# Homogeneous Biodata Extraction using Cluster Analysis & Natural Language Processing Techniques

**Rabeea Ijaz[1], Sheikh Muhammad Aamir[2], Amnah Riaz Chohan[3]**

Department of Computer Science, Government College University, Faisalabad, Pakistan

***Abstract:*** *It has always been a big challenge to manage large volumes of text documents in organizations. During recruitment process organizations are challenged with massive amount of structured and unstructured datasets in the form of Biodata or resumes. Therefore, this is time consuming and more effort demanding process to analyze resumes data separately. This study presents an exploratory automated filtration of a search result clustering methodology by the assistance of Natural Language Processing (NLP) techniques to group the specific resumes on the basis of certain keywords for multiple departments. It is based on contents of text documents and label these resulting groups significantly. In this technique effectiveness and efficiency has been judged. The proposed mining methodology is constructed on the principles of text mining using RapidMiner7 open source tool. RapidMiner7 utilizes data extraction strategies to extract specific information from text for decision making. The study proposes the most appropriate cluster analysis algorithm for document clustering and evaluates these methods to decide ideal number of clusters to have effective categorization and tag the categories meaningfully by comparing two algorithms that are K-Means and DBSCAN. This study evaluates the feasibility of proposed methodology specifically for the textual data sets by performing various experiments and results which states that the proposed method has higher precision.*

***Keywords:*** *Cluster Analysis; Data mining; DBSCAN clustering; K-Means; Natural Language Processing; Text mining.*

## 1. Introduction

This research reports the discoveries and explorations in text mining. Text mining (also acknowledged as smart & intelligent script examination, documented, word-based or text-based data mining, amorphous datasets organization, and knowledge discovery in text script) remains a subcategory of Information Retrieval (IR), which is a universal subcategory of Artificial Intelligence subdivision of computer science. Text mining can be defined as "non-trivial mining of formerly unidentified, implicit, and hypothetically valuable knowledge from huge no. of "text dataset", that differs from data mining in that it is organized, unformatted and frequently looking for information. Text mining is an interdisciplinary domain that involves not only machine learning, standard data mining and statistics, but also Natural Language Processing (NLP) & Computational Linguistics (CL)[14].

Data Mining have 6 general categories:

A. **Anomaly Detection (Outlier/change/deviation detection) –** Recognizing the infrequent & uncommon dataset archives which may be stimulating or data faults or variations that needs some additional exploration.

B. **Association Rule Learning (Dependency modelling)** – Retrieves associations & relations between different variables e.g. a departmental store can collect data about a client's buying behaviors. By adopting association rule learning, store identifies products that are frequently purchased and uses these information for advertising purposes. Sometimes, this is called “Market Basket Analysis”.

C. **Clustering –** (the descriptive approach) is the method of exploring collections and patterns in data which are approximately "similar" in some means, deprived of consuming identified patterns in datasets.

D. **Classification –** (the predictive approach) is a generalization of identified structures for applying them to new dataset e.g., an e-mail application can aim towards categorizing e-mail in place of "spam" Or "legitimate".

E. **Regression –** attempts towards discovery of a function that models the minimum error-contained data.

F. **Summarization –** provides more concise illustration of dataset, as well as report generation and visualization.

Text Mining (TM) is Knowledge Discovery process for extracting effective high quality hidden information sets for efficient retrieval of specific required information, decision making or predictive analysis from textual big data sources[16]. Text mining has been considered to initiate as of data mining; though, a couple of methods have originated from different domains i.e. information visualization and data science. Text mining endeavors to tackle data over-burden issue by utilizing strategies from data mining, Information Extraction (IE), Knowledge Management (KM), Machine Learning, Natural Language Processing (NLP) and Information Retrieval (IR). TM comprises pre-processing of massive text document gatherings (text classification, feature or term mining etc.), storing intermediate representations, analyzing these middle expressions (association rules, trend analysis, clustering and distribution analysis) and results visualization [2].

### 1.1. Cluster Analysis

Clustering is a method of distributing an arrangement of information into similar, related sub-classes and dissimilar groups, called clusters that useful for data experts and database administrators to understand the common congregation or organization in an information set. Developed either as per a stand-alone tool to access understanding about information diffusion or as a pre-processing scheme for different calculations. A group of data items can be considered as a single set. Although, performing a cluster analysis, first divide the dataset into groups according to the resemblance of the data, and then allocate the tags to the clusters. The key benefit and objective of cluster analysis is that it is flexible to variations and assists to pick valuable features that differentiate heterogeneous datasets [21].

Clustering is a method of distributing an arrangement of information into similar, related sub-classes and dissimilar groups, called clusters that useful for data experts and database administrators to understand the common congregation or organization in an information set. Developed either as per a stand-alone tool to access understanding about information diffusion or as a pre-processing scheme for different calculations. A group of data items can be considered as a single set. Although, performing a cluster analysis, first divide the dataset into groups according to the resemblance of the data, and then allocate the tags to the clusters. The key benefit and objective of cluster analysis is that it is flexible to variations and assists to pick valuable features that differentiate heterogeneous datasets [1].

1. Connectivity based clustering that is hierarchical clustering technique.
2. Partitioning clustering that is the centroid based cluster analysis.
3. Density-Based clustering that is stated as region of greater density.
4. Grid based cluster analysis which is grounded on the size or dimension of grid not on dataset.
5. Optics and Expectation Maximization algorithm.

These clustering technique's algorithms are compared based on different factors that are: size or dimensions of the data-set, total amount of clusters, different types of data & kind of software that is used [15]. Clustering technique can be useful in knowledge areas i.e. Machine pattern recognition, machine learning, image analysis & feature extraction from image, bioinformatics and information retrieval [12].

### 1.2. Natural Language Processing for Text

Most of the text used in online forums (such as chat rooms, Twitter, discussion forums, and social media) is inaccurate because senders & receivers both can be very used to with the shorthand structures of all words and reduce the time & struggle of the sender & receiver. Maximum of the text is formed and warehoused, so that individuals could recognize it, and it is not at all the times relaxed for the computer to

process the textual data. With the growth of noisy text-based data produced from several social media, the cleaning of this text had turn out to be essential, and because existing NLP techniques are used generally due to several reasons (such as sparseness, vocabulary words) and the inconsistent syntactic structure in the text [23].

Some of the texts cleansing methods are as follows:

- Removing Stop Words (removing most of the common words such as "a", "the", "and", etc.).
- Tokenization (Converting text and words into tokens by ignoring the spaces & punctuations)
- Stemming (techniques of joining words that has the identical linguistic stem or root).
- POS Tagging (also recognized as word class uncertainty or ambiguity, is the procedure of allocating text parallel to some specific POS, such as an adjective, noun, pronoun, verb, adverb, preposition, or other vocabulary class-marker.)
- Syntactical Parsing (procedure of executing a parsing of a string of phrases, words, or sentences as stated by firm rules of syntax and grammar)

### 1.3. Research Questions & Problem Statement

All organizations' most of the critical job now a day is recruitment and appointment of new people. For recruitment, large number of resumes, any organization receives as job applicant or career submission, are much larger than the no. of persons allocated to analyze and process these. There is a need of text mining model that sorts & filter keywords, like study domain, specific work experience, department they are applying for, internships & interests, awards & achievements etc. Grounding upon those keywords several classes could be identifies and CVs could be categorized, ultimately leading to selection of better individuals.

In general, the research paper will answer following research questions.

It should be clear that research specialist should make numerous decisions which might affect the construction for a clustering solution. These decisions could be gathered in the following comprehensive groups:

A. ***Data Conversion Problems***
- Which is the degree of similarity / non-similarity from the textual data sets?
- Do I need to standardize my text data? How should the critics of the metrics be treated?
- How should data interdependencies be handled?

B. ***Solution Problems***
- What number of clusters would be acquired?
- Which clustering algorithm had better to be used for text based clustering?
- Would all cases be involved in the cluster analysis or what subset would be disregarded?

C. ***Rationality Problems***
- Is the cluster analysis resultant dissimilar from what was predicted?
- Is the cluster resultant stable or reliable through the sample?
- Are clusters interrelated to variables other than derived variables? Are clusters beneficial?

D. ***Variable selection Issue***

What are most appropriate attributes or variables for producing a clustering solution for the textual data set?

## 2. Related Work

Bhanuse, S. S., et al. (2016) gives the idea of using metadata to generate sub-information in text mining. Can be classified and categorized according to various clustering and classification algorithms. Metadata information has been used by the text mining techniques to mine text-based datasets. To plan clustering, conventional segmentation utilizes the probability models for creating the efficient clusters. The presented experimental outcomes in this paper were based on cluster number, execution time & precision. A series of classical segmentation and probability models has been used to design an extended clustering method.

This also provides a scope to provide security for cluster-side information and to explore the implementation of filtering methods that work in the training set model [9].

Al-Anazi, S., et al. (2016) indicates that several clustering models have been established for the Saudi International University graduation project document. Three sets of similarity measurements were tested, compared & evaluated. They discover that better performance could be gained by using the K-mean & cosine similarity from K-neutral joint. The text documents in their data set have different lengths and are divided into various themes. Subsequently the cosine similarity calculation is independent of length of the document, this is possible to improved handling of data set. On the basis of usage of cluster analysis assessment calculations, the quality of clustering is different. They also discovered that, as per the rate of 'k' rises, quality of cluster analysis has been enhanced. Lastly, they conclude that project concepts are typically divided among following classes: eHealth applications, Religious and Arabic applications, location based applications, signal, voice & image recognition, gaming and E-learning applications [5].

Thomas, A. M., & Resmipriya, M. G. (2016) presented a text categorization method for semi - supervised clustering, and analyzes the accuracy of the similarity metric obtained in the classification algorithm. The elementary hypothesis is that every class of documents originates from several modules and could be recognized through cluster analysis. To illustrate clustering procedure, unmarked documents had been utilized to iteratively adjust the centroid of the clustering candidate & used a tagged document to capture the outline of the cluster. This is a semi-supervised clustering method based on search. It provides better classification results. Use different similarity measures in the classification process and get better accuracy values from SMTP, because one can use future enhanced size reduction techniques. And the dimension of the term document matrix can be reduced. As a result, better execution time can be achieved, and a large number of documents can be easily handled [22].

Abin, A. A. (2016) presented a combined framework of constraint cluster analysis and active assortment constraints. The proposed method adopts the alternating way to constrain the clustering and give the idea of fuzzy clustering through the multi - core learning constraint information. The constraints have been chosen on the basis of the reality that it is healthier to actively query the constraints if the clustering algorithm is clustering. Taking into account this hypothesis, the proposed method of active constraint assortment had implanted in the technique of querying constraints according to the present clustering condition [3].

Pohl, D. et al. (2016) proposed a framework based on crisis-related data recognition sub-events. Identify the use of online clustering & automatic indexing algorithms. They examined 3 online indexing approaches (skewness, incremental TF-TDF & learning overlook). Experimentations represents that for more metric terms, the "learning and forgetting" approach is preferable than the skew & incremental TF-IDF technique. Can identify small events related to major events, such as power outages, floods, evacuation, and so on. The demonstration activities held in 2013 in September, which similarly show the role of proposed framework in real-time backup reply [17].

Song, X., et al. (2016) proposed to extract the context information into text mining, they introduced a new concept as a semantic continuous sequence pattern. A novel semantic model mining problem is proposed and a solution is provided. By incorporating the location information in the definition of semantic patterns, their algorithm can obtain semantic meaningful text units as a model. It also provides a novel perspective to generate a candidate pattern with a suffix array and a longest public prefix, which is in full compliance with their theoretical framework and their problem. Although more information is included, the algorithm is still running in linear time. In addition, experience had been shown that the semantic pattern is more compact and representative than the continuous sequence pattern of the prior art [20].

Sheng, X. et al. (2016) presented in their paper, a new & innovative algorithm for text mining grounded on ANN (Artificial Neural Network). It is a significant means for data mining. One of the foundations of Neural Network model and differentiation is text mining. The ANN model of data mining mostly includes forward multilayered neural network, RBF-based neural network, self-organized feature map based NN, learning quantitative NN, NN adaptive resonance and cyclic NN. According to Sheng et al. some of the signals of the areas could range the concept of synaptic strength & cerebral cortex. By adding this concept to reduce the distance of the neurons, a new neural network dimension algorithm was designed grounded on fundamental change of the NN signal transmission mechanism. They apply this theory to text mining applications. Experiments showed that the results of proposed algorithm were acceptable [19].

Kapugama, K. D. C. G. et al. (2016) shows that Wikipedia searching queries & results could be sorted & marked by utilizing machine learning and text mining methods. K-means clustering is superior to other clustering techniques, for text documents categorization. In the article from Wikipedia, investigational results show that by choosing the first paragraph of the Wikipedia article text, better accuracy could be received, rather than the K-means and clustering of the complete text. In addition, a common frequent pattern could be seen in graph of the typical sum of the TF-IDF scores from the TF-IDF matrix. Some of the clusters in the generated cluster contain similar tags, consequently these clusters would increase readability cluster. In k means algorithm for clustering, the concluding last production or output be contingent on early discovery of centroid. Therefore, they tried to improve their results by detecting the initial centroid k-means by using the results from "text document clustering on the basis of centroid" of the early centroid. Their method used 400 documents. They trust that by growing the no. of considered documents, they'll be able to enhance the resultant clusters quality. A large no. of document collections takes lengthy period to collect the outcomes, so the additional expansion for this research was in a distributed environment in a distributed manner in a small time to experiment. In this study, they compared 2 clustering algorithms, and they had planned to compare more clustering algorithms to discover the utmost appropriate technique or algorithm. It was a continuous study, the further expansion of the research was to utilize further methods, e.g. silhouette analysis for comparing the various algorithms. To find the no. of clusters, they used their own technique, and they wanted to make more improvements to get improved outcomes [13].

Alksher, M. A. et al. (2016) outlines the continuous expansion of text data leads to mining methods and methods for analyzing the secreted data since the unstructured textual data. The concept of "mining" is a young and exposed field of study and research, few of which are not yet completed. Therefore, some use of text mining technology research, and the application of automatic concept mining method to extract ideas. Thus, in general, improvements in text model creation seem to help to extract ideas by combining mining ideas [6].

Brindha, S. et al. (2016) research presents that classification technique of data mining is too much supportive in the text mining field, increasing number of digitalized data or information and mining valuable knowledge from this huge amount of data. In machine learning and data mining literature, classification issue is utmost significant problem. From the textual data context, one can consider failure in a separate class to set the convenience of the perceived attribute, while ignoring the frequency of the word. Nearly all data classification methods, such as Bayesian methods, Decision trees, Nearest Neighbors, Neural Networks, and SVM classifiers, all apply to features of text-based data. Classification development could greatly improve the quality of text results and the precise data in least access time [10].

Rezaeian, M. et al. (2017) has introduced (I) lifecycle analysis, (ii) text mining, and (iii) identification of knowledge gaps through automatic clustering. They showed that proposed method could deliver beneficial knowledge for assessment of innovative investigation & development actions in explicit

research areas. Though the previous personal method had applied in the situation of prospective research, the chief role of this research is the combination of its framework. Through the development of text mining and the expansion of technical level & cluster analysis, the related key terms & their connections in the systematic literature have been traced in the growth track of the exploration research area. By merging this (quantitative) automated analysis (i.e., the occurrence or nonoccurrence of common) and (qualitative) skilled input, the information gap in this area is found in effective means, as well as the potential for future research and development. The elements that link these three steps together are scientific research methods. First, it can increase insight into the impact of procedural tendencies on technological lifecycle growth and systematic consideration. In addition, it effectively transfers the analyst's attention to areas that have not been previously explored [18].

Dong, G. et al. (2017) proposed the subject of the explosion in the case of Twitter is the problem of mining outbreaks. In order to express the topological topology of a huge amount of Twitter user's topology, a prominent theme user graph design model is projected. Also, hierarchical cluster analysis is implemented on cluster burst themes, revealing burst patterns from a macro viewpoint. Alternatively, from the perspective of the prominent theme of the information flow model, frequent subgraph mining is used. The investigational outcomes demonstrates that there are numerous exciting burst patterns that could show bursts of frequent and frequent information flows from different subject groups [11].

Alghamdi, Turki et al. (2015) anticipated a method to improve and enhance the performance in finding the appropriate and related document from heterogeneous records or databases. The method comprises on two features that is; text mining based on retrieval & the query extension based on users feedback. The investigational outcome demonstrates that their anticipated algorithm carries important developments in the outcome. In this method extended query based on user feedback the precision is enhanced by 0.4% on average from 36.4% to 36.8% and the recall is enhanced by 0.6% from 41.7% to 42.3%. Their anticipated technique could produce considerable enhancements over prevailing methods [24].

## 3. Materials & Methods

The dataset utilized for the research under study was gathered from several people of various institutes and of different fields. Manual information is collected, pre-processed and rationalized it as stated in to the obligatory of this research. Information gathered by google forms and some data is gathered from data banks available on internet. There are 2 types of datasets have been used in this study i.e. data is both in structured and in unstructured form.

- Google form:

https://docs.google.com/forms/d/e/1FAIpQLScDGhoFUOMLsIYbxcHJtoGCohBZIR4WGzWhxrv4nWlke-KyQ/viewform?usp=sf_link

- Data Sources:

http://barbizonmodeling.com/resumes/
http://www.circuitgallery.com/resumes/
http://www.vmwp.com/resumes/

More than 200 responses have been recorded from these different sources. Data was in structured, semi structured and unstructured format. Preprocessing applied to data to make is in single useable format for analysis.

### 1.1. Data Cleaning & Pre-Processing (ETL):

A. ***Extraction***

Large datasets collection from multiple sources and extraction of required information from the huge amount of raw facts and figures is known as “Extraction” stage. The dataset is available in structured, semi-structured and unstructured form which is very heterogeneous, noisy and complex. I digitize these raw datasets from hard form to soft form by making MS Excel files.

B. ***Transformation & Cleaning***

To create a useful database, these extracted attributes are stored and integrated in Microsoft Excel. To clean, removed empty, duplicate and unnecessary data from datasets. It also removed complexities and noisy data. Various machine learning algorithms are used to generate unstructured dataset into a structured

dataset. Such as, mapping table has used to change textual form data into numeric form, because many clustering algorithms accept only numeric and binary data.

C. ***Loading***

After cleaning and transforming the data, it is ready to load in the tool for analysis. The transformed and cleaned Excel spreadsheet is loaded into the RapidMiner 7 and this excel sheet is transformed into multiple homogenous format of documents by using "data to documents" operator available in RapidMiner discussed in next sections. And make a corpus and stored in the repository for analysis.

D. ***NLP Techniques:***

After loading data into RapidMiner, NLP techniques are applied to data that are necessary for the Natural Language textual datasets. To process textual datasets, NLP techniques are necessary for the machine(Computer) to learn these words. These techniques include tokenization, stop words removal, stemming, POS tagging, transformation of cases etc.

After data pre-processing, it is ready for analysis. Data loaded into the RapidMiner in .xlsx form and then converted to multiple documents by using "Data to Document" operator and stored as corpus of multiple documents. These generated documents are then analyzed by applying NLP techniques and using different clustering models. Term Frequency and Term occurrences are calculated for attribute "Position Interested" and according to this attribute similarity of different documents is calculated by using "Data to Similarity" operator in RapidMiner will discussed later in this chapter. And then K-Means and DBSCAN clustering models are applied. Both algorithms produces the same result but for K-Means it takes 25 clusters. In contrast, DBSCAN revenues only 12 clusters i.e. equals to 12 categories in dataset.

Finally, in this study the clustering model has been used for categorization of resumes data according to the study domain or position interested for specific department. There are 12 different categories or departments for positions interested for applicants and each cluster have single category resumes. i.e. homogenous biodata according to position interested by the applicants. And the resultant clusters are saved into the local repository in folder form. A unique folder is generated for each cluster and each folder has single category resumes. When new applicants are added to the catalogue, they automatically stored into the related folder or cluster according to position interested.

Large number of text mining tools for the text analysis and applying different data mining techniques on textual datasets are available. Such as R, GATE, Weka, RapidMiner, Orange, KNIME etc. RapidMiner 7.5. and Microsoft Excel 2016 has been used in this study for text mining analysis and document clustering.

Two approaches are used to text mining.

- Semantic Parsing
- Bag of Words

Semantic parsing is grounded on syntax of word. In this approach, word type and order has been considered to be cared. This methodology generates a great deal of structures & features for research & study. E.g. solo word could be marked (tagged) as fragment of the sentence, then a noun, likewise a proper noun, or named entity. Thus, that solo word had three features linked through it. This outcome made semantic parsing features powerful and rich. To do the tagging, semantic parsing trails a tree format to endure to break up the text. Words are broken down with different and unique attributes annotated with them.

In divergence, the bag of words methodology does not overhaul about word order or type. At this point, words are only attributes of some text document. In the following example, a sentence is parsed. "Stephen Johns missed a tough shot."

In bag of words technique, any term is handled as a solo token in the sentence, irrespective of the type or order of words. In this study, simple bag of words method had been utilized.

E. ***Indexing:***

The TF-IDF technique is commonly used intended for the calculation of term frequency of all or some specific or selected keywords. The keywords are then organized in descending

order. The value of keyword is total-TF and computed by adding all the term frequencies of that keyword. Fig. 1 shows some of the TF-IDF values of the keywords.

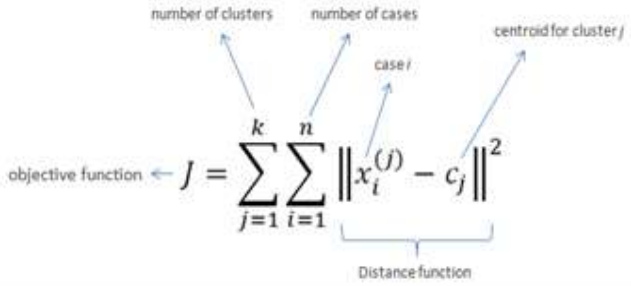
| Word | Attribute Name | Total Occurences | Document Occurences |
|---|---|---|---|
| Accounts | Accounts | 12 | 12 |
| Administration | Administration | 15 | 15 |
| Architecture | Architecture | 10 | 10 |
| DatabaseManagement | DatabaseManagement | 8 | 8 |
| Development | Development | 22 | 22 |
| Education | Education | 33 | 33 |
| Engineering | Engineering | 8 | 8 |
| GraphicDesigning | GraphicDesigning | 13 | 13 |
| HumanResourceManagement | HumanResourceManagement | 20 | 20 |
| InformationTechnology | InformationTechnology | 36 | 36 |
| Networks | Networks | 15 | 15 |
| QualityAssurance | QualityAssurance | 8 | 8 |

Fig. 1 : Word List on the Basis of Selected Attribute "Position Interested"

F. ***K-Means***

Step1: Randomly choose K cluster centers.
Step2: Assign each object to closest center.
Step3: Recalculate the centers.
Step4: Repeat step1 and step2 until stop condition is reached.

Fig. 2 : K-Means Algorithm




G. ***DBSCAN***

**for** each $o \in D$ **do**
  **if** $o$ is not yet classified **then**
    **if** $o$ is a core-object **then**
      collect all objects density-reachable from $o$
      and assign them to a new cluster.
    **else**
      assign $o$ to NOISE

Fig. 3 : DBSCAN Algorithm

Objects and operators used in cluster model are as follows:

i. Read Data (Format) or Retrieve from repository
ii. Then convert data into suitable format for analysis by using various operators available.
iii. Creating a documents corpus in the repository and store for future use in other processes.
iv. Then create word vector and count term frequencies in document occurrences.
v. Process documents by using some NLP techniques like tokenization etc.
vi. Then calculate data to similarity: similarity between multiple documents.
vii. The generate a cluster model by using any clustering algorithm.
viii. Setting parameters accurately can leads to a successful cluster model.
ix. By executing the generated model, we can get results in form of example set, WordList of TF-IDF and cluster model with advanced graphs and folder view of clusters.

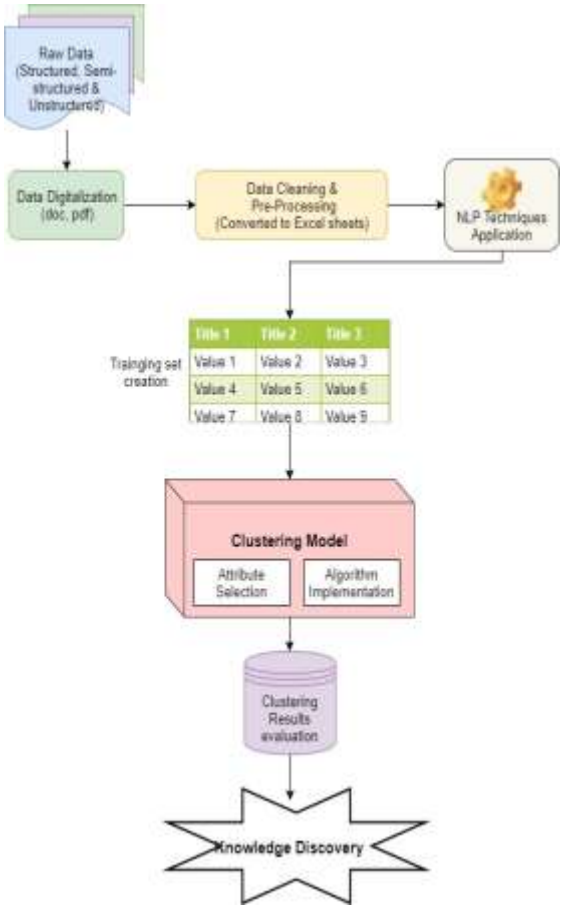



Fig. 4 : Proposed Framework Flow for Cluster Analysis Model

Proposed Framework as shown in Fig. 4, the cluster model or framework can categorize the any type of textual datasets based on some variables or attributes and enhance the power of data mining by reducing the manual work. In clustering model, different algorithms are available, but in this study, 2 algorithms K-Means and DBSCAN have been used, and compared their results based on the efficiency and amount of clusters produced by the algorithm & accuracy of cluster generation. Steps involved for the generation of cluster model in RapidMiner are:

a) Preparation & organization of the dataset
b) Selection of Attribute on the base of which clusters will be generated.
c) Selection of clustering technique and applying appropriate algorithm
d) Generation of clusters according to selected attributes
e) Cluster Analysis
    a. K-Means
    b. DBSCAN
f) Obtained output
g) Knowledge Discovery

## 2. Results and Discussion

### *A. Experiment 1: K-Means Cluster Model*

First, the dataset has been analyzed and clustered by using K-means clustering algorithm. There are 12 categories in selected attribute named “position_interested”, therefore, the value of ‘k’ given as 12 initially as parameter. It created 12 clusters in cluster model. But due to very less Euclidian distance it merges some categories in single cluster zero. Then value of ‘k’ gradually increased and evaluated one by one. At last value of ‘k’ given 25 as parameter, 25 clusters have been generated with each cluster having different and single type of category for the selected attribute / variable to reduce the Euclidian distance between clusters. K-Means cluster model results are shown in Table 1.

### *B. Experiment 2: DBSCAN Cluster Model*

In second experiment, algorithm has been changed for the same dataset. Now, the dataset has been analyzed and clustered by using DBSCAN clustering algorithm. There are 12 categories in selected attribute named “position_interested”. DBSCAN clustering operator in RapidMiner have some parameters. One is epsilon, whose default value is 1.0. The parameter defines the epsilon (real) parameter of the DBSCSN clustering algorithm. Epsilon defines

The region or neighborhood size which is another is min. points (integer) parameter defines the minimal no. of points generating some cluster. Next is ‘add cluster attribute’ (Boolean) If this parameter is established to true, a novel attribute with cluster role is produced in the resulting ExampleSet, else this operator doesn’t improve the cluster attribute. In the

concluding situation, have to avail the Apply Model operator to create cluster attribute. DBSCAN cluster model results are shown in Table 2.

## 3. Conclusion

The motivation for this study is organizational facilitation and reducing manual work. Each organization's most of the critical job is recruiting the new people. For recruitment, large number of resumes an organization receives for a career application are much higher than the no. of persons allocated to examine these. In this era of Computer Science and Information Technology, there is a need of text mining model that sorts & filter keywords, like study domain, department, internship experiences, explicit job or work experience, awards, interests, achievements etc. On the basis those keywords, numerous categories could be defined & biodata or resumes could be clustered, ultimately leading to selection of better individuals. The utmost significant disturbance of the specified problem is that there's a hole between theoretical work and realistic one.

The chief purpose behind this research has become to make organizations' HR department fastest, within the time, and the opportunity of handing over complete tasks on the cease of the recruitment process. Application of the selected procedure might in addition made the implementation in their recruitment process in a powerful manner.

In this study, the dataset used, has been obtained in form of resumes and biodata, from various applicants from online portal i.e. google form and some data is collected through online data repositories.

The dataset comprised of applicants' demographic data, their academic qualification data, their experiences and position in which they are interested in. The number of samples used to construct a cluster model is 200 and 12 categories are obtained in sense of "Position Interested" by Applicants and only this attribute is used for analysis purpose.

Only the attribute that was required intended for data mining progression was selected. The variables used for categorizing the data may be multiple but in this study only one variable is used for analysis i.e. position interested by the candidate. Two algorithms are used and compared for this model, they are K-Means and DBSCAN clustering algorithms. In this study, DBSCAN algorithm has been found fast for the specific structured dataset and gives exact results with exact number of clusters according to categories. K-Means clustering model is useful in case of unstructured dataset where there is lot of noisy and unnecessary data.

## 4. Future Work

Future work in this perspective might be a new ranking algorithm can be proposed for each cluster to rank the resume according to credibility, experience and qualifications. In the algorithm, higher the qualification, higher will be the rank and it will increase efficiency and automation more. Also these biodata datasets can be categorized or classified according to their educational details and a prediction based classification model will be introduced to suggest and predict the most appropriate job or position or department for the each individual candidate.

**Table 1 :** K-Means Result-Cluster Model Description and Details

| Cluster # | No. of Items | Category Name / Position Interested by Applicants |
|---|---|---|
| Cluster No. 0 | 36 | Information Technology |
| Cluster No. 1 | 20 | Human Resource Management |
| Cluster No. 2 | 0 | Empty |
| Cluster No. 3 | 0 | Empty |
| Cluster No. 4 | 33 | Education |
| Cluster No. 5 | 13 | Graphic Designing |
| Cluster No. 6 | 0 | Empty |
| Cluster No. 7 | 12 | Accounts |
| Cluster No. 8 | 0 | Empty |
| Cluster No. 9 | 0 | Empty |
| Cluster No. 10 | 0 | Empty |
| Cluster No. 11 | 8 | Quality Assurance |
| Cluster No. 12 | 10 | Architecture |
| Cluster No. 13 | 0 | Empty |
| Cluster No. 14 | 8 | Engineering |
| Cluster No. 15 | 0 | Empty |
| Cluster No. 16 | 0 | Empty |
| Cluster No. 17 | 15 | Administration |
| Cluster No. 18 | 15 | Networks |
| Cluster No. 19 | 0 | Empty |
| Cluster No. 20 | 22 | Development |
| Cluster No. 21 | 0 | Empty |
| Cluster No. 22 | 0 | Empty |
| Cluster No. 23 | 0 | Empty |
| Cluster No. 24 | 8 | Database Management |
| Total No. of Items | 200 | 12 Categories |

**Table 2.** DBSCAN Result - Cluster Model Description and Details

| Cluster # | No. of Items | Category Name / Position Interested by Applicants |
|---|---|---|
| C-0 | 0 | Empty (Cluster zero is for noisy data in DBSCAN clustering) |
| C -1 | 36 | Information Technology |
| C -2 | 20 | Human Resource Management |
| C -3 | 22 | Development |
| C -4 | 8 | Engineering |
| C -5 | 12 | Accounts |
| C -6 | 33 | Education |
| C -7 | 15 | Networks |
| C -8 | 13 | Graphic Designing |
| C -9 | 8 | Quality Assurance |
| C -10 | 15 | Administration |
| C -11 | 10 | Architecture |
| C -12 | 8 | Database Management |
| Total No. of Items | 200 | 12 categories |

**References**


[1] Abbas, Osama Abu. "Comparisons Between Data Clustering Algorithms." *International Arab Journal of Information Technology (IAJIT)* 2008;5(3).

[2] Abdullah, Z., and A. R. Hamdan. "Hierarchical Clustering Algorithms in Data Mining." *International Journal of Computer, Electrical, Automation, Control and Information Engineering* 2015; 9(10)

[3] Abin, Ahmad Ali. "Clustering with side information: Further efforts to improve efficiency." *Pattern Recognition Letters* 2016; 84 :252-258.

[4] Aghabozorgi, Saeed, Ali Seyed Shirkhorshidi, and Teh Ying Wah. "Time-series clustering–A decade review." *Information Systems* 2015; 53: 16-38.

[5] Al-Anazi, Sumayia, Hind AlMahmoud, and Isra Al-Turaiki. "Finding Similar Documents Using Different Clustering Techniques." *Procedia Computer Science* 2016; 82:28-34.

[6] Alksher, Mostafa A., Azreen Azman, Razali Yaakob, Rabiah Abdul Kadir, Abdulmajid Mohamed, and Eissa M. Alshari. "A review of methods for mining idea from text." In *Information Retrieval and Knowledge Management (CAMP), 2016 Third International Conference IEEE on*, 2016; 88-93.

[7] Alzghool, Muath, and Diana Inkpen. "A novel class-based data fusion technique for information retrieval." *Journal of Emerging Technologies in Web Intelligence* 2010;2(3): 160-166.

[8] Basu, Tanmay, and C. A. Murthy. "A similarity assessment technique for effective grouping of documents." *Information Sciences* 2015; 311: 149-162.

[9] Bhanuse, Shraddha S., Shailesh D. Kamble, and Sandeep M. Kakde. "Text Mining Using Metadata for Generation of Side Information." *Procedia Computer Science* 2016; 78: 807-814.

[10] Brindha, S., K. Prabha, and S. Sukumaran. "A survey on classification techniques for text mining." In *Advanced Computing and Communication Systems (ICACCS), 2016 3rd International IEEE Conference on*, 2016; 1: 1-5.

[11] Dong, Guozhong, Wu Yang, Feida Zhu, and Wei Wang. "Discovering burst patterns of burst topic in twitter." *Computers & Electrical Engineering* 2017; 58: 551-559.

[12] Jain, Anoop Kumar, and Satyam Maheswari. "Survey of recent clustering techniques in data mining." *Int J Comput Sci Manag Res* 2012; 3: 72-78.

[13] Kapugama, K. D. C. G., S. A. S. Lorensuhewa, and M. A. L. Kalyani.

"Enhancing Wikipedia search results using Text Mining." In *Advances in ICT for Emerging Regions (ICTer), 2016 Sixteenth International IEEE Conference on*, 2016; 168-175.

[14] Kaur, Manpreet, and Usvir Kaur. "Comparison between k-means and hierarchical algorithm using query redirection." *International Journal of Advanced Research in Computer Science and Software Engineering* 2013; 3(7).

[15] Kaushik, Manju, and Mrs Bhawana Mathur. "Comparative Study of K-Means and Hierarchical Clustering Techniques." *International Journal of Software & Hardware Research in Engineering (IJSHRE)* 2014; 2(6).

[16] Mahgoub, Hany, and D. Rösner. "Mining association rules from unstructured documents." In *Proc. 3rd Int. Conf. on Knowledge Mining, ICKM, Prague, Czech Republic*, 2006; 167-172.

[17] Pohl, Daniela, Abdelhamid Bouchachia, and Hermann Hellwagner. "Online indexing and clustering of social media data for emergency management." *Neurocomputing* 2016; 172: 168-179.

[18] Rezaeian, Mina, H. Montazeri, and R. C. G. M. Loonen. "Science foresight using life-cycle analysis, text mining and clustering: A case study on natural ventilation." *Technological Forecasting and Social Change* 2017; 118: 270-280.

[19] Sheng, Xiaobao, Xin Wu, and Yimin Luo. "A novel text mining algorithm based on deep neural network." In *Inventive Computation Technologies (ICICT), International IEEE Conference on*, 2016; 2: 1-6.

[20] Song, Xiaoli, XiaoTong Wang, and Xiaohua Hu. "Semantic pattern mining for text mining." In *Big Data (Big Data), 2016 IEEE International IEEE Conference on*, 2016; 150-155.

[21] Tan, Ah-Hwee. "Text mining: The state of the art and the challenges." In *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, 1999; 8: 65-70.

[22] Thomas, Anisha Mariam, and M. G. Resmipriya. "An efficient text classification scheme using clustering." *Procedia Technology* 2016; 24: 1220-1225.

[23] Yuan, Man, and Yong Shi. "Text clustering based on a divide and merge strategy." *Procedia Computer Science* 2015; 55: 825-832.

[24] Alghamdi, Turki, Mohammad Husain, and Ahmad Alkhodre. "A Novel Approach for Filtering Appropriate Document from Most Relevant Query Terms." *MAGNT Research Report (ISSN. 1444-8939)* 2015; Vol.3 (4): 359-367