psychology*, vol. 9, no. 3, pp. 621–640, 2016.

[11] A. Amado, P. Cortez, P. Rita, and S. Moro, "Research trends on Big data in marketing: A text mining and topic modeling based literature analysis," *European Research on Management and Business Economics*, vol. 24, no. 1, pp. 1–7, 2018.

[12] M. Mezzanzanica, "Italian web job vacancies for marketing-related professions," *Symphonya. Emerging Issues in Management*, vol. 3, no. 1, pp. 110–124, 2017.

[13] L. Guo, C. J. Vargo, Z. Pan, W. Ding, and P. Ishwar, "Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling," *Journalism & Mass Communication Quarterly*, vol. 93, no. 2, pp. 332–359, 2016.

[14] Y. Kino, H. Kuroki, T. Machida, N. Furuya, and K. Takano, "Text analysis for job matching quality improvement," *Procedia Computer Science*, vol. 112, no. 1, pp. 1523–1530, 2017.

[15] I. Karakatsanis, W. AlKhader, F. MacCrory, A. Alibasic, M. A. Omar, and Z. Aung, "Data mining approach to monitoring the requirements of the job market: A case study," *Information Systems*, vol. 65, no. 4, pp. 1–6, 2016.

[16] O. Müller, T. Schmiedel, E. Gorbacheva, and J. vom Brocke, "Towards a typology of business process management professionals: Identifying patterns of competences through latent semantic analysis," *Enterprise Information Systems*, vol. 10, no. 1, pp. 50–80, 2016.

[17] F. Amato, R. Boselli, M. Cesarini, F. Mercorio, M. Mezzanzanica, and V. Moscato, "Challenge: Processing web texts for classifying job offers," in *Semantic Computing (ICSC), 2015 IEEE International Conference on Semantic Computing*, 2015, pp. 460–463.

[18] R. Boselli, M. Cesarini, F. Mercorio, and M. Mezzanzanica, "Classifying online job advertisements through machine learning," *Future Generation Computer Systems*, vol. 86, no. 9, pp. 319–328, 2018.

[19] X. Xu, X. Wang, Y. Li, and M. Haghigh, "Business intelligence in online customer textual reviews: Understanding consumer perceptions and influential factors," *International Journal of information management*, vol. 37, no. 6, pp. 673–683, 2017.

[20] W. He, S. Zha, and L. Li, "Social media competitive analysis and text mining: A case study in the pizza industry," *International Journal of Information Management*, vol. 33, no. 3, pp. 464–472, 2013.

[21] B. Jeong, J. Yoon, and J. M. Lee, "Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis," *International Journal of Information Management*, vol. 48, no. 1, pp. 280-290, 2019.

[22] M. Pejić Bach, Ž. Krstić, S. Seljan, and L. Turulja, "Text mining for big data analysis in financial sector: A literature review," *Sustainability*, vol. 11, no. 5, pp. 1-27, 2019.

[23] H. C. Chang, C. Y. Wang, and S. Hawamdeh, "Emerging trends in data analytics and knowledge management job market: Extending KSA framework," *Journal of Knowledge Management*, vol. 23, no. 4, pp. 664-686, 2018.

[24] I. Kregel, N. Ogonek, and B. Matthies, "Competency profiles for lean professionals-an international perspective," *International Journal of Productivity and Performance Management*, vol. 68, no. 2, pp. 423–446, 2019.

[25] A. de Mauro, M. Greco, M. Grimaldi, and P. Ritala, "Human resources for big data professions: A systematic classification of job roles and required skill sets," *Information Processing and Management*, vol. 54, no. 9, pp. 807–817, 2017.

[26] A. Gardiner, C. Aasheim, P. Rutner, and S. Williams, "Skill requirements in big data: A content analysis of job advertisements,"

Journal of Computer Information Systems, vol. 58, no. 4, pp. 374–384, 2018.

[27] P. A. Todd, J. D. McKeen, and R. B. Gallupe, "The evolution of IS job skills: A content analysis of IS job advertisements from 1970 to 1990," *MIS quarterly*, vol. 19, no.1, pp. 1-27, 1995.

[28] A. Amado, P. Cortez, P. Rita, and S. Moro, "Research trends on Big data in marketing: A text mining and topic modeling based literature analysis," *European Research on Management and Business Economics*, vol. 24, no. 1, pp. 1-7, 2018.

[29] A. AlAlwan, N. P. Rana, Y. K. Dwivedi, and R. Algharabat, "Social media in marketing: A review and analysis of the existing literature," *Telematics and Informatics*, vol. 34, no. 7, pp. 1177-1190, 2017.

[30] Y. K. Dwivedi, K. K. Kapoor, and H. Chen, "Social media marketing and advertising," *The Marketing Review*, vol. 15, no. 3, pp. 289-309, 2015.

[31] L. Guo, C. J. Vargo, Z. Pan, W. Ding, and P. Ishwar, "Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling," *Journalism & Mass Communication Quarterly*, vol. 93, no. 2, pp. 332-359, 2016.

[32] K. K. Kapoor, K. Tamilmani, N. P. Rana, P. Patil, Y. K. Dwivedi, and S. Nerur, "Advances in social media research: Past, present and future," *Information Systems Frontiers*, vol. 20, no. 3, pp. 531-558, 2018.

[33] W. L. Shiau, Y. K. Dwivedi, and H. S. Yang, "Co-citation and cluster analyses of extant literature on social networks," *International Journal of Information Management*, vol. 37, no. 5, pp. 390-399, 2017.

[34] W. L. Shiau, Y. K. Dwivedi, and H.H. Lai, "Examining the core knowledge on Facebook," *International Journal of Information Management*, vol. 43, no. 1, pp. 52-63, 2018.

[35] I. Rahhal, I. Makdoun, G. Mezzour, I. Khaouja, K., Carley, and I. Kassou, "Analyzing cybersecurity job market needs in Morocco by mining job ads. In *2019 IEEE Global Engineering Education Conference (EDUCON)*, 2019, pp. 535-543.