QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (2), Spring 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

Correlation Evaluation Scale Through Text Mining Algorithms and Implementation on the Kurdish Language: A Review

Kazheen Ismael Taher
Department of Information Technology, Akre Technical College of Informatics, Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq
kajeen.ismael@gmail.com

Assist. Prof. Dr. JIHAN Abdulazeez Ahmed
Department of Computer Science, College of Science, University of Duhok , Duhok, Kurdistan Region, Iraq
Drjihanrasool@uod.ac

ARTICLE INFO

Article History:
Received: 26/5/2022
Accepted: 31/8/2022
Published: Spring 2023

Keywords: *Text Mining, Classification, Document clustering, Association rule mining,* Keyword extraction.

Doi:
10.25212/lfu.qzj.8.2.53

ABSTRACT

In recent times, because of the many articles that are found on the web, text extraction has become an interesting area of research. Text mining is a technique that can be used to extract useful information or knowledge from text documents that are not typically in an unstructured form. There are some studies conducted to use different techniques of text production for unstructured data sets. This study will provide an overview of the different methods and algorithms that are related to Text mining and also some studies on the mining of Kurdish web documentation. In addition, a collection of research problems and research methodologies will assist scholars in tracking their future research.

1. Introduction

Nowadays, text mining is the most prevalent way of exchanging information. However, deciphering the meaning of the writing is not a simple task. Text mining is often referred to as Text Analytics. Text mining is used to extract the related patterns, information, or knowledge from an unstructured form. Text mining is used to extract information from text which is complex formed and unstructured. There are two

QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (2), Spring 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

different components of the general framework of text mining, Text refining is the process of turning free-form text documents into an intermediate form, whereas knowledge distillation is the extraction of patterns or information from the intermediate form. Intermediate form (IF) can be semi-structured, as in conceptual graph representation, or organized, as in relational data representation. IF can be document-based (each entity represents a document) or concept-based (each entity represents an item or concepts of interest in a certain domain) (Inzalkar & Sharma, 2015; Sukanya & Biruntha, 2012).

Text mining is the use of automated technologies to comprehend the knowledge contained in text documents (Aggarwal & Zhai, 2012a, 2012b). Text mining may also be used to teach a computer to recognize, organized or unstructured material. Qualitative data, often known as unstructured data, is information that cannot be quantified. These data often include information such as colour, texture, and text. Quantitative data, often known as data structured, is that data can be easily quantified. Text mining is a variety of disciplines that combines data mining, machine learning, information retrieval, statistics, and other disciplines. Text mining is a distinct field from data mining (Wallace, Paul, Sarkar, Trikalinos, & Dredze, 2014). Text mining is a new field of research in computer science that seeks to address issues in data mining, natural language processing, information extraction, machine learning, information retrieval, classification, and knowledge management. The Text mining technique is depicted in Fig. 1 (Vijayarani, Ilamathi, & Nithya, 2015).

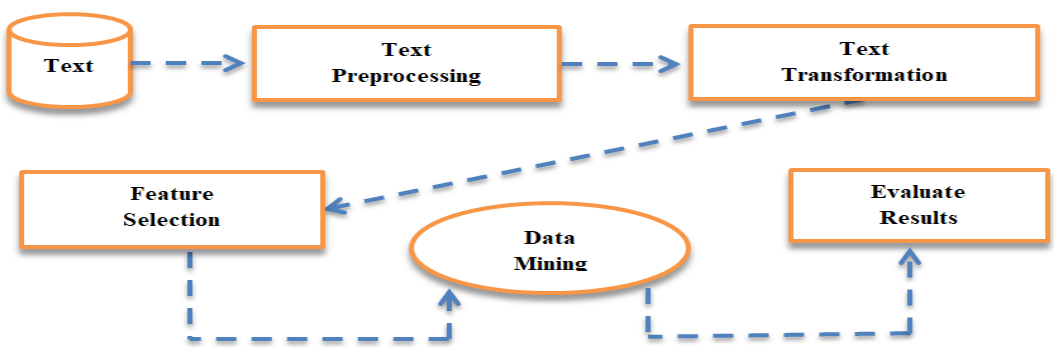

Figure 1: The Process of Text mining (A. Rashid, Shoaib, & ShahzadSarfraz).

QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (2), Spring 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

One of the primary reasons for the delay in using Text mining over the Kurdisis due to the online resources, lack of parallel corpora, and language processing tools (Ahmadi, 2019). Kurdistan is a hilly region that includes northern Iraq, south-eastern Turkey, northern Syria, and northwestern Iran. In contrast to English, the Sorani dialect of Kurdistan is written from right to left. The Kurdish language has 34 letters and several dialects, including Sorani and Kurmanji. Kurdish people in Iraq and Iran speak Kurdish Sorani (Malmasi, 2016). Vowels and consonants make up the Kurdish language. Vowels can be short or long, and there are 10 vowel phonemes in Kurdish. Consonants can assume several forms depending on where they appear in a word. Kurdish Sorani has a right-to-left writing style, comparable to Arabic script, whereas Kurmanji utilizes Latin script, which is left-to-right. Word order, nouns, absolute, indefinite, and definite states are all part of the Kurdish grammatical rules. In most cases, word arrangement or ordering includes the subject, object, and verb (Hamarashid, Saeed, & Rashid, 2021).

This paper is structured as follows: Sections 2 to 6 of this paper present the theory of Text mining, methodology of Text mining, and Kurdish text; Section 7 relates to related work. Section 8 discusses and compares, while Section 9 concludes the paper.

2. Text Mining

Text mining is the extraction of high-quality information from text. It is sometimes referred to as text data mining and is roughly equivalent to text analytics. High-quality data is typically obtained through deducing patterns and trends using techniques such as statistical pattern learning. Text mining is the process of organizing incoming text (typically parsing, with the inclusion of some derived linguistic components and the exclusion of others, and subsequent insertion into a database), generating patterns within the structured data, and lastly analyzing and interpreting the output. In text mining, "high quality" generally refers to a mix of relevance, interest, and distinctiveness (Berry & Castellanos, 2004; Ghosh, Roy, & Bandyopadhyay, 2012).

3. Text Mining methodology

Many technologies are used to assist Text_Mining such as Natural Language Processing, information extraction, Information Retrieval, categorization,

QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (2), Spring 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

summarization, Information Extraction, information visualization, and clustering are used in the text mining process. The relevance of each of these methods in text mining will be discussed in this section.

3.1 Information Retrieval (IR)

These technologies are generally used in libraries, where the papers are digital records containing information on the books rather than the books themselves. We may use IR systems to narrow down the texts that are relevant to a certain circumstance. Because text mining requires the use of costly algorithms on big collection of documents, IR can significantly speed up the analysis by lowering the document numbers to be analyzed (Hotho, Nürnberger, & Paaß, 2005) The Information Retrieval technique is depicted in Fig. 2.

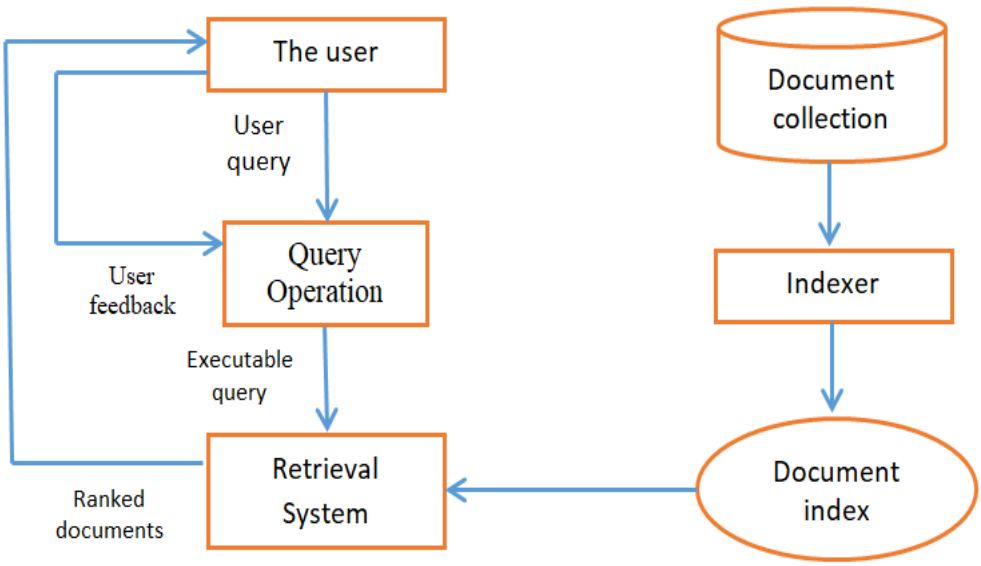

Figure 2: The Information Retrieval technique (Liu, 2011)

3.2 Natural Language Processing (NLP)

Is one of the most challenging and oldest challenges in the field of artificial intelligence. NLP is the study of human language for computers to comprehend natural languages. The function of NLP in Text mining is to give the linguistic data

QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (2), Spring 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

required by the systems in the information extraction phase. Shallow parsers detect just the most important grammatical parts in a sentence. Deep parsers provide a comprehensive representation of a sentence's grammatical structure. Although this aim is still a long way off, NLP can successfully do some forms of analysis (Runeson, Alexandersson, & Nyholm, 2007). Fig. 3 depicted the steps of NLP.

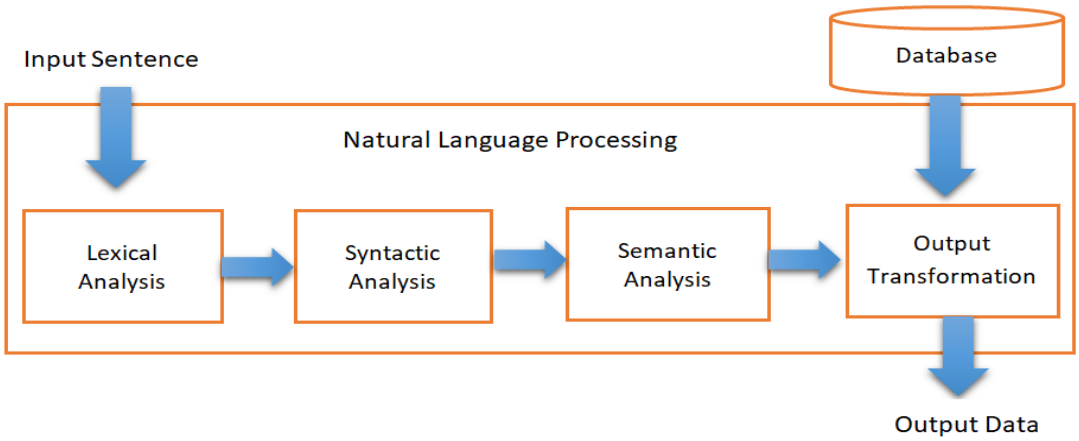

Figure 3: Steps Natural Language Processing (Tapsai, Meesad, & Haruechaiyasak, 2016).

3.3 Information-Extraction (IE)

Is the technique of extracting data structured from an unstructured natural language material mechanically. It accomplishes this by searching for predetermined sequences in text. The data to be "mined" in traditional data mining assumes that it already exists in the form of a relational database. Because IE performs the task of organizing a corpus of textual documents, the database built by an IE module may be sent to the KDD module for knowledge mining, as shown in Fig.4 (Gupta & Lehal, 2009).

QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (2), Spring 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

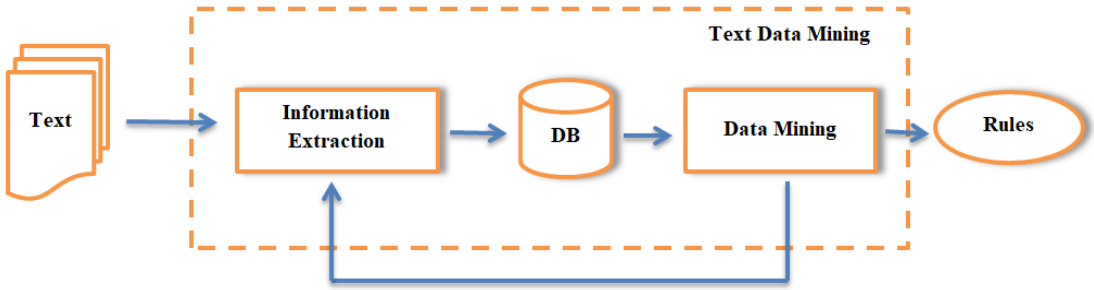

Figure 4: An overview of the Text mining framework based on IE (Kanya & Geetha, 2007).

3.4 Clustering

The clustering technique is an unsupervised procedure that uses several clustering algorithms to classify documents. Clustering involves organizing and extracting comparable words or patterns from several texts in both a top-down and bottom-up fashion (Shehata, Karray, & Kamel, 2006). Each cluster produces a division called clusters P, and each cluster contains several documents d, as seen in Fig, 3. When the contents of documents inside one cluster are more similar and those within clusters are more different, the clustering quality is rated higher. Although clustering is a technique used to group related articles, it varies from categorization in that documents in clustering are clustered on the fly rather than using predetermined themes (Gaikwad, Chaugule, & Patil, 2014; Solka, 2008).

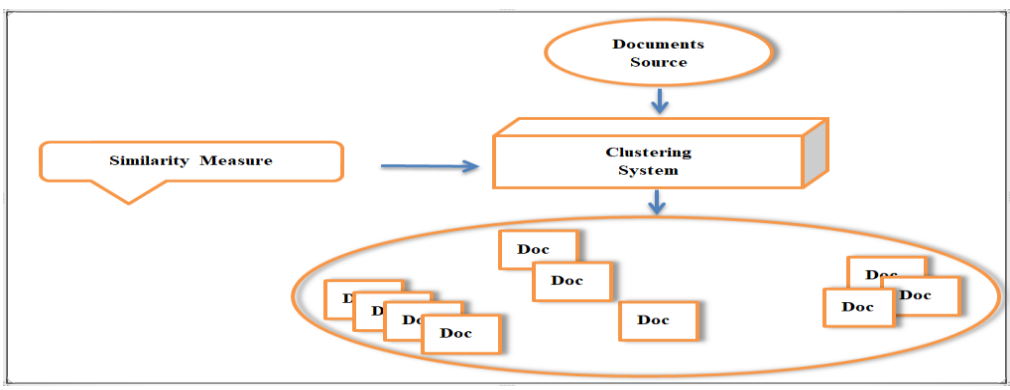

Figure 5: Clustering (Inzalkar & Sharma, 2015)

QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (2), Spring 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

3.5 Association Rule

Association rule mining identifies interesting association or correlation links among a huge range of data objects. In a nutshell, the association rule is predicated on linked connections. The finding of interesting association links among massive volumes of transaction information can aid in numerous decision-making processes. Association rules are developed based on two key terms: minimum support threshold and minimum confidence level (Kamruzzaman, Haider, & Hasan, 2010).

3.6 Information Visualization

Also known as information visualization, organizes enormous amounts of text into a visual hierarchy or map and allows for browsing in addition to simple searching. When a user has to narrow down a large number of papers and investigate related themes, information visualization can help. The user may interact with the page by performing a variety of actions like zooming, scaling, and so on (Inzalkar & Sharma, 2015). The government can use information visualization to discover terrorist networks or to learn more about incidents that were previously assumed to be unrelated. Text mining requires three critical processes in information visualization: data preparation, data analysis and extraction, and visualization mapping (Tandel, Jamadar, & Dudugu, 2019).

3.7 Text summarization

Text summarizing is the process of reducing the length and detail of a text while maintaining the most significant elements and overall meaning. It analyzes if a long document meets the user's requirements and is worth reading for further information. Text summary software analyzes and summarizes vast text content in the time it takes the user to read the first paragraph. Even while computers can detect people and places, teaching software to study semantics and decipher written documents is difficult (Gaikwad et al., 2014).

3.8 Categorization

Text documents are allocated predefined classifications based on their content. Categorization assigns one or more categories to a free-text document automatically. Categorization is used to train a classifier on known examples and then automatically

QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (2), Spring 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

categorize unexpected cases. Text may be classified using statistical classification techniques such as the Naive Bayesian classifier, the Decision Tree, and Support Vector Machines (Aggarwal & Zhai, 2012a; Kak, Mustafa, & Valente, 2018).

4. Text Mining Versus Data Mining

Text mining is similar to data mining in certain ways. In data mining, methods are developed to deal with structured data obtained from databases (Suresh & Harshni, 2017). Text mining, on the other hand, can be used to analyze semi-structured or unstructured data sets like full-text documents, emails, HTML files, etc. The main challenge with Text mining is that natural language was developed to allow people to interact and keep track of information as seen in Fig 6, but machines are still far behind in comprehending natural language (Gupta & Lehal, 2009; Inzalkar & Sharma, 2015).

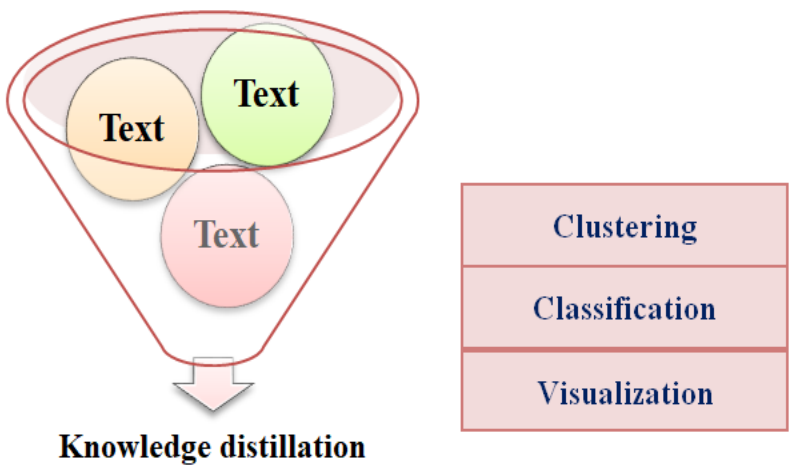

Figure 6: Knowledge extraction from unstructured text (Salloum, AlHamad, Al-Emran, & Shaalan, 2018).

The below compares and contrasts data mining with Text mining. according to some features such as (Salloum et al., 2018; Sukanya & Biruntha, 2012):

QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (2), Spring 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

4.1 Overview

A set of functions is used in data mining to seek patterns and correlations in structured data. Text mining uses a variety of procedures to convert unstructured textual material into structured information that can then be analyzed.

4.2 Type of data

The structured data from huge datasets found in systems such as databases, spreadsheets, ERP, CRM, and accounting applications is mined in data mining. Text mining is the process of analyzing unstructured textual data found in emails, papers, presentations, films, file transfers, social media, and the Internet.

4.3 Data retrieval

Structured data in data mining is homogeneous and ordered, making it easier to retrieve. Unstructured textual data in Text mining comes in a variety of forms and content kinds and is used in a wider range of applications and systems.

4.4 Data preparation

Structured data in data mining is formal and standardized, easing the process of feeding data into analytical models. Linguistic and statistical approaches, such as NLP keywording and meta-tagging, must be used in Text mining to convert unstructured data into useable structured data.

4.5 Taxonomy is needed.

There is no need to construct an overarching taxonomy for Text mining in data mining. Because unstructured text may take many various forms and formats in Text mining, there must be an overarching taxonomy for the data for it to be organized into a common framework.

5. Kurdish language

Kurdistan is the Kurds' homeland. Kurds are sometimes referred to as "a nation without a state". Kurdish usage and popularity have suffered as a result of this predicament. The situation appears to be changing now that the Iraqi Kurdistan region has begun to establish its regional administration (Hassani & Kareem, 2011;

QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (2), Spring 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

Hassani & Medjedovic, 2016). Kurdish is a language like any other in that it has its letters or characters. The Kurdish script, on the other hand, is similar to that of Arabic, Persian, and Urdu. It is made up of many dialects, including Sorani (Arabic script-based) and Kurmanji (Latin script-based). Sorani is spoken by Kurds in Iraq and Iran (MacKenzie, 1962). Fig. 7 depicts the Kurdish alphabet for both dialects (Hamarashid et al., 2021).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Arabic-based	ئ	ب	ج	چ	د	ی	ف	گ	ژ	ک	ل	م	ن	ڤ	پ	ق	ر	س	ش	ت	وو	ڤ	خ	ز
Latin-based	A	B	C	Ç	D	Ê	F	G	J	K	L	M	N	O	P	Q	R	S	Ş	T	Û	V	X	Z

(a) One-to-One Mapping

	25	26	27	28
Arabic-based	ێ / ی	و	ی	ه
Latin-based	I	U/W	C	Ç

(b) One-to-Two Mapping

	29	30	31	32	33
Arabic-based	ڕ	ڵ	ع	غ	ح
Latin-based	(RR)	-	(E)	(X)	(H)

(c) One-to-Zero Mapping

Figure 7: The two Alphabets of Standard Kurdish (Esmaili & Salavati, 2013)

6. Challenges the Kurdish language

The main challenges in the Kurdish language will be divided into five groups. The first two groups are concerned with the diversity aspect of the Kurdish language, the third and fourth highlight the processing difficulties and the last one examines the depth of resource scarcity for Kurdish (Kamal & Hassani, 2020).

6.1 Dialect Diversity

The multiplicity of Kurdish dialects is the first and most difficult issue in processing Kurdish texts. In terms of the number of speakers and degree of standardization, Kurmanji and Sorani are the two most major Kurdish dialects (Haig & Matras, 2002). The fundamental distinction between these two dialects is morphological. The most significant morphological distinctions are (Haig & Matras, 2002; Kamal & Hassani, 2020; MacKenzie, 1962):

QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (2), Spring 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

- Kurmanji is more orthodox in that he employs nouns and pronouns with both gender (feminine: masculine) and case opposition (absolute: oblique). Sorani has mostly abandoned this technique in favor of employing pronominal suffixes to complete the responsibilities of Case 3.
- Kurmanji has full ergative alignment in past-tense transitive verbs, but Sorani uses pronominal enclitics due to the loss of oblique pronouns.
- Passive and causative are generated purely by verb morphology in Sorani; they may also be formed in Kurmanji using the verbs hatin (to come) and dan (to give).
- Only in Sorani can the definite suffix -eke exist.

6.2 Script Diversity

Each of the two aforementioned dialects has its own writing system for geopolitical reasons. Kurdish is a bi-standard language, with Kurmanji written in Latin and Sorani written in Arabic. These two systems are virtually entirely phonetic. (Esmaili et al., 2013). As noted before, Sorani and Kurmanji are not morphologically identical and since these systems react to the phonology of their corresponding dialects, there is no bijective mapping between them (Edmonds, 1971).

6.3 Normalization

The Arabic-based Kurdish script's Unicode identifiers feature two potential causes of confusion that should be properly addressed:

- There are many Unicode characters for certain letters, such as ye and ka (Shamsfard, 2011). These multi-code letters should be united during the normalization phase.
- As in Urdu, the solitary and all variations of the Arabic letter ha compose one letter (pronounced e). The beginning and middle versions of the same Arabic letter, however, comprise another letter (pronounced h), for which a separate Unicode encoding is supplied (Walther & Sagot, 2010). Many electronic works employ just the ha, which is differentiated by a zero-width non-joiner character that prevents a character from being joined to its following. This distinction must be kept in mind during the normalization process.

QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (2), Spring 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

6.4 Segmentation and Tokenization

The process of identifying the borders of text elements, such as sentences, phrases, and words, is referred to as segmentation. In comparison to Persian and Arabic, this procedure is easier in Kurdish, owing to the explicit representation of short vowels in Kurdish writing systems. Despite including short vowels, the Kurdish alphabet, which is based on Arabic, has two flaws that it received from the Arabic writing system. (Volodina et al., 2019):

- Because the Arabic letter lacks capitalization, it is more difficult to distinguish between sentence boundaries and Named Entities. (Shamsfard, 2011).
- Space does not serve as a deterministic delimiter and sign for boundaries. It can be found within a word, between words, or between syllables. In Persian, there are various suggestions about how to address this issue.(Shamsfard, Jafari, & Ilbeygi, 2010) and Urdu (Rehman, Anwar, & Bajwa, 2011).

7. Related Work

7.1 Applications of data mining algorithms and Text mining

(Jalal & Ali, 2021) developed a method for categorizing documents that may cluster the text documents of research articles into relevant groups that have a common scientific topic. The document weight is influenced by the frequency of word tokens in documents, which is determined using a numerical statistic of term frequency-inverse document frequency (TF-IDF). The suggested technique to execute the categorization process employs the title, abstract, and keywords of the publication, as well as the category themes. Documents are sorted and grouped into major categories based on the greatest measure of cosine similarity between category weight and document weight.

(Belwal, Rai, & Gupta, 2021), suggested a novel graph-based summarization technique that, in addition to considering phrase similarity, also takes into consideration sentence similarity and the whole (input) text. Two qualities are taken into account while distributing the weight across the graph's edges. The nodes that make up the edges of the graph's initial attribute are comparable to one another.

QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (2), Spring 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

The second property is the weight assigned to a component that indicates how much a certain edge resembles the overall document's subjects for which topic modeling was used. They utilized the semantic measure in conjunction with these adjustments to determine how similar the nodes were. The evaluation results of the suggested method show a considerable increase in summary quality over existing text summarizing techniques.

(Gao et al., 2020), proposed a new framework for topic evolution mining in short texts They initially propose an Encoder Only Transformer Language Model (ETLM) to quantify the link between words. Next, suggested an unique topic model CCTM that integrates global and local semantic correlations by employing a Weighted Conditional Random Field (WCRF) to promote semantically similar words to share the same subject label. Finally, OCCTM is utilized to automatically detect topics and topic relationships at each time slice, as well as to generate topic evolutionary graphs. Ex820 experiments using real-world short text collections verify the efficacy of their suggested strategies. On the future, they plan to create a parallel version of OCCTM in a distributed system and test it on more full and large-scale datasets.

(Shao, Li, Wang, Zhao, & Guo, 2020) proposed TA-ARM, a new concept map automated generation technique based on text analysis and association rules mining. The algorithm generates concept maps by combining the association rule mining approach with the text classification algorithm in text analysis technology. The experimental results demonstrate that the TA-ARM algorithm can construct the concept map automatically and quickly, reducing the effect of outside experts while also dynamically adjusting the concept map based on criteria like the confidence threshold between test questions.

(Dai, 2021) proposed an algorithm of text data mining based on a convolutional neural network (CNN) model and a Deep Boltzmann Machine (DBM) model was suggested. This technique combines CNN and DBM models with effective feature extraction to accomplish double feature extraction. It may actualize the tag tree by creating the tag tree and establishing an appropriate hierarchical network to achieve classification. Simultaneously, the model may minimize the input noise in categorisation. The

QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (2), Spring 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

results of the experiments reveal that the modified model has a positive influence on the text in the given area.

(Zhang, Fleyeh, Wang, & Lu, 2019) suggested five basic models: support vector machine (SVM), Naive Bayes (NB), linear regression (LR), decision tree (DT), k-nearest neighbor (KNN), and an ensemble model to diagnose the causes of accidents. In addition, the weight of each classifier in the ensemble model is optimized using the Sequential Quadratic Programming (SQP) approach. According to the findings of the experiments, the optimized ensemble model outperforms the other models in terms of average weighted F1 score. The results also suggest that the proposed method is more resistant to low support scenarios. Furthermore, based on the grammatical rules revealed in the reports, an unsupervised chunking strategy is presented to identify common items that cause accidents.

(J. Kim, Jang, Park, & Choi, 2020) recommended the use of capsule networks in the domain of text categorization, as well as the use of a static routing alternative They categorize text using capsule networks with dynamic routing and get results that are comparable to previous techniques. Then they suggested an alternate routing strategy that outperforms dynamic routing in terms of accuracy. In addition, they propose using an ELU-gate to spread crucial information. Based on seven common benchmark datasets, they compared the suggested model to CNNs and demonstrated that capsule networks are useful for text categorization. They also provided static routing as an alternative to dynamic routing, which results in higher classification accuracy while needing less computation.

(Agnihotri, Verma, & Tripathi, 2014) focused on extracting the most critical information from text data for the experimental investigation. The authors employed the tales data set from Project Guttenberg's William Shakespeare stories dataset. In the Ubuntu 12.04 LTS Linux Operating System, R is employed as a text mining and statistical analysis tool. Frequent pattern mining is a technique for locating often occurring phrases in documents and text. A threshold value is used to assess the degree of association between two or more words. Before grouping, their method measures the distance between words using cosine similarity. The algorithm may be

QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (2), Spring 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

used to compare articles, news, and emails for resemblance. The cluster is formed using the hierarchical agglomerative clustering and k-means algorithms.

(Saura, Palos-Sanchez, & Grilo, 2019) used new technologies in the suggested study technique to discover the critical criteria for the success of start-up enterprises. A Latent Dirichlet Allocation (LDA) model was employed, which is a cutting-edge theme modeling method written in Python that selects the database subject by analyzing tweets with the hashtag #Startups on Twitter. A sentiment analysis was carried out in Python using a Supervised Vector Machine (SVM) technique that works with Machine Learning. This study indicated that startup tools, technology-based startups, founder attitude, and startup methodology development are the topics with good feelings for defining critical components for startup firm success. The identified neutral themes include the creation of the business strategy, the type of startup project, and the geography of the incubator and company.

(J.-C. Kim & Chung, 2019) proposed to extract associative feature information from large health data sets using Text mining. Text mining is used to extract meaningful information from health documentation. Health papers are acquired as raw data via Web scraping and saved on a file server. To assess the relevance of words in a series of texts, the TF-C-IDF method is applied to the candidate corpus. The association rules of keywords in the formed transaction are examined using an Apriori mining method, and associative keywords are generated. The proposed method is a foundational technology for adding value to the healthcare business in the fourth industrial revolution. Its performance was rated highly in terms of F-measure and efficiency.

(Goh & Ubeynarayana, 2017) aimed to evaluate the efficacy of several text mining classification approaches in classifying 1000 publically available construction accident tales acquired from the US OSHA website Six machine learning approaches were explored in the study, including support vector machine (SVM), k-nearest neighbor (KNN), linear regression (LR), decision tree (DT), random forest (RF), and Naive Bayes (NB), and discovered that SVM performed the best in categorizing the test set of 251 instances. Experimentation with tokenization of the processed text and nonlinear SVM was also carried out. Because of its simplicity, the linear SVM is chosen. The

QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (2), Spring 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

linear SVM's accuracy ranged from 0.5 to 1. The causes of the misclassification were examined, as well as ideas for how to improve performance.

(Boukil, Biniz, El Adnani, Cherrat, & El Moutaouakkil, 2018) presented a novel method for classifying Arabic text from large datasets. As a baseline, they evaluate their dataset using CNNs and other classic machine learning models. To extract, select, and reduce the amount of information that is required, they employ an Arabic stemming algorithm. Then, as a feature weighting strategy, they employ the Term Frequency Inverse Document Frequency technique. Finally, for the classification stage, they employ Convolutional Neural Networks, a deep learning method that is quite strong in other fields such as image processing and pattern recognition, but is still infrequently used in text_mining. They can obtain outstanding performance on various benchmarks with this combination and some hyperparameter modifications in the Convolutional Neural Networks technique.

Now, will discuss the comparison among ten new research on applications of data mining algorithms and Text mining, as shown in Table 1.

Table (1): Summary of review related to data mining and Text mining algorithms.

Ref.	Dataset	Tools	Methodologe	Finding
(Jalal & Ali, 2021)	Text documents	similarity measure	web mining cosine similarity TF-IDF data extraction	Using the cosine similarity approach, it was discovered that more than 96% of publications could be classified into comparable ranges.
(Shao et al., 2020)	Selects 617,940 original answer records from 6,866 students in a Computer Culture Foundation large-scale test as an empirical dataset	Python 3.6 PyCharm Community Edition 2018 SQL Server 2008	(k-NN and Apriori) Algorithms	the TA-ARM algorithm showed that can generate a concept map automatically and quickly
(Dai, 2021)	D1(No. samples = 9666, No. catagories = 39, Multiclass type)	recall rate Fsimilarity S rate P Rmeasure value	(CNN and DBM) Models	The advanced model has been shown to have a positive effect on the text special domain.

QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (2), Spring 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

	D2 (No. samples = 1000, No. catagories = 168, Multi categories) D3(No. samples = 1000000, No. catagories =150, Multilabel type)			
(Zhang et al., 2019)	OSHA is (800 training samples and 200 testing samples)	Python 2.7 matplotlib v2.1.2 sklearn v0.19.1 matplotlib v2.1.2 nltk v3.2.5 pandas v0.22.0	(SVM, LR, KNN, DT, NB) and an ensemble model Sequential Quadratic Programming (SQP) algorithm	Showed that the optimal group model is superior to the rest models considered also, the proposed approach is more robust for less supportive situations
(J. Kim et al., 2020)	7 benchmark (20,000 news, Reuters 10, MR (2005), MR (2004), TREC-QA, IMDb, and MPQA.	Static routing variant Python natural language	Capsule networks	higher classcation accuracies with less computation.
(Agnihotri et al., 2014)	William Shakespeare Story Data for Experimental Research from Guttenberg.	R used as Text-Mining LTS Linux Operating System Ubuntu 12.04	k-means hierarchical agglomerative clustering	Much of the information is stored as text such as technical papers, web pages, books, news articles, emails, blogs, and digital libraries. Hence, Text-mining search has been very active.
(Saura et al., 2019)	User-Generated Content (UGC) hashtag #Startups on Twitter (n = 35.401)	Python Nvivo software	Latent Dirichlet Allocation (LDA) SVM	Displayed the name and description of the subject and specific feelings (negative, positive, or neutral)
(J.-C. Kim & Chung, 2019)	Health Big Data documents are written in natural language	R 3.4.1. F-measure efficiency values	TF-C-IDF TF–IDF TF Apriori	The suggested TF-C-IDF approach produced better average values than other methods
(Goh & Ubeyn arayan	A total of 1000 construction accident reports were acquired from the US OSHA website.	N-grams Python 2.7	6 machine learning algorithms (SVM, DT, LR,	found that SVM performed better in the distribution of the test group of 251 cases. The SVM score was 0.5 to 1,

QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (2), Spring 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

a, 2017)			RF, NB, and KNN)	the memory was 0.36 to 0.9 and the F1 point was between 0.45 and 0.92.
(Boukil et al., 2018)	319 million Arabic words corpus D_27k: Size (No.doc.=27,932) D_55k: Size (No.doc.= 55,864) D_83k: Size (No.doc= 83,796) D_111k: Size (No.doc. = 111,728)	Softmax max-pooling	TF-IDF CNN	CNNs(D_27k: 86.30, D_55k: 87.12, D_83k: 89.75, D_111k: 92.94)

7.2 Implementation of the Kurdish language

(Ahmadi, 2019) suggested a rule-based approach for transliterating Kurdish text. Kurdish faces several obstacles when it comes to transliterating its two most common orthographies, Arabic-based and Latin-based. They provided a solution for overcoming these obstacles by utilizing the Wergor transliteration system. They highlight many obstacles in Kurdish text mining and suggest fresh methods for the Sorani Kurdish translation task. The issue of transliteration of the two most often used orthographies for Sorani Kurdish, Arabic-based and Latin-based, Werger, their transliteration system, achieves an overall accuracy of 82.79% and a detection rate of more than 99% for double-usage letters. They also give a Kurdish corpus that has been painstakingly transliterated.

(T. A. Rashid, Mustafa, & Saeed, 2017) focused on text categorization of Kurdish text documents to place articles or emails in the appropriate class based on their contents. In the study, a new dataset entitled KDC-4007 was produced that may be widely used in text categorization research of Kurdish news and articles. On the KDC-4007, comparisons of three well-known techniques for text classification (such as Support Vector Machine (SVM), Naive Bays (NB), and Decision Tree (DT) classifiers) and the TF X IDF feature weighting approach are made. The experimental findings show that the SVM classifier provides an excellent accuracy value of 91.03%, especially when stemming and TF X IDF feature weighting are used in the preprocessing phase.

(Saeed et al., 2018) dedicated to refining a technique for categorizing Kurdish languages using Reber Stemmer Thus, an unique technique for obtaining the stems

QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (2), Spring 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

of Kurdish words by eliminating the longest suffixes and prefixes is being investigated. The stemming method is used on the eight-class KDC-4007 dataset. Support Vector Machines (SVM) and Decision Trees (DT or C 4.5) are used for categorization. The longest match stemmer technique has been effectively compared to this stemmer. The F-measure of the Reber stemmer and the Longest-Match approach in SVM are larger than DT, according to the data. In SVM, Reber stemmer obtained a higher F-measure for classes (religion, sport, health, and education), whilst the rest of the classes obtained a lower F-measure in Longest-Match.

(Abdulla & Hama, 2015) proposed data mining classification algorithms such as the Naive Bayes classifier At work, comments were collected from the microblogs Twitter, Facebook, and Google+, and Naive Bayes was used to classify "documents" into positive and negative sentiment. The social network remarks are divided into positive or negative polarity in the experimental findings. The accuracy of sentiment analysis is 66% when employing a Naive Bayes classifier for the unigram feature on a Kurdish text sample. They also showed the research that adequate accuracy may be obtained when compared to similar works. The techniques presented in the work can be extended to other areas of Kurdish text categorization.

Now, will discuss the comparison among other four new research on applications of an Implementation of the Kurdish language, as shown in Table2.

Table (2): Summary of review related to the implementation of the Kurdish language.

Ref.	Problem	Datasets	Dialect	Method	Finding
(Ahmadi, 2019)	Translation of two famous orthographies, Latin-based and Arabic-based.	36 top-ranked Kurdish websites	Sorani Kurdish	Wergor transliteration CCS Concepts	achieves an overall precision of 82.79 % and a detection rate of more than 99 % for double-usage characters, as well as a manually transliterated corpus for Kurdish.

QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (2), Spring 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

(T. A. Rashid et al., 2017)	The large number of text documents uploaded daily to the Internet.	KDC 4007	Sorani	SVM NB DT TF x IDF	The SVM classifier has a good accuracy value of 91.03 %.
(Saeed et al., 2018)	The high-dimensionality of the linguistic difference advantage of space for the word	KDC-4007	Sorani	SVM F-measure DT or C 4.5	10.11 s and 9.97 s for SVM in the suggested technique. The time it takes to construct a model in C4.5 is 668.35 s.
(Abdulla & Hama, 2015)	Positive or negative user reviews of text information for social networks	15,000 text files	Sorani	Naive Bayes	Sentiment analysis accuracy of 66% was obtained using Naive Bayes classifier for unigram feature on Kurdish script

8. Discussion

After a review of several papers on the two sections, in general, researchers used various methods of Text mining for their proposal. The first section of the survey is on the applications of data mining and algorithms of Text mining. Table 1 discussed the data mining and Text mining algorithms, researchers depend on some features such as the tools, methods, and dataset, they have different results such as [39] used OSHA is (800 training samples and 200 testing samples) dataset, Python 2.7, Python 2.7, matplotlib v2.1.2, sklearn v0.19.1, matplotlib v2.1.2, , nltk v3.2.5, pandas v0.22.0 and software, SVM, LR, KNN, DT, NB) and an ensemble model, Sequential Quadratic Programming (SQP) algorithm, their result showed that the optimal group model is superior to the rest models considered, also the proposed approach is more robust for less supportive situations. Also [44] 1000 available construction accident accounts from the US OSHA website obtained, N-grams, Python 2.7 tools, 6 machine learning algorithms (SVM, LR, RF, KNN, NB, DT), and they found that SVM performed better in the distribution of the test group of 251 cases. . The SVM score was 0.5 to 1, the memory was 0.36 to 0.9 and the F1 point was between 0.45 and 0.92. Table 2 discussed the implementation of the Kurdish language, this language has a challenge

in writeng, the reason for these challenges has multi dialects such as Sorani, Kurmanji, Zazaki, and Gorani. After reviewing most researchers focused on Sorani like [47], The large number of text documents uploaded daily to the Internet, KDC 4007 dataset, Sorani dialect, SVM, NB, and DT methods, SVM classifier has a good accuracy value of 91.03 %.

The objective of this work is to use an automated scale instead of critics to evaluate articles and direct readers to read useful and good press articles by generating a correlation evaluation scale through data mining methods by finding the frequency of title keywords within the text, which represents the degree of attachment of the text to its title. Relying on a dictionary of non-key words and a dictionary of candidate words to be keywords, The accuracy ratios of the scale were verified by relying on tables comparing the accuracy achieved by the scale and what the scale achieved. It is decided by experienced surfers by the correlation ratio.

9. Conclusion

Currently, the Internet is the primary source of information for a vast number of individuals. Therefore, users need to be able to easily and simply find their interesting requests that represent the most relevant information of the query. However, the search engine returns more unrelated pages based on a few user query keywords, which results in long URLs. Searching the web pages to discover the knowledge according to the user's question is not an easy task. Text reduction is the process of extracting interesting information, knowledge, or templates from irregular text that comes from a variety of sources. This study will provide an overview of the various approaches and algorithms used in the field of writing, with particular emphasis on metamorphic methods. In addition, attention has been paid to the Sorani dialect and web documents, in general to numerous studies on the excavation of Kurdish texts. In addition, the synthesis of research questions and the methodology of the researched papers will support the authors of the text in their research.

References:

Abdulla, S., & Hama, M. H. (2015). Sentiment analyses for Kurdish social network texts using Naive Bayes classifier. *Journal of University of Human Development, 1*(4), 393-397.

Aggarwal, C. C., & Zhai, C. (2012a). An introduction to text mining *Mining text data* (pp. 1-10): Springer.

Aggarwal, C. C., & Zhai, C. (2012b). *Mining text data*: Springer Science & Business Media.

Agnihotri, D., Verma, K., & Tripathi, P. (2014). *Pattern and Cluster Mining on Text Data*.

Ahmadi, S. (2019). A rule-based Kurdish text transliteration system. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 18*(2), 1-8.

Belwal, R. C., Rai, S., & Gupta, A. (2021). A new graph-based extractive text summarization using keywords or topic modeling. *Journal of Ambient Intelligence and Humanized Computing, 12*(10), 8975-8990.

Berry, M. W., & Castellanos, M. (2004). Survey of text mining. *Computing Reviews, 45*(9), 548.

Boukil, S., Biniz, M., El Adnani, F., Cherrat, L., & El Moutaouakkil, A. E. (2018). Arabic text classification using deep learning technics. *International Journal of Grid and Distributed Computing, 11*(9), 103-114.

Dai, R. (2021). Text Data Mining Algorithm Combining CNN and DBM Models. *Mobile Information Systems, 2021*.

Edmonds, C. J. (1971). Kurdish nationalism. *Journal of contemporary history, 6*(1), 87-107.

Esmaili, K. S., Eliassi, D., Salavati, S., Aliabadi, P., Mohammadi, A., Yosefi, S., & Hakimi, S. (2013). *Building a test collection for Sorani Kurdish.* Paper presented at the 2013 ACS International Conference on Computer Systems and Applications (AICCSA).

Esmaili, K. S., & Salavati, S. (2013). *Sorani Kurdish versus Kurmanji Kurdish: an empirical comparison.* Paper presented at the Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).

Gaikwad, S. V., Chaugule, A., & Patil, P. (2014). Text mining methods and techniques. *International Journal of Computer Applications, 85*(17).

Gao, W., Peng, M., Wang, H., Zhang, Y., Han, W., Hu, G., & Xie, Q. (2020). Generation of topic evolution graphs from short text streams. *Neurocomputing, 383*, 282-294.

Ghosh, S., Roy, S., & Bandyopadhyay, S. K. (2012). A tutorial review on Text Mining Algorithms. *International Journal of Advanced Research in Computer and Communication Engineering, 1*(4), 7.

Goh, Y. M., & Ubeynarayana, C. (2017). Construction accident narrative classification: An evaluation of text mining techniques. *Accident Analysis & Prevention, 108*, 122-130.

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence, 1*(1), 60-76.

Haig, G., & Matras, Y. (2002). Kurdish linguistics: a brief overview. *STUF-Language Typology and Universals, 55*(1), 3-14.

Hamarashid, H. K., Saeed, S. A., & Rashid, T. A. (2021). Next word prediction based on the N-gram model for Kurdish Sorani and Kurmanji. *Neural Computing and Applications, 33*(9), 4547-4566.

Hassani, H., & Kareem, R. (2011). *Kurdish text to speech (KTTS).* Paper presented at the Tenth International Workshop on Internationalisation of Products and Systems.

Hassani, H., & Medjedovic, D. (2016). Automatic Kurdish dialects identification. *Computer Science & Information Technology, 6*(2), 61-78.

Hotho, A., Nürnberger, A., & Paaß, G. (2005). *A brief survey of text mining.* Paper presented at the Ldv Forum.

Inzalkar, S., & Sharma, J. (2015). A survey on text mining-techniques and application. *International Journal of Research In Science & Engineering, 24*, 1-14.

Jalal, A. A., & Ali, B. H. (2021). Text documents clustering using data mining techniques. *International Journal of Electrical & Computer Engineering (2088-8708), 11*(1).

Kak, S. F., Mustafa, F. M., & Valente, P. (2018). A review of person recognition based on face model. *Eurasian Journal of Science & Engineering, 4*(1), 157-168.

Kamal, Z., & Hassani, H. (2020). *Towards Kurdish text to sign translation.* Paper presented at the Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives.

Kamruzzaman, S., Haider, F., & Hasan, A. R. (2010). Text classification using data mining. *arXiv preprint arXiv:1009.4987*.

Kanya, N., & Geetha, S. (2007). Information Extraction-a text mining approach.

Kim, J.-C., & Chung, K. (2019). Associative feature information extraction using text mining from health big data. *Wireless Personal Communications, 105*(2), 691-707.

Kim, J., Jang, S., Park, E., & Choi, S. (2020). Text classification using capsules. *Neurocomputing, 376*, 214-221.

Liu, B. (2011). Information retrieval and Web search *Web Data Mining* (pp. 211-268): Springer.

MacKenzie, D. N. (1962). *Kurdish Dialect, Studies 1* (Vol. 2): Oxford University Press.

Malmasi, S. (2016). *Subdialectal differences in sorani kurdish.* Paper presented at the Proceedings of the third workshop on nlp for similar languages, varieties and dialects (vardial3).

Rashid, A., Shoaib, U., & ShahzadSarfraz, M. KNOWLEDGE DISCOVERY IN DATABASE USING INTENTION MINING.

Rashid, T. A., Mustafa, A. M., & Saeed, A. M. (2017). *Automatic Kurdish text classification using KDC 4007 dataset.* Paper presented at the International Conference on Emerging Internetworking, Data & Web Technologies.

Rehman, Z., Anwar, W., & Bajwa, U. I. (2011). *Challenges in Urdu text tokenization and sentence boundary disambiguation.* Paper presented at the Proceedings of the 2nd Workshop on South Southeast Asian Natural Language Processing (WSSANLP).

Runeson, P., Alexandersson, M., & Nyholm, O. (2007). *Detection of duplicate defect reports using natural language processing.* Paper presented at the 29th International Conference on Software Engineering (ICSE'07).

Saeed, A. M., Rashid, T. A., Mustafa, A. M., Agha, R. A. A.-R., Shamsaldin, A. S., & Al-Salihi, N. K. (2018). An evaluation of Reber stemmer with longest match stemmer technique in Kurdish Sorani text classification. *Iran Journal of Computer Science, 1*(2), 99-107.

Salloum, S. A., AlHamad, A. Q., Al-Emran, M., & Shaalan, K. (2018). A survey of Arabic text mining *Intelligent Natural Language Processing: Trends and Applications* (pp. 417-431): Springer.

Saura, J. R., Palos-Sanchez, P., & Grilo, A. (2019). Detecting indicators for startup business success: Sentiment analysis using text data mining. *Sustainability, 11*(3), 917.

Shamsfard, M. (2011). Challenges and open problems in Persian text processing. *Proceedings of LTC, 11*.

Shamsfard, M., Jafari, H. S., & Ilbeygi, M. (2010). *STeP-1: A Set of Fundamental Tools for Persian Text Processing.* Paper presented at the LREC.

Shao, Z., Li, Y., Wang, X., Zhao, X., & Guo, Y. (2020). Research on a new automatic generation algorithm of concept map based on text analysis and association rules mining. *Journal of ambient intelligence and humanized computing, 11*(2), 539-551.

Shehata, S., Karray, F., & Kamel, M. (2006). *Enhancing text clustering using concept-based mining model.* Paper presented at the Sixth International Conference on Data Mining (ICDM'06).

Solka, J. L. (2008). Text data mining: theory and methods. *Statistics Surveys, 2*, 94-112.

Sukanya, M., & Biruntha, S. (2012). *Techniques on text mining.* Paper presented at the 2012 IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT).

Suresh, R., & Harshni, S. (2017). *Data mining and text mining—a survey.* Paper presented at the 2017 International Conference on Computation of Power, Energy Information and Commuincation (ICCPEIC).

Tandel, S. S., Jamadar, A., & Dudugu, S. (2019). *A survey on text mining techniques.* Paper presented at the 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS).

Tapsai, C., Meesad, P., & Haruechaiyasak, C. (2016). *TLS-ART: Thai language segmentation by automatic ranking trie.* Paper presented at the 9th International Conference Autonomous Systems.

Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks, 5*(1), 7-16.

Volodina, E., Granstedt, L., Matsson, A., Megyesi, B., Pilán, I., Prentice, J., . . . Sundberg, G. (2019). The SweLL Language Learner Corpus: From Design to Annotation. *Northern European Journal of Language Technology, 6*, 67-104.

Wallace, B. C., Paul, M. J., Sarkar, U., Trikalinos, T. A., & Dredze, M. (2014). A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. *Journal of the American Medical Informatics Association, 21*(6), 1098-1103.

Walther, G., & Sagot, B. (2010). *Developing a large-scale lexicon for a less-resourced language: General methodology and preliminary experiments on Sorani Kurdish.* Paper presented at the Proceedings of the 7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages (LREC 2010 Workshop).

QALAAI ZANISTSCIENTIFIC JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil, Kurdistan, Iraq
Vol. (8), No (2), Spring 2023
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

Zhang, F., Fleyeh, H., Wang, X., & Lu, M. (2019). Construction site accident analysis using text mining and natural language processing techniques. *Automation in Construction, 99*, 238-248.

پێوەری هەڵسەنگاندنی پەیوەندی لە ڕێگەی ئەلگۆریزمەکانی دەرهێنانی دەق و جێبەجێکردن لەسەر زمانی کوردی: پێداچوونەوەیەک

پـوختـه:

دەرهێنانی دەقەکان بوتە جێگای سەرنج لەبواری لێکولینەوەکان لەبەر هەبونی چەندین بابەت لە ویبسایتی ئەنتەرنێت لەچەند سالی رابردوو دا. (Text Mining) پێناسە دەکرێت وەک تەکنیکێک وە بەکار دێت بو دەرهێنانی زانیاریەکان یان زانستە سەرنج راکێشەکان لەبەلگەنامە دەقییەکان کە بە شێوەیەکی گشتی ڕێکخراو نین. هەندێک لە لێکولینەوەکان ئەنجامدراون بە بەکارئێنانی تەکنیکی جیاواز بو دەرخستنی دەقەکان لە داتای نا ڕێکخراو. ئەم لێکولینەوە تێروانینێکی گشتی لە دور شێواز و ئەلگوریزما پێشکەش دەکات بە وانەی کەپەیوەندی بە دەرهێنانی دەقەکانەوە هەیە، وە هەروەها هەندێک لێکولینەوەکان لە دور (Kurdish web documentation). وە لەهەمان کاتدا، کومەلێک کێشەو و میتودۆلیژیای توێژینەوەکان هاوکاری زانایەکان دەبن لە بەدواداچونی توێژینەوەکانی لە ئایندەدا.

مقياس تقييم الارتباط من خلال خوارزميات التنقيب عن النص والتطبيق على اللغة الكردية: مراجعة

الملخص:

أصبح استخراج النص موضوعًا ساخنًا في البحث نظرًا لتوافر العديد من المقالات على الإنترنت في السنوات الأخيرة. يُعرَّف تعدين النص على أنه تقنية تُستخدم لاستخراج معلومات مثيرة للاهتمام أو معرفة من مستندات نصية غير منظمة بشكل عام. أجريت بعض الدراسات باستخدام تقنيات مختلفة لاستخراج النصوص من البيانات غير المنظمة. تقدم هذه الدراسة لمحة عامة عن الطرق والخوارزميات البعيدة في الموضوعات المتعلقة باستخراج النص ، وكذلك بعض الأبحاث عن بعد (توثيق الويب الكردي). في الوقت نفسه ، سيساعد عدد من قضايا ومنهجيات البحث العلماء على متابعة البحث في المستقبل.