International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 28, No. 5, October, 2023, pp. 1265-1273
Journal homepage: http://iieta.org/journals/isi

Leveraging Text Mining for Analyzing Students' Preferences in Computer Science and Language Courses

Alex Alfredo Huaman Llanos[1*], Lenin Quiñones Huatangari[2], Jeimis Royler Yalta Meza[3], Alexander Huaman Monteza[4]

[1] Informatic and Language Center, National University of Jaen, Jaen 06801, Peru
[2] Data Science Research Institute, National University of Jaen, Jaen 06801, Peru
[3] Direction of Production Center of Goods and Services, National University of Jaen, Jaen 06801, Peru
[4] Social Science Academic Department, National University of Jaen, Jaen 06801, Peru

Corresponding Author Email: alex.huaman@unj.edu.pe

https://doi.org/10.18280/isi.280515

Received: 11 May 2023
Revised: 21 August 2023
Accepted: 23 August 2023
Available online: 31 October 2023

Keywords:
decision making, educational policy, R language, students' preferences, text mining WEKA, word cloud

ABSTRACT

In an increasingly competitive, globalized world, educational institutions must strategically offer courses that align with the skill-acquisition needs of their target audience. As such, the application of text mining techniques to extract valuable insights and patterns from structured data across various knowledge domains becomes paramount. This study employed text mining to scrutinize students' preferences for course offerings at the Computer and Language Center of the National University of Jaen. The analysis was based on data collected from a Google Forms survey of 315 students. The employed methodology facilitated the unearthing of patterns, trends, and semantic relationships within a large corpus of students' opinions. Frequency distributions and word clouds were generated using R programming language. Furthermore, the WEKA software and Python were utilized for cluster analysis, enabling the detection of groupings and trends within the data. Although other methods such as sentiment analysis and statistical methodologies exist, text mining was deemed most suitable for identifying patterns and relationships within students' opinions. The study revealed that students predominantly favored advanced Excel, AutoCAD, ArcGIS, Nutrition, and Revit courses, which appeared to correlate with their professional aspirations and prevailing course trends. Therefore, the application of text mining tools to analyze structured institutional data can significantly contribute to informed decision-making processes.

1. INTRODUCTION

The advent of technological innovation and scientific advancements has catalyzed various strategies aimed at addressing the exigencies of public management. A critical enabler in this context is the electronic government, a platform that amplifies service delivery and provides reliable information to citizens. Consequently, national governments are grappling with the challenge of implementing electronic government to optimize public services.

This study is located within this broader context and seeks to analyze students' opinions on Computer Science and Language courses, and identify patterns related to their professional education. The primary research question guiding this inquiry is: How can text mining techniques be utilized to effectively scrutinize students' opinions and course preferences at the Computer Science and Languages Centre, and how can significant patterns and relationships be discerned to augment educational quality and institutional decision-making?

Text Mining (TM), a process employed to discover and extract previously unknown knowledge from a large corpus of text data [1], is crucial to this study. Similarly, Computational Linguistics, an interdisciplinary field straddling linguistics, computer science, and machine learning [2], and Natural Language Processing (NLP), a subcategory of artificial intelligence (AI) that automates the understanding, representation, and manipulation of text and human speech [3, 4], are integral to the study.

TM, which involves extracting non-trivial patterns from massive text documents [5], informed the text preprocessing phase methodology of this study. Various TM approaches, many employing computational linguistics and Python libraries and R language functions [6], were used as exemplars for creating word clouds in R. As a growing field, TM aims to enhance human understanding by revealing hidden text patterns [7], primarily through stages such as normalization, cleaning, and elimination of stop words. The flexibility of TM methodologies allow for diverse utilizations, such as text understanding and knowledge extraction [8], which facilitated the development of the data collection phase.

Applications of TM are widespread and diverse, including cancer research analysis [9], opinion analysis of digital banking applications [10], sentimental analysis [11, 12] vaccine hesitancy studies [13], and analysis of information management systems research topics [14]. Other applications include Twitter content sharing studies [15], text classification [16], database vulnerability identification [17], Business Intelligence article reviews [18], Peruvian professional CV analysis [19], and construction sector research [20].

In the educational field, TM has been utilized for systematic review using classification and clustering methods [21]. The K-means clustering algorithm has been employed to group student outcomes [22] and to identify the cyberbullying [23], machine learning has been applied in the RFID supply chain [24], and signal processing, data mining, and pattern recognition have been used [25]. Also, a mixed methodology based on integrating data structure and students' academic performance has been used to determine factors influencing student performance [26], and methods used in 21st-century education have been explored [27]. Other studies have proposed a new recruitment system [28], analyzed online questions and chat messages [29], and examined China's educational policy network [30].

This study is critical because the Computer Science and Languages Center, as a public institution, needs to discern trends or guidelines for managing, planning, and offering courses tailored to the specific professional development and skill enhancement needs of their target audience. Therefore, the findings will aid in decision-making, ensuring that course offerings align with students' opinions and/or suggestions. The application of TM techniques will contribute to identifying patterns, trends, and needs that can shape educational decision-making and curriculum planning across various professional disciplines. The detailed methodological approach proposed will enable a comprehensive analysis and understanding of textual data. Furthermore, other institutions within the same field can adopt it as a reference to enhance service delivery, improve educational quality in both disciplines, and meet end-user needs.

2. PROPOSED METHODOLOGY

For the development of the research, data creation of frequencies and word cloud, as well as the application of unsupervised algorithms (K-means) of data mining were considered; additionally, clustering and statistics were used to evaluate their behavior.

2.1 Data collection

The data collected show the opinion and suggestions of university students enrolled in courses offered by the Computer and Language Centre. Therefore, a sample of 315 students from the Office and English courses belonging to the VII to X cycle of the professional careers of Civil Engineering, Mechanical and Electrical Engineering, Food Industry Engineering, Forestry and Environmental Engineering and Medical Technology of the National University of Jaen was used as a sample. The chosen sample, represents approximately one fifth of the total population of students at the National University of Jaen, who were sent a link in Google Forms and the survey was applied during the period from November 2022 to May 2023. The information collected was stored in a sheet in an Excel format and the study variables (career and recommended course) were checked, see Figure 1.

2.2 Frequency and word cloud creation

To analyze the information the information and form the word cloud, RStudio software was used (Figure 2). It consists of the following steps:

(1) Installation of libraries to create the word cloud: SnowballC, RColorBrewer, RCurl, NLP, tm, WordCloud, Corpus and ggplot2. SnowballC was used to reduce words to a common root in order to facilitate vocabulary comparison, RColorBrewer to provide color to the generated graphics, RCurl provided functions to obtain URLs and process results returned by the web server, NLP for working with basic classes and methods for Natural Language Processing, Tm for text mining applications, WordCloud for graphing the world cloud, Corpus for searching term occurrences and calculating term frequency and ggplot2 for creating graphs declaratively based on the graph grammar.

(2) Use previously installed libraries.

(3) Load the data using the read.csv function. The data consisted of two columns and 315 instances, delimited by commas.

(4) Data cleaning process: The following items were removed: numerical data, words without meaning on their own, scores, blanks and words of little relevance to the study. The data cleaning was carried out using R, and the following functions were used: tolower to convert text from uppercase to lowercase, removeNumbers to remove numerical data, stopwords to remove words without relevance such as articles, pronouns, prepositions and adverbs within the corpus. Then, punctuation marks were removed using removePunctuation function and finally stripWhitspace to remove blank spaces. It should be noted that there were no missing data values due to the survey in Google Forms was configured in such a way that all the fields' questions were obligatory.

(5) Elaboration of the graphical representation of word frequency, using the corpus worked on and the data processed in the previous steps.

(6) Finally, the word cloud was plotted taking into account the word and word frequency.

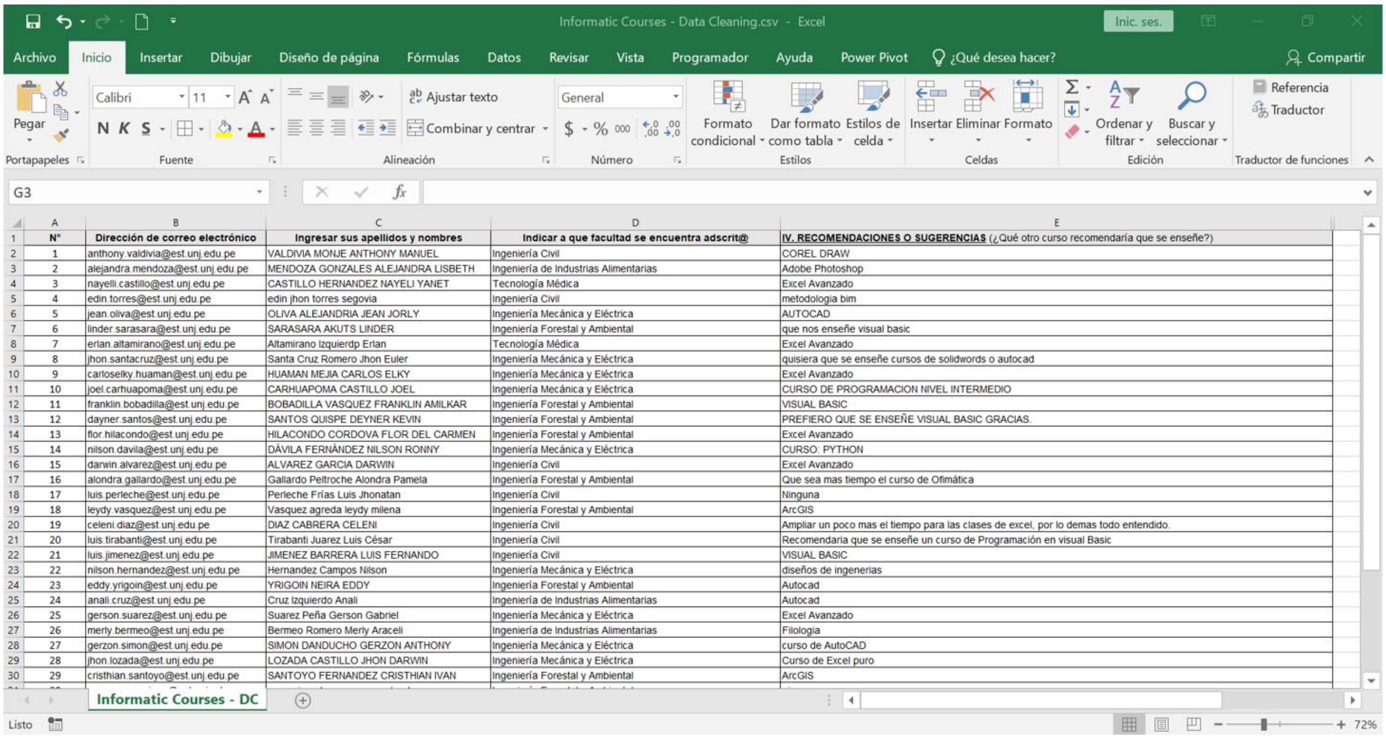
2.3 Data mining use

From the databse, the data was cleaned using text ming and machine learning techniques in order to filter it for further analysis. The Waikato Environment for Knowledge Analysis (WEKA) software was used for this phase, considering the following steps:

(1) Loading and exploration of data in .csv format. See Figure 3.

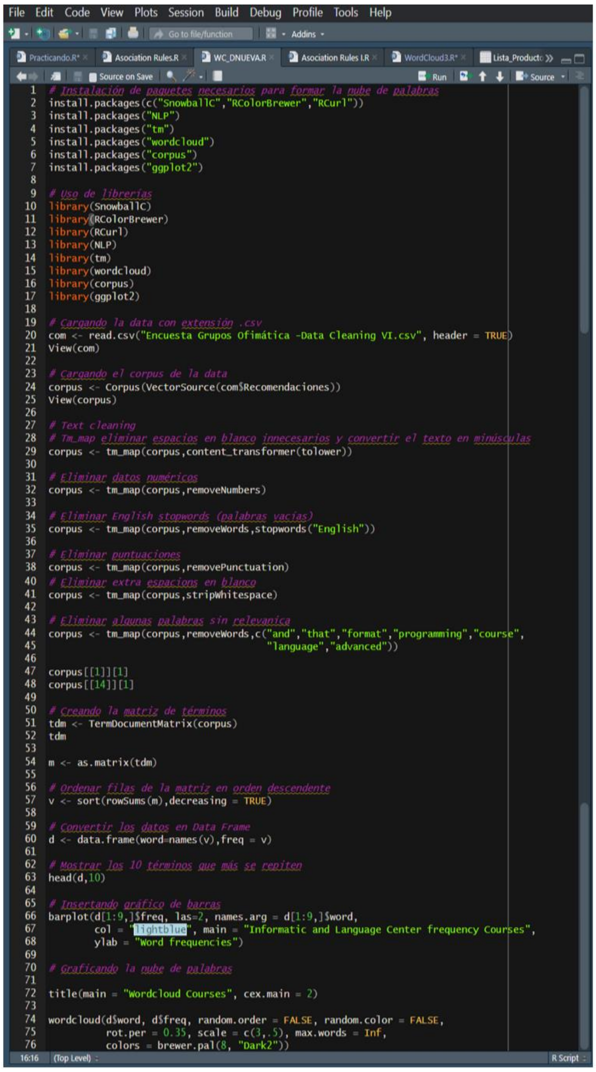
(2) Data filtering: Consisted of using the NominalToString attribute of unsupervised learning to transform the data values, changing the data type (nominal to string), finally the StringToWordVector class to filter the strings into N-grams using the Word Tokenizer class, which helped to provide strings, four was considered as the minimum frequency of words. In addition, articles, prepositions and pronouns were removed by choosing the Snowball Stemmer and Rainbow options for the stop words (empty words). Finally, the words were converted to lower case, see Figure 4.

(3) Cluster generation: A clustering model was chosen for data segmentation, using the K-means algorithm. This was due to, this algorithm facilitates the identification of opinion patterns, preferences and trends in large amount of textual data in order to break down the information into meaningful groups, contributing to the deep understanding and data structure. For example, it was possible to verify which courses were recommended by students taking into account the professional career, such as Civil Engineering, Food Industry Engineering, Forestry and Environmental Engineering, Mechanical and Electrical Engineering and Medical Technology (Figure 5).



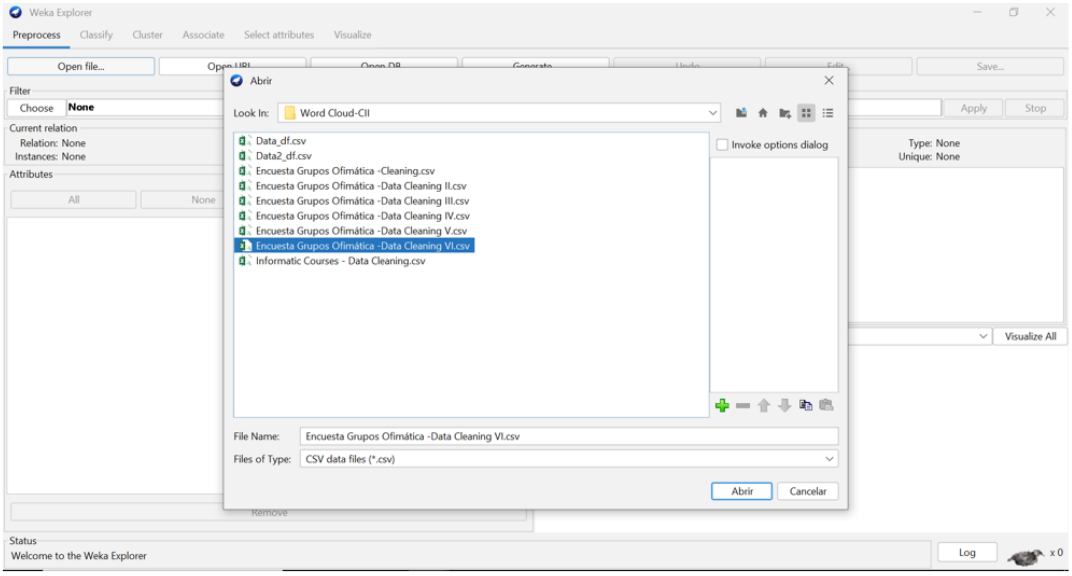
	A	B	C	D	E
1	N°	Dirección de correo electrónico	Ingresar sus apellidos y nombres	Indicar a que facultad se encuentra adscrit@	IV. RECOMENDACIONES O SUGERENCIAS (¿Qué otro curso recomendaría que se enseñe?)
2	1	anthony.valdivia@est.unj.edu.pe	VALDIVIA MONJE ANTHONY MANUEL	Ingeniería Civil	COREL DRAW
3	2	alejandra.mendoza@est.unj.edu.pe	MENDOZA GONZALES ALEJANDRA LISBETH	Ingeniería de Industrias Alimentarias	Adobe Photoshop
4	3	nayelli.castillo@est.unj.edu.pe	CASTILLO HERNANDEZ NAYELI YANET	Tecnología Médica	Excel Avanzado
5	4	edin.torres@est.unj.edu.pe	edin jhon torres segovia	Ingeniería Civil	metodologia bim
6	5	jean.oliva@est.unj.edu.pe	OLIVA ALEJANDRIA JEAN JORLY	Ingeniería Mecánica y Eléctrica	AUTOCAD
7	6	linder.sarasara@est.unj.edu.pe	SARASARA AKUTS LINDER	Ingeniería Forestal y Ambiental	que nos enseñe visual basic
8	7	erlan.altamirano@est.unj.edu.pe	Altamirano Izquierdp Erlan	Tecnología Médica	Excel Avanzado
9	8	jhon.santacruz@est.unj.edu.pe	Santa Cruz Romero Jhon Euler	Ingeniería Mecánica y Eléctrica	quisiera que se enseñe cursos de solidwords o autocad
10	9	carloselky.huaman@est.unj.edu.pe	HUAMAN MEJIA CARLOS ELKY	Ingeniería Mecánica y Eléctrica	Excel Avanzado
11	10	joel.carhuapoma@est.unj.edu.pe	CARHUAPOMA CASTILLO JOEL	Ingeniería Mecánica y Eléctrica	CURSO DE PROGRAMACION NIVEL INTERMEDIO
12	11	franklin.bobadilla@est.unj.edu.pe	BOBADILLA VASQUEZ FRANKLIN AMILKAR	Ingeniería Forestal y Ambiental	VISUAL BASIC
13	12	dayner.santos@est.unj.edu.pe	SANTOS QUISPE DEYNER KEVIN	Ingeniería Forestal y Ambiental	PREFIERO QUE SE ENSEÑE VISUAL BASIC GRACIAS.
14	13	flor.hilacondo@est.unj.edu.pe	HILACONDO CORDOVA FLOR DEL CARMEN	Ingeniería Forestal y Ambiental	Excel Avanzado
15	14	nilson.davila@est.unj.edu.pe	DÁVILA FERNÁNDEZ NILSON RONNY	Ingeniería Mecánica y Eléctrica	CURSO: PYTHON
16	15	darwin.alvarez@est.unj.edu.pe	ALVAREZ GARCIA DARWIN	Ingeniería Civil	Excel Avanzado
17	16	alondra.gallardo@est.unj.edu.pe	Gallardo Peltroche Alondra Pamela	Ingeniería Forestal y Ambiental	Que sea mas tiempo el curso de Ofimática
18	17	luis.perleche@est.unj.edu.pe	Perleche Frías Luis Jhonatan	Ingeniería Civil	Ninguna
19	18	leydy.vasquez@est.unj.edu.pe	Vasquez agreda leydy milena	Ingeniería Forestal y Ambiental	ArcGIS
20	19	celeni.diaz@est.unj.edu.pe	DIAZ CABRERA CELENI	Ingeniería Civil	Ampliar un poco mas el tiempo para las clases de excel, por lo demas todo entendido.
21	20	luis.tirabanti@est.unj.edu.pe	Tirabanti Juarez Luis César	Ingeniería Civil	Recomendaria que se enseñe un curso de Programación en visual Basic
22	21	luis.jimenez@est.unj.edu.pe	JIMENEZ BARRERA LUIS FERNANDO	Ingeniería Civil	VISUAL BASIC
23	22	nilson.hernandez@est.unj.edu.pe	Hernandez Campos Nilson	Ingeniería Mecánica y Eléctrica	diseños de ingenerias
24	23	eddy.yrigoin@est.unj.edu.pe	YRIGOIN NEIRA EDDY	Ingeniería Forestal y Ambiental	Autocad
25	24	anali.cruz@est.unj.edu.pe	Cruz Izquierdo Anali	Ingeniería de Industrias Alimentarias	Autocad
26	25	gerson.suarez@est.unj.edu.pe	Suarez Peña Gerson Gabriel	Ingeniería Mecánica y Eléctrica	Excel Avanzado
27	26	merly.bermeo@est.unj.edu.pe	Bermeo Romero Merly Araceli	Ingeniería de Industrias Alimentarias	Filologia
28	27	gerzon.simon@est.unj.edu.pe	SIMON DANDUCHO GERZON ANTHONY	Ingeniería Mecánica y Eléctrica	curso de AutoCAD
29	28	jhon.lozada@est.unj.edu.pe	LOZADA CASTILLO JHON DARWIN	Ingeniería Mecánica y Eléctrica	Curso de Excel puro
30	29	cristhian.santoyo@est.unj.edu.pe	SANTOYO FERNANDEZ CRISTHIAN IVAN	Ingeniería Forestal y Ambiental	ArcGIS

Figure 1. Database collected from students' opinions



```r
# Instalación de paquetes necesarios para formar la nube de palabras
install.packages(c("SnowballC","RColorBrewer","RCurl"))
install.packages("NLP")
install.packages("tm")
install.packages("wordcloud")
install.packages("corpus")
install.packages("ggplot2")

# Uso de librerías
library(SnowballC)
library(RColorBrewer)
library(RCurl)
library(NLP)
library(tm)
library(wordcloud)
library(corpus)
library(ggplot2)

# Cargando la data con extensión .csv
com <- read.csv("Encuesta Grupos Ofimática -Data Cleaning VI.csv", header = TRUE)
View(com)

# Cargando el corpus de la data
corpus <- Corpus(VectorSource(com$Recomendaciones))
View(corpus)

# Text cleaning
# Tm_map eliminar espacios en blanco innecesarios y convertir el texto en minúsculas
corpus <- tm_map(corpus,content_transformer(tolower))

# Eliminar datos numéricos
corpus <- tm_map(corpus,removeNumbers)

# Eliminar English stopwords (palabras vacías)
corpus <- tm_map(corpus,removeWords,stopwords("English"))

# Eliminar puntuaciones
corpus <- tm_map(corpus,removePunctuation)
# Eliminar extra espacions en blanco
corpus <- tm_map(corpus,stripWhitespace)

# Eliminar algunas palabras sin relevanica
corpus <- tm_map(corpus,removeWords,c("and","that","format","programming","course",
                                      "language","advanced"))

corpus[[1]][1]
corpus[[14]][1]

# Creando la matriz de términos
tdm <- TermDocumentMatrix(corpus)
tdm

m <- as.matrix(tdm)

# Ordenar filas de la matriz en orden descendente
v <- sort(rowSums(m),decreasing = TRUE)

# Convertir los datos en Data Frame
d <- data.frame(word=names(v),freq = v)

# Mostrar los 10 términos que más se repiten
head(d,10)

# Insertando gráfico de barras
barplot(d[1:9,]$freq, las=2, names.arg = d[1:9,]$word,
        col = "lightblue", main = "Informatic and Language Center frequency Courses",
        ylab = "Word frequencies")

# Graficando la nube de palabras

title(main = "Wordcloud Courses", cex.main = 2)

wordcloud(d$word, d$freq, random.order = FALSE, random.color = FALSE,
          rot.per = 0.35, scale = c(3,.5), max.words = Inf,
          colors = brewer.pal(8, "Dark2"))
```

Figure 2. RStudio code for data processing



Abrir

Look In: Word Cloud-CII

- Data_df.csv
- Data2_df.csv
- Encuesta Grupos Ofimática -Cleaning.csv
- Encuesta Grupos Ofimática -Data Cleaning II.csv
- Encuesta Grupos Ofimática -Data Cleaning III.csv
- Encuesta Grupos Ofimática -Data Cleaning IV.csv
- Encuesta Grupos Ofimática -Data Cleaning V.csv
- Encuesta Grupos Ofimática -Data Cleaning VI.csv
- Informatic Courses - Data Cleaning.csv

□ Invoke options dialog

File Name: Encuesta Grupos Ofimática -Data Cleaning VI.csv

Files of Type: CSV data files (*.csv)

Abrir | Cancelar

Filter: Choose None

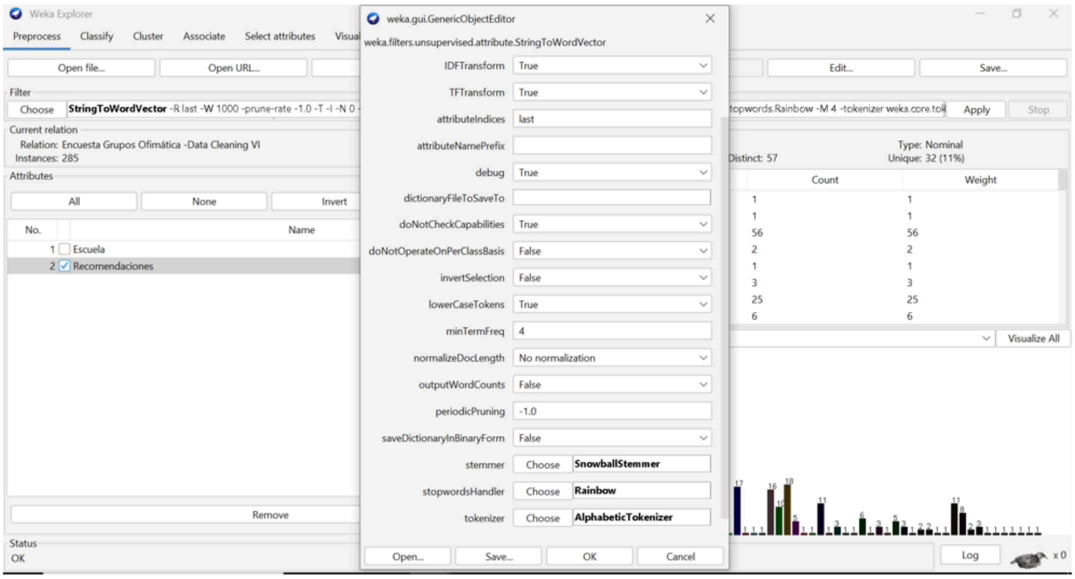
Current relation
Relation: None
Instances: None

Type: None
Unique: None

Attributes

Status
Welcome to the Weka Explorer

Figure 3. Loading and exploration data using WEKA



Filter: Choose **StringToWordVector** -R last -W 1000 -prune-rate -1.0 -T -I -N 0 - ... topwords.Rainbow -M 4 -tokenizer weka.core.tok

Current relation
Relation: Encuesta Grupos Ofimática -Data Cleaning VI
Instances: 285

Attributes

No.		Name
1	□	Escuela
2	☑	Recomendaciones

Distinct: 57
Type: Nominal
Unique: 32 (11%)

Count	Weight
1	1
1	1
56	56
2	2
1	1
3	3
25	25
6	6

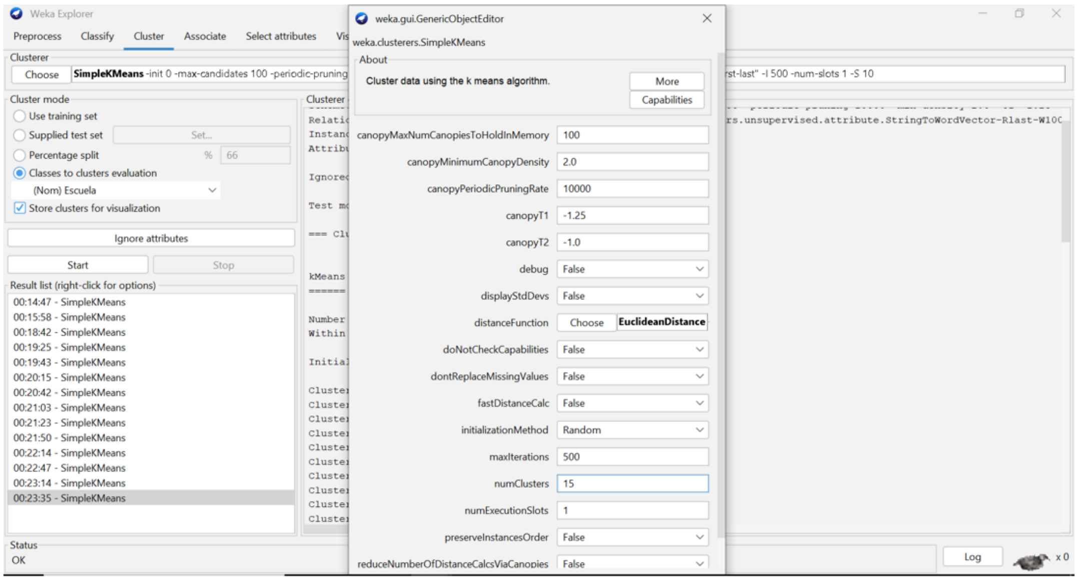
weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.StringToWordVector

IDFTransform	True
TFTransform	True
attributeIndices	last
attributeNamePrefix	
debug	True
dictionaryFileToSaveTo	
doNotCheckCapabilities	True
doNotOperateOnPerClassBasis	False
invertSelection	False
lowerCaseTokens	True
minTermFreq	4
normalizeDocLength	No normalization
outputWordCounts	False
periodicPruning	-1.0
saveDictionaryInBinaryForm	False
stemmer	Choose **SnowballStemmer**
stopwordsHandler	Choose **Rainbow**
tokenizer	Choose **AlphabeticTokenizer**

Status
OK

Figure 4. Data cleaning using WEKA



Clusterer: Choose **SimpleKMeans** -init 0 -max-candidates 100 -periodic-pruning ... rst-last" -I 500 -num-slots 1 -S 10

Cluster mode
○ Use training set
○ Supplied test set
○ Percentage split % 66
● Classes to clusters evaluation
(Nom) Escuela
☑ Store clusters for visualization

Result list (right-click for options)
00:14:47 - SimpleKMeans
00:15:58 - SimpleKMeans
00:18:42 - SimpleKMeans
00:19:25 - SimpleKMeans
00:19:43 - SimpleKMeans
00:20:15 - SimpleKMeans
00:20:42 - SimpleKMeans
00:21:03 - SimpleKMeans
00:21:23 - SimpleKMeans
00:21:50 - SimpleKMeans
00:22:14 - SimpleKMeans
00:22:47 - SimpleKMeans
00:23:14 - SimpleKMeans
00:23:35 - SimpleKMeans

```
rs.unsupervised.attribute.StringToWordVector-Rlast-W10C
```

weka.gui.GenericObjectEditor

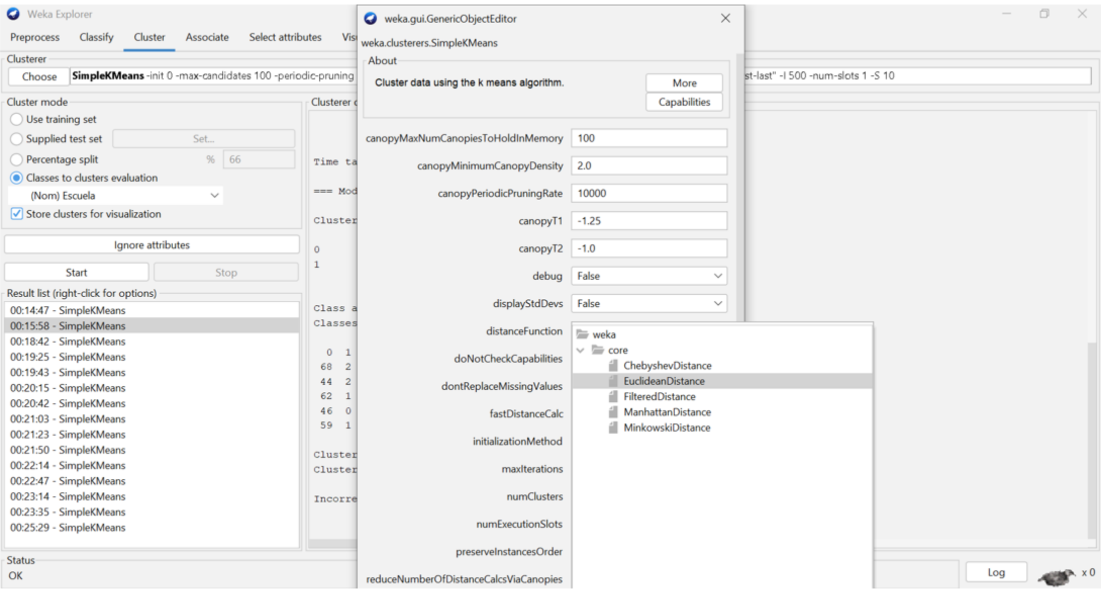
weka.clusterers.SimpleKMeans

About
Cluster data using the k means algorithm.

canopyMaxNumCanopiesToHoldInMemory	100
canopyMinimumCanopyDensity	2.0
canopyPeriodicPruningRate	10000
canopyT1	-1.25
canopyT2	-1.0
debug	False
displayStdDevs	False
distanceFunction	Choose **EuclideanDistance**
doNotCheckCapabilities	False
dontReplaceMissingValues	False
fastDistanceCalc	False
initializationMethod	Random
maxIterations	500
numClusters	15
numExecutionSlots	1
preserveInstancesOrder	False
reduceNumberOfDistanceCalcsViaCanopies	False

Status
OK

Figure 5. K-means generating using WEKA



Clusterer: SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning … st-last" -I 500 -num-slots 1 -S 10

Cluster mode
○ Use training set
○ Supplied test set
○ Percentage split % 66
● Classes to clusters evaluation
(Nom) Escuela
☑ Store clusters for visualization

Result list (right-click for options)
00:14:47 - SimpleKMeans
00:15:58 - SimpleKMeans
00:18:42 - SimpleKMeans
00:19:25 - SimpleKMeans
00:19:43 - SimpleKMeans
00:20:15 - SimpleKMeans
00:20:42 - SimpleKMeans
00:21:03 - SimpleKMeans
00:21:23 - SimpleKMeans
00:21:50 - SimpleKMeans
00:22:14 - SimpleKMeans
00:22:47 - SimpleKMeans
00:23:14 - SimpleKMeans
00:23:35 - SimpleKMeans
00:25:29 - SimpleKMeans

Status
OK

Clusterer o

```
Time ta

=== Mod

Cluster

0
1


Class a
Classes

  0  1
 68  2
 44  2
 62  1
 46  0
 59  1

Cluster
Cluster

Incorre
```

weka.clusterers.SimpleKMeans
About
Cluster data using the k means algorithm.

canopyMaxNumCanopiesToHoldInMemory	100
canopyMinimumCanopyDensity	2.0
canopyPeriodicPruningRate	10000
canopyT1	-1.25
canopyT2	-1.0
debug	False
displayStdDevs	False
distanceFunction	
doNotCheckCapabilities	
dontReplaceMissingValues	
fastDistanceCalc	
initializationMethod	
maxIterations	
numClusters	
numExecutionSlots	
preserveInstancesOrder	
reduceNumberOfDistanceCalcsViaCanopies	

weka
core
ChebyshevDistance
EuclideanDistance
FilteredDistance
ManhattanDistance
MinkowskiDistance

Figure 6. Iteration of K-means using WEKA

2.4 Clustering

Clustering is the process of grouping a set of elements into clusters of similar objects [31]. The research used clustering algorithms built into WEKA and K-means was chosen as the clustering method. As well as, the use of Elbow Method to determine the optimal cluster number using python [32]. In addition, the Euclidean distance function, iterations number: 10 and the class for cluster evaluation: School were considered as parameters (Figure 6).

2.4.1 Statistics for assessing cluster performance

Sum Squared Error (SSE): It is one of the statistical methods used to measure the total difference between the actual value and the achieved value [33], which is defined as [34], see Eq. (1):

$$SSE\ (X, \Pi) = \sum_{i}^{k} \sum_{x_i \epsilon C_i} \parallel x_j - m_i \parallel^2 \qquad (1)$$

where:

$\parallel . \parallel$: Euclidean distance

m_i: Centroid of cluster C_i

In this study, the choice of using the K-means algorithm and the elbow method is based on their effectiveness in addressing the task segmentation textual data and determining the optimal number of clusters in the analysis of opinions and preferences related to Computer and Language courses. The K-means algorithm is a widely recognized technique in text mining for clustering data and revealing patterns in large text sets. On the other hand, the elbow method is employed to identify the appropriate number of clusters, allowing for a more meaningful results interpretation. This choice aligns with the unsupervised nature of opinion and preference data, and seeks to capture intrinsic relationships of participant's responses. The combination of these two methods seeks to ensure a robust analysis and an important representation of the different perspectives and trends in the data, thus contributing to a deeper understanding of student's course and career preferences.

3. RESULTS

A frequency bar chart (Figure 7) and word cloud (Figure 8) were generated. Figure 7 lists the most repeated courses after the data analysis. As a result, students show their preferences in the following courses: advanced Excel, AutoCAD, ArcGIS, Nutrition, Revit, Portuguese, Python language, French and Arduino, which shows the trend and importance of these courses for students of different professional careers.

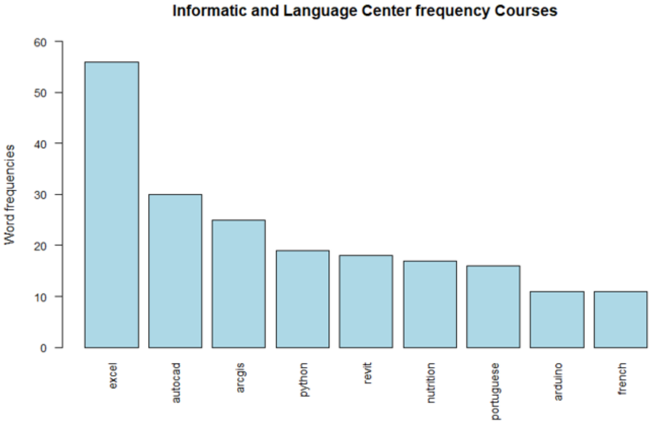


Figure 7. Frequency of words on course preferences

This study uses data collected from surveys that were administered to students at the Computer and Language Centre. The key features of this data include the variety of terms used to express opinions and preferences. Thus, by processing texts and analyzing word frequency, a word cloud was created as a visual representation that highlights the most frequent and relevant words in the dataset. The word cloud generated in the study condenses the student's responses into an image in which the size of each word reflects its frequency in the text. In this way, the word cloud provides a quick and visual insight into trends in students' course preferences. Each term in the word cloud is a clue to understanding the predominant preferences and areas of interest, giving an overview of emerging patterns in the data and highlighting the most relevant aspects of the analysis.

In the word cloud (Figure 8), it is observed that the advanced Excel is the one with the outstanding size, then, there are courses like AutoCAD, Revit, Python, Nutrition, ArcGIS that have a lower preference. The results also show courses such as Quality Control, languages such as French, Mandarin Chinese and Portuguese, which have an intermediate preference. Finally, courses such as Visual Basic, Graphic Design, SolidWorks have a low preference.

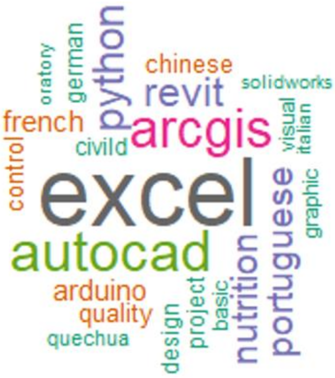


Figure 8. Word cloud on course preferences

Table 1 shows the results obtained after running 10 iterations using WEKA software and then applying the Elbow method.

Figure 9 shows the application of the elbow method, which consists of plotting the relationship between the number of clusters and the sum of squared error (SSE). In order to plot the graph, the data shown in Table 1, were considered an it was possible to observe that the inflection point is observed in the cluster k=2 (see Figure 9), indicating that this is the optimal number of clusters for the given data.

Figure 10 shows the graph generated in WEKA for cluster (k=2), where the x-axis shows the school or professional career and y-axis shows cluster 0 in blue and cluster 1 in red.

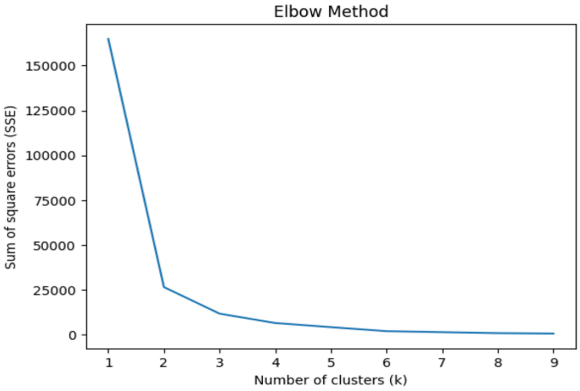
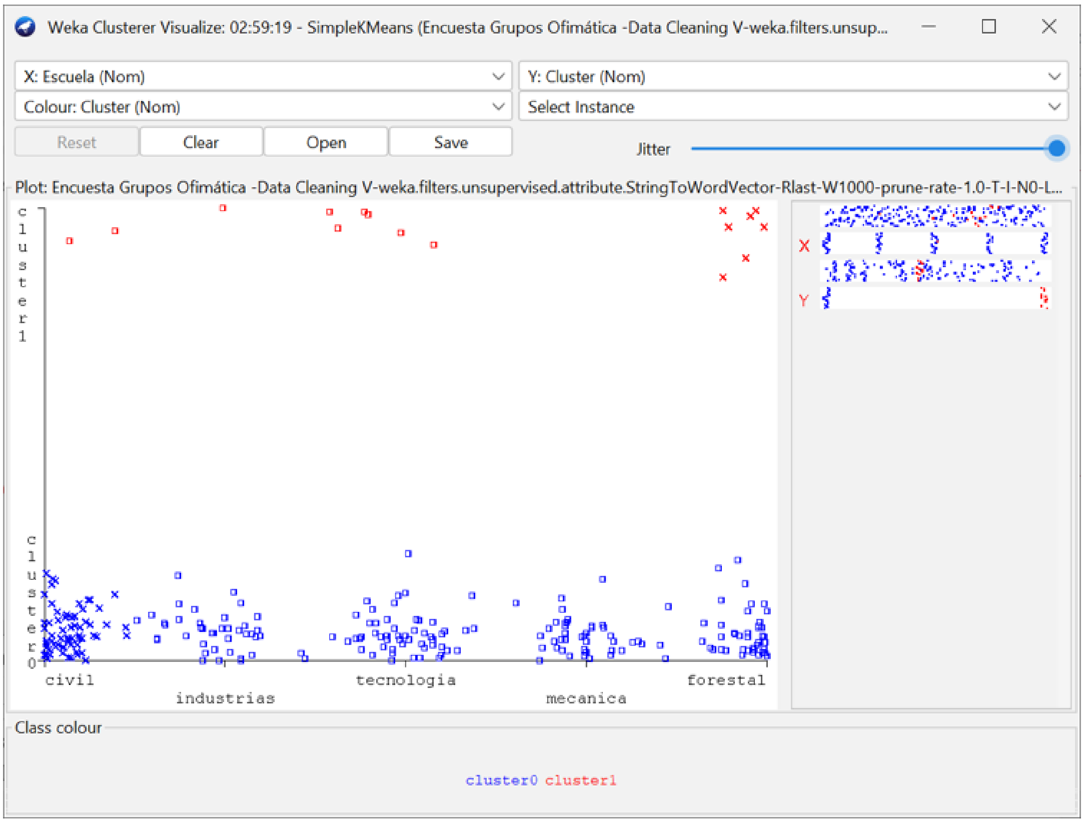


Figure 9. Elbow method to find the cluster

Table 1. Number of iterations and SEE

Number of Iterations	Sum of Square Errors (SSE)
1	336.48
2	279.80
3	253.17
4	248.27
5	228.15
6	217.60
7	209.38
8	176.38
9	156.71
10	82.70



Weka Clusterer Visualize: 02:59:19 - SimpleKMeans (Encuesta Grupos Ofimática -Data Cleaning V-weka.filters.unsup...

X: Escuela (Nom)

Y: Cluster (Nom)

Colour: Cluster (Nom)

Select Instance

Plot: Encuesta Grupos Ofimática -Data Cleaning V-weka.filters.unsupervised.attribute.StringToWordVector-Rlast-W1000-prune-rate-1.0-T-I-N0-L...

Figure 10. Cluster generated for k=2

4. DISCUSSION

The emergence and convergence of technologies like artificial intelligence, the Internet of Things, and automation in Industry 4.0 are also boosting parallel developments in the education sector [35]. For instance, ChatGPT is a type of artificial intelligence chatbot that can understand and respond to natural language inputs [36]. Thus, this tool has been designed to generate well-crafted texts indistinguishable from those produced by humans, applicable to a wide range of knowledge fields [37]. As a result, in recent years, the importance of integrating natural language processing (NLP) techniques for opinion extraction has been recognized [38]. In this context, text mining techniques were used to assess organization engagement with technologies like 5G networks, advanced robotics, artificial intelligence, autonomous driving, blockchain, and drones [39].

The academic paper is related to other research [18], since text mining tools such as the use of R libraries, tm package and WordCloud were used to form the word cloud and to verify the words frequency. On the other hand, four steps were used in the methodology, of which the preprocessing and data visualization was used in the present study [19].

The words with the highest frequency (Figure 7) are related to the academic training of students at the National University of Jaen in the five professional careers: Civil Engineering, Food Industry Engineering, Forestry and Environmental Engineering, Forestry and Environmental Engineering and Medical Technology. AutoCAD and Revit courses were the most recommended software by Civil Engineering students because engineering projects to be developed successfully make it difficult for any company to deal with a single program.

In the same way, the development of digital competences makes it essential for students to take on challenges that demonstrate continuous changes with the efficient performance of digital tools [40]. Consequently, ArcGIS was the software choice for Forestry and Environmental Engineering students due to the positioning accuracy of data collected by remote sensing platforms, which is of great importance in forestry and wildlife studies, salvage logging, soil disturbance after logging operations and fire risk management [41]. In health science, Nutrition was the recommended course for Medical Technology and Food Industry Engineering students; for example, nutrition education has a significant impact on material knowledge and child nutritional status [42]. Thus, the nutritional recommendations for bariatric and metabolic surgery aim to provide knowledge on different surgical techniques in the treatment obesity and metabolic diseases [43].

In relation to Python and Arduino, these were the tools needed by professional engineering careers, since the use of computational software like Python, role an important play in different areas knowledge, such as science, physics, chemistry and genetics [44]. In contrast, the digitalization in healthcare requires applications that use human sensors that are monitored in everyday life due to Internet of Things [45].

On the other hand, advanced Excel program was the most recommended by students of different professional careers, making it the most widely used for data analysis and visualization [46, 47]. Finally, languages such as Portuguese and French were the most requested, since the use of foreign language in professional communication facilitates participation in cultural, commercial, political and economic exchanges [48].

It is important to highlight the interconnection between the use of ChatGPT, NLP and text mining in relation to the objectives and results of the study. The incorporation of ChatGPT for interactive survey suggestion, NLP for text preprocessing, allowing for appropriate cleaning and extraction of relevant features for subsequent application of text mining techniques, which allowed the identification of patterns and trends within students' responses. The application of clustering techniques, like K-means algorithm, allowed the classification responses into meaningful groups. Furthermore, the word cloud creation allowed a concise and effective visualization frequent terms, providing a course preference. Theses interconnected technologies not only improved the quality results, but also provided a deeper understanding of common students' preferences for academic and curricular decision-making in educational institutions.

5. CONCLUSIONS

Text mining emerges as a field that seeks to improve people's perception about different topics, in which information and text patterns of great interest were presented [49]. In the context of this study, tools such as WEKA software, as well as R and Python programming languages were used to analyze the results of students' opinions from the Computer and Language Center of the National University of Jaen.

The results indicate that the most suggested and plausible courses by students form the professional career of Civil Engineering, Food Industry Engineering, Mechanical and Electrical Engineering, Forestry and Environmental Engineering were advanced Excel, AutoCAD, ArcGIS, Nutrition and Revit. However, it was important to note that beyond the mention of courses or programs, this study has deeper implication for the educational field. The findings provide valuable information for improving curriculum planning, teacher training, students' satisfaction, current courses trends and contribute to the existing text mining literature. It is recommended to explore how this result can influence in educational decision-making, contributed for the academic discussions and educational quality in general.

The research proves to be a seminal study for the exploration of future research on how course preferences related to academic performance and students' satisfaction, providing a more holistic view of the impact of these preferences on students' vocational training. Finally, it is recommended to consider incorporate sentiment analysis to capture emotions in students' opinions and to use a larger sample for future works.

REFERENCES

[1] Forman, H., Kerr, J., Norman, G.J., Saelens, B.E., Durant, N.H., Harris, S.K., Sallis, J.F. (2008). Reliability and validity of destination-specific barriers to walking and cycling for youth. Preventive medicine, 46(4): 311-316. https://doi.org/10.1016/j.ypmed.2007.12.006

[2] Church, K., Liberman, M. (2021). The future of computational linguistics: On beyond alchemy. Frontiers in Artificial Intelligence, 4: 625341. https://doi.org/10.3389/frai.2021.625341

[3] Nadkarni, P.M., Ohno-Machado, L., Chapman, W.W. (2011). Natural language processing: An introduction.

Journal of the American Medical Informatics Association, 18(5): 544-551. https://doi.org/10.1136/amiajnl-2011-000464

[4] Cambria, E., White, B. (2014). Jumping NLP curves: A review of natural language processing research. IEEE Computational Intelligence Magazine, 9(2): 48-57. https://doi.org/10.1109/MCI.2014.2307227

[5] Preethi, M., Radha, P. (2016). A survey paper on text mining-techniques, applications and issues. IOSR Journal of Computer Engineering (IOSR-JCE): 46-51.

[6] Vel, S.S. (2021). Pre-processing techniques of text mining using computational linguistics and python libraries. In 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, pp. 879-884. https://doi.org/10.1109/ICAIS50930.2021.9395924

[7] Hossain, A., Karimuzzaman, M., Hossain, M.M., Rahman, A. (2021). Text mining and sentiment analysis of newspaper headlines. Information (Switzerland), 12(10): 1-15. https://doi.org/10.3390/info12100414

[8] Isaeva, E., Aldarova, D. (2021). Text-mining in terms of methodology and development. In 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus), St. Petersburg, Moscow, Russia, pp. 413-416. https://doi.org/10.1109/ElConRus51938.2021.9396437

[9] Zhu, F., Patumcharoenpol, P., Zhang, C., Yang, Y., Chan, J., Meechai, A., Vongsangnak, W., Shen, B. (2013). Biomedical text mining and its applications in cancer research. Journal of Biomedical Informatics, 46(2): 200-211. https://doi.org/10.1016/j.jbi.2012.10.007

[10] Cheng, L.C., Sharmayne, L.R. (2020). Analysing digital banking reviews using text mining. In 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), The Hague, Netherlands, pp. 914-918. https://doi.org/10.1109/ASONAM49781.2020.9381429

[11] Gupta, K., Jiwani, N., Afreen, N. (2023). A Combined Approach of Sentimental Analysis Using Machine Learning Techniques. Revue d'Intelligence Artificielle, 37(1): 1-6. https://doi.org/10.18280/ria.370101

[12] Yadu, R., Shukla, R. (2022). A Hybrid Model Integrating Adaboost Approach for Sentimental Analysis of Airline Tweets. Revue d'Intelligence Artificielle, 36(4): 519-528. https://doi.org/10.18280/ria.360402

[13] Qorib, M., Oladunni, T., Denis, M., Ososanya, E., Cotae, P. (2023). Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. Expert Systems with Applications, 212: 118715. https://doi.org/10.1016/j.eswa.2022.118715

[14] Delen, D., Crossland, M.D. (2008). Seeding the survey and analysis of research literature with text mining. Expert Systems with Applications, 34(3): 1707-1720. https://doi.org/10.1016/j.eswa.2007.01.035

[15] Nanath, K., Joy, G. (2023). Leveraging Twitter data to analyze the virality of Covid-19 tweets: A text mining approach. Behaviour & Information Technology, 42(2): 196-214. https://doi.org/10.1080/0144929X.2021.1941259

[16] Shabat, H.A., Abbas, N.A. (2020). Independent component analysis based on natural gradient algorithm for text mining. In 2020 1st. Information Technology to Enhance e-learning and Other Application (IT-ELA), Baghdad, Iraq, pp. 72-76. https://doi.org/10.1109/IT-ELA50150.2020.9253072

[17] Wang, P., Zhou, Y., Sun, B., Zhang, W. (2019). Intelligent prediction of vulnerability severity level based on text mining and XGBboost. In 2019 Eleventh International Conference on Advanced Computational Intelligence (ICACI), Guilin, China, pp. 72-77. https://doi.org/10.1109/ICACI.2019.8778469

[18] Ishikiriyama, C.S., Miro, D., Gomes, C.F.S. (2015). Text mining business intelligence: A small sample of what words can say. Procedia Computer Science, 55: 261-267. https://doi.org/10.1016/j.procs.2015.07.044

[19] Chire, J., Apaza, H. (2020). Text mining over curriculum vitae of peruvian professionals using official scientific site DINA. In 2020 International Computer Symposium (ICS), Tainan, Taiwan, pp. 105-109. https://doi.org/10.1109/ICS51289.2020.00030

[20] Yan, H., Ma, M., Wu, Y., Fan, H., Dong, C. (2022). Overview and analysis of the text mining applications in the construction industry. Heliyon, 8(12): e12088. https://doi.org/10.1016/j.heliyon.2022.e12088

[21] Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., Romero, C. (2019). Text mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(6): e1332. https://doi.org/10.1002/widm.1332

[22] Chi, D. (2021). Research on the application of k-means clustering algorithm in student achievement. In 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), Guangzhou, China, pp. 435-438. https://doi.org/10.1109/ICCECE51280.2021.9342164

[23] Muhariya, A., Riadi, I., Prayudi, Y., Saputro, I.A. (2023). Utilizing K-Means Clustering for the Detection of Cyberbullying Within Instagram Comments. Ingénierie Des Systèmes d Information, 28(4): 939-949. https://doi.org/10.18280/isi.280414

[24] Alfian, G., Syafrudin, M., Yoon, B., Rhee, J. (2019). False positive RFID detection using classification models. Applied Sciences, 9(6): 1154. https://doi.org/10.3390/app9061154

[25] Li, Y., Wu, H. (2012). A clustering method based on K-means algorithm. Physics Procedia, 25: 1104-1109. https://doi.org/10.1016/j.phpro.2012.03.206

[26] Okoye, K., Daruich, S.D.N., De La O,J.F.E., Castaño, R., Escamilla, J., Hosseini, S. (2023). A text mining and statistical approach for assessment of pedagogical impact of students' evaluation of teaching and learning outcome in education. IEEE Access, 11: 9577-9596. https://doi.org/10.1109/access.2023.3239779

[27] Ahadi, A., Singh, A., Bower, M., Garrett, M. (2022). Text mining in education-A bibliometrics-based systematic review. Education Sciences, 12(3): 210. https://doi.org/10.3390/educsci12030210

[28] Thirumoorthy, K., Muneeswaran, K. (2021). An application of text mining techniques and outcome based education: student recruitment system. Journal of Ambient Intelligence and Humanized Computing, 14: 1359-1371. https://doi.org/10.1007/s12652-021-03162-4

[29] He, W. (2013). Examining students' online interaction in a live video streaming environment using data mining and text mining. Computers in Human Behavior, 29(1): 90-102. https://doi.org/10.1016/j.chb.2012.07.020

[30] Zhang, D., Zhang, J., Zhang, Y., Wu, Y. (2021).

Sentiment analysis of China's education policy online opinion based on text mining. In 2021 9th International Conference on Information and Education Technology (ICIET), Okayama, Japan, pp. 73-77. https://doi.org/10.1109/ICIET51873.2021.9419585

[31] Kusrini, K. (2015). Grouping of retail items by using k-means clustering. Procedia Computer Science, 72: 495-502. https://doi.org/10.1016/j.procs.2015.12.131

[32] Marutho, D., Handaka, S.H., Wijaya, E. (2018, September). The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. In 2018 International Seminar on Application for Technology of Information and communication, Semarang, Indonesia, pp. 533-538. https://doi.org/10.1109/ISEMANTIC.2018.8549751

[33] Nainggolan, R., Perangin-angin, R., Simarmata, E., Tarigan, A.F. (2019). Improved the performance of the K-means cluster using the sum of squared error (SSE) optimized by using the Elbow method. In Journal of Physics: Conference Series, 1361(1): 012015. https://doi.org/10.1088/1742-6596/1361/1/012015

[34] Kwedlo, W. (2011). A clustering method combining differential evolution with the K-means algorithm. Pattern Recognition Letters, 32(12): 1613-1621. https://doi.org/10.1016/j.patrec.2011.05.010

[35] Bhatia, A., Asthana, A., Bhattacharya, P., Tanwar, S., Singh, A., Sharma, G. (2023). A sentiment analysis-based recommender framework for massive open online courses toward education 4.0. Lecture Notes in Networks and Systems, 421: 817-827. https://doi.org/10.1007/978-981-19-1142-2_64

[36] Salvagno, M., Taccone, F.S., Gerli, A.G. (2023). Can artificial intelligence help for scientific writing? Critical Care, 27(1): 1-5. https://doi.org/10.1186/s13054-023-04380-2

[37] Dwivedi, Y.K., Kshetri, N., Hughes, L., Slade, E.L., Jeyaraj, A., Kar, A.K., Baabdullah, A.M., Koohang, A., Raghavan, V. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management, 71: 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

[38] Baroutsou, V., Cerqueira Gonzalez Pena, R., Schweighoffer, R., et al. (2023). Predicting openness of communication in families with hereditary breast and ovarian cancer syndrome: natural language processing analysis. JMIR Formative Research, 7: e38399. https://doi.org/10.2196/38399

[39] Spinazzola, M., Scuotto, V., Farronato, N., Pironti, M. (2022). Identifying synergies and barriers to the adoption of disruptive technologies for sustainable societies-an innovation ecosystem perspective. In 2022 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD), Marrakech, Morocco, pp. 1-6. https://doi.org/10.1109/ICTMOD55867.2022.10041880

[40] Santamaría-Peña, J., Rojo-Vea, S., Sanz-Adán, F. (2022). BIM workflows in the classroom: A topographical and earthworks experience with Autodesk Revit® and AutoCAD Civil3D®. Advances in Design Engineering, pp. 358-365. https://doi.org/10.1007/978-3-030-92426-3_41

[41] Đuka, A., Tomljanović, K., Franjević, M., Janeš, D., Žarković, I., Papa, I. (2022). Application and accuracy of unmanned aerial survey imagery after salvage logging in different terrain conditions. Forests, 13(12): 2054. https://doi.org/10.3390/f13122054

[42] Prasetyo, Y.B., Permatasari, P., Susanti, H.D. (2023). The effect of mothers' nutritional education and knowledge on children's nutritional status: A systematic review. International Journal of Child Care and Education Policy, 17(1): 11. https://doi.org/10.1186/s40723-023-00114-7

[43] Pereira, S.E., Rossoni, C., Cambi, M.P.C., et al. (2023). Brazilian guide to nutrition in bariatric and metabolic surgery. Langenbecks Archives of Surgery, 408(1): 143. https://doi.org/10.1007/s00423-023-02868-7

[44] Almosa, N.A.A., Al-Jilawi, A.S. (2023). Python optimization code for solving a mathematical modeling of Covid-19. In AIP Conference Proceedings, 2591: 050026. https://doi.org/10.1063/5.0119637

[45] Velichko, A., Huyut, M.T., Belyaev, M., Izotov, Y., Korzun, D. (2022). Machine learning sensors for diagnosis of COVID-19 disease using routine blood values for internet of things application. Sensors, 22(20): 1-29. https://doi.org/10.3390/s22207886

[46] Seid, G., Alemu, A., Dagne, B., Gamtesa, D.F. (2023). Microbiological diagnosis and mortality of tuberculosis meningitis: Systematic review and meta-analysis. PLoS One, 18: 1-14. https://doi.org/10.1371/journal.pone.0279203

[47] Duong, C.B., Van Tran, N., Nguyen, A.H., Le, T.N., Ha, B.H., Do, C.N.P., Huynh, K., Le, T.M., Nguyen, T.P., Nguyen, H.T.T. (2023). Impacts of COVID-19 crisis and some related factors on the mental health of 37150 Vietnamese students: A cross-sectional online study. BMC Public Health, 23(1): 1-12. https://doi.org/10.1186/s12889-023-15317-3

[48] Matijašević, M. (2022). Languages for specific purposes at the faculty of law, university of Zagreb. Folia Linguistica Et Litteraria, XIII (42): 73-93. https://doi.org/10.31902/fll.42.2022.6

[49] Rahman, A. (2019). Statistics-based data preprocessing methods and machine learning algorithms for big data analysis. International Journal of Artificial Intelligence, 17(2): 44-65.