Journal of Advanced Computer Science and Technology Research, Vol. 9 No. 4, December 2019, 41 - 59

ARABIC TEXT MINING AND ROUGH SET THEORY FOR DECISION SUPPORT SYSTEM.

Hasanin Alwan Malik Ahmed T. Sadiq
Department of Computer Sciences University of Technology – Iraq.

ISSN (Printed): 2314-7350
ISSN (Online): 2231-8852

Abstract The present work based on an implement classification system for Arab Complaint. The basis of this system is the construction of Decision Support Systems, Text Mining, and Rough Set, a specialized semantic approach of correlation which is referred to as the model of ASN for creating semantic structure between all document's words. A particular semantic weight represents true significance of this term in the text after each feature is probable to be utilized the semantic selection of a suggested feature based on 2 threshold values, the first being the maximum weight in a document and the other one representing the characteristic having the maximum semantic scale with the first one of the thresholds. The task of classification in the stage of testing depends on the dependency degree concept in the raw group theory than before to treat all the characteristics of every one of the classes that results from the stage of training as particular condition rules and therefore, considered in the lower rounding group. The outputs of this classification system are sufficient, performance is quite good persuasive, and assessment of this the system through measurement precision**.**

Keywords: **text-mining, text classification, decision support system.**

1. Introduction

Underwriting is considered as a very critical task for all companies. During years, data maintenance can either be performed in systems of legacy or in hard-copy files for the transactions of underwriting. Systems of legacy were modeled more for the sake of connecting workflow than for harvesting knowledge about transaction of underwriting, which makes the accessibility to data underwriting onerous and manual. With the advancements in the technology, there is a variety of new means for data extraction. With the progress that the data extraction capability has witnessed, there has been a maturing and counter-weighing idea that not every historical data which may be improved is actually related to what one needs or wants (Ryan, 2015).

The basic motive for more efficient performance is growing beyond trivial computations to analytics and business intelligence fields. Typically, underwriting examination factor is combined with data for enhancing the abilities of the predictive analytics. The underwriting process's complexity was enhanced as data increased. A decision support system is needed, and that may play the role of underwriter and it has to be sufficiently capable for making decisions according to trained data (Samaneh, 2013).

Therefore, the era of the underwriting which is intuition based is coming to an end and the role that the underwriting is playing keeps changing exponentially. The growth in the large data amounts started a more predictive environment for the risk management that drives the demands for cultural shifts in the conventional method of the underwriting. Soon, the best underwriting process execution is going to require developing new sets of skills for the sake of meeting the requirements of the progressing model of underwriting (M. Sridevi, 2016).

Text Mining (TM) can be defined as a semi-automatic procedure of the extraction of non-trivial patterns from the unstructured documents. Text Mining presents numerous advantages for the organizations, like understanding the opinions of the customers and the on-line study of the reputation of a brand. TM is vital for the majority of the web sectors due to the fact that the majority of information which is shared on-line is in text format. Which is why, an approach which is capable of structuring text information is necessary for uncovering basic topics and terms which are highlighted in a text (J. Dorri, 1999).

According to Sprague and Carlson, "DSS comprise a class of information system that draws on transaction processing systems and interacts with the other parts of the overall information system to support the decision-making activities of managers and other knowledge workers in organizations"**.**

DSSs are broadly characterized as interactive computer based systems, which are helpful for people in the utilization of data, documents, computer communications, models, and knowledge for solving tasks and making decisions. DSSs are auxiliary or ancillary systems (Arora, 2017).

On the other side, rough set theory play a big role in several machine learning applications, also some researchers used it in text mining applications and obtained good results (Ahmed, 2012; Ayad, 2019; Hasanen, 2019).

In the presented study, the aim is to develop categorization system based on the word's meaning when differentiating information, which can differentiate in every one of them with the other attempting to fix the trouble of noise information.

2. Related Works

Kelwin Fernandes et al. (Kelwn, 2015), has proposed an innovative proactive Intelligent DSS (IDSS) which could analyze articles before publishing them. With the use of a wide range of extracted characteristics (such as, digital media content, key-words, earlier popularity of the news that have been referenced in an article). The Intelligent DSS initially performs a prediction of whether or not an article is going to be popular. After that, it enhances a sub-set of articles properties which may easier to change by the author, searching for an improvement of the predicted probability of popularity.

Mathias Kraus et al. (Mathias, 2017), has studied utilizing DNNs in financial decision support. In addition to that, the paper has experimented with the learning of transfer, where the network has been pre-trained on a different corpus with a length of 139.10 million words. the results of this study have shown better directional precision in comparison with conventional ML in the prediction of the movements of stock prices as a response to the financial disclosures.

Mohammed Alkahtani et al. (AKahtani, 2019), has proposed a DSS which has been based on innovative integrated stepwise approaches, which include self-organizing maps, ontology-

based TM, cost and reliability enhancement to identify faults in manufacturing, mapping those faults to design information and ultimately enhancing the parameters of the design for minimal costs and maximal reliability. The DSS performs the analysis of warranty data-bases that gather information of warranty failures from customers in a text format. For the extraction of the hidden knowledge from this, an ontology-based TM based method has been employed. A DM based method with the use of the SOMs was presented for drawing information from warranty data-base and for relating that information to the data of manufacturing. Clusters that have been obtained by the use of the SOM have been analyzed for the sake of identifying critical regions, in other words, map sections in which maximal defects happen.

Qing Zhu et al. (Qing, 2017) has proposed a court judgment DSS (CJ-DSS) based on the medical TM and the technology of the automatic classification. This study has taken documents of medical damage judgment in China for an example. The system has been capable of predicting the results of the trail of the new documents of the lawsuit based on verdicts of previous cases - rejected and non-rejected. Combined with cases, the research has concluded that the method of the combined feature extraction in fact improves the efficiency of 3 classifier types, which are – the SVMs, ANNs and K-NN, the level of the enhanced efficiency differs from the use of the DF-CHI approach of the combined feature extraction. Moreover, integrated algorithm of learning enhanced the efficiency of the classification of the entire system as well.

Marta Nave et al. (Marta, 2018), focused developing a DSS for assisting tourism managers to improve their offer associated with the explicit on-line behavior of the consumers. This research has used a TM and the method of sentiment analysis for structuring on-line reviews and presenting those results on a DSS with 2 distinct dashboards for assisting in decision-making. This type of systems could be helpful for the managers in developing new visions and strategies that are aligned with the expectations of consumers in a considerably more sustainable and flexible pace.

Aman Dubey et al (Aman, 2018), has proposed technique which has the fundamental aim of reducing amount of time an underwriter takes to find correlations and use his previous experience for the sake of giving the most extensive plan to the client. Briefly, an extensive amount of data has to be obtained and cleaned for features. Those e-mails are obtained as a collection of texts. The application of the concepts of NLP to data like POS tagging, tokenization, and Tf-Idf feature for the determination of word importance. Those features are utilized for model training.

Ahmed T. et al (Ahmed, 2013), suggest a system for social network conversation (benefit for DSS), that works by assigning scores to sentences in the document to be summarized, and using the highest ranking sentences in the summary. Highest ranking values are based on features extracted from the sentence. A linear combination of feature scores is used. In addition to basic summarization, some attempt is made to address the issue of targeting the text at the user. The intended user is considered to have little background knowledge or reading ability. The system helps by simplifying the individual words used in the summary. In the proposed system, classification of keywords by higher ranking of topics has contributed to an active role for the extraction of summarization.

3. Decision Support System

A DSS can be described as a flexible, adaptable, and interactive system which uses models, decision rules, and model base combined with a comprehensive data-base and the own insights of the decision maker, which lead to implementable, specific decisions to solve tasks which wouldn't be amenable to the science models of the management. Which is why, DSSs support complicated decision-making and increase its efficiency (Banerjee, 1995). An application of DSS may include following subsystems as depicted in Figure (1).

1. Model of sub-system of Management: The model base provides the decision makers with the ability for accessing various models and assisting them in the process of the decision making. The model base may be including the model base management software (MBMS) which is responsible for the coordination of using models in the DSSs. This element may be related to the external data storage.
2. Data Management sub-system: The management sub-system of the data-base includes a data-base, containing related data for the case and is regulated with a software known as the data-base management system (DBMS). The data-base management sub-system may be inter-connected with the data warehouse of the corporate, a repository for corporate related decision-making data.
3. User Interface sub-system: referred to as dialog management facility as well, it provides users with the ability of interacting with the DSSs for obtaining the information. This sub-system needs 2 abilities; action language which tells the DSS the requirements and passes data to DSS and the language of presentation which can transfer and present the results of the user. The generators of the DSSs act as buffers between users and the rest of the components of the DSS, which interact with the database, the user interface, the model base.
4. Knowledge-based Management sub-system: which is capable of supporting any other sub-system or acting as a standalone component. It gives intelligence to the augment the decision maker's own. It may be inter-connected with the repository of knowledge of the organization, referred to as the knowledge base of the organization.

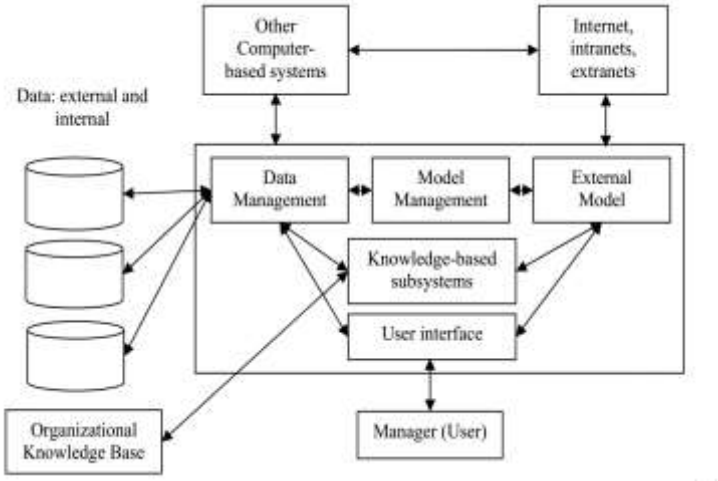

Figure (1): A Scheme of the DSS.

5. TM-Based Decision Support: Methods and Architecture

TM-based DSSs (TMbDSSs) are responsible for the integration of unstructured text data with predictive analytics for providing an medium to arrive at citizen-centric well-informed decisions.

Technologies that have been utilized in the Text Mining comprise information extraction (IE), information retrieval (IR), summarization, topic tracking, concept linkage, categorization, question answering, and information visualization. The most commonly utilized approaches of TM have been discussed in short for the enabling of a more sufficient comprehension of their applications in e-governance, e-democracy, and citizen participation areas (G. Koteswara, 2011).

1. Categorization: this process is involved with the identification of the fundamental document themes via placing it in a predefined group of subjects. It doesn't try processing the information itself as does the information extraction. This process only performs counting of words appearing in text and, from counts, characterizes the basic subjects which are covered by the document. This process is typically dependent on thesaurus for which topics are pre-defined, and correlations are specified via searching for narrower terms, broader terms, related terms, and synonyms.
2. Information extraction: those algorithms specify relationships and key phrases in the text. Which is performed via searching for specified sequences in the text, with the use of a process which is referred to as "pattern matching". Algorithms conclude the correlations amongst all sequences that have been identified for providing users with relevant insights. This technology may be quite beneficial in the case of dealing with massive text volumes.
3. Topic tracking: this type of systems operates via maintaining profiles of user, according to the documents that are viewed by the users, makes prediction of other documents which might be relevant to the user. Some of more efficient tools of text mining permit users in selecting specific categories of interest, and can even automatically infer the user's interests based on their click-through information and reading history.
4. Association detection: In the Rules of Association, the emphasis is made on researching the correlations and implications amongst the subjects, or descriptive concepts that are utilized for the characterization of a group of associated texts. The aim is discovering valuable rules of association in a corpus in a way that the existence of a group of topics in an article implies the existence of another topic.
5. Clustering: this approach is utilized for grouping documents that have similarity, however, it is different from categorization in the fact that the clustering of documents is performed according to similarity to one another rather than via using pre-defined subjects. A fundamental algorithm of clustering produces a topic vector for every one of the documents and performs measuring of how sufficiently the document fits to every one of the clusters.
6. Question answering: this application is concerned with the way of finding the optimal answer to a certain question. This application might use a combination of text mining approaches.
7. Text Summarization: is greatly helpful for the attempt to understand if or not a long document satisfies the requirements of users and deserved to be read for additional information. The base of summarization is the reduction of the detail and length of a document at the same time as keeping its basic points and general meaning.

Figure 2 includes an illustration of the fundamental for Text-Mining based DSS for e-governance technical architecture.

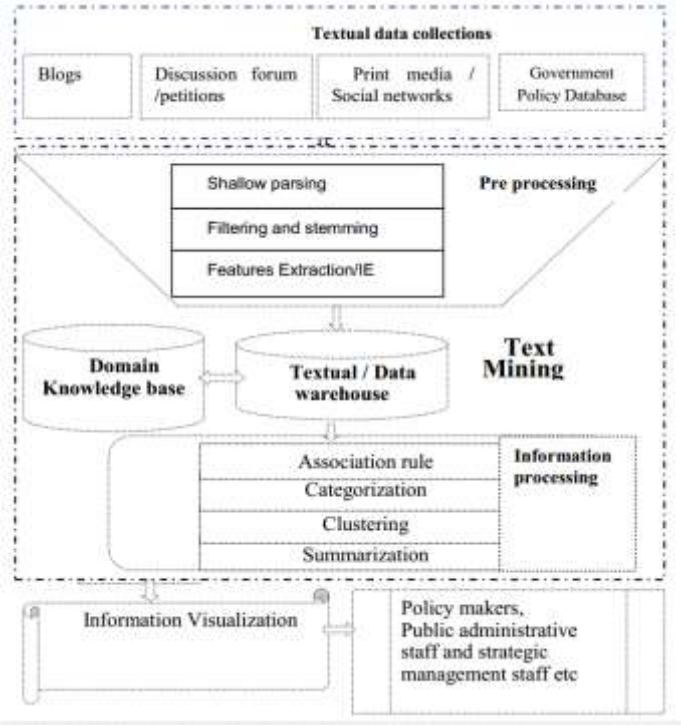

Figure 2: TM-Based DSS for E-Government: Technical Architecture.

6. The Suggested System Architecture

The suggested system had been develop particularly for classifying Citizens' requests written in Arabic. Such a system is adopting the ML concepts with the use of prior knowledge-based as the foundation for the Citizens' request categorization. Such prior knowledge had been acquire from a set related to pre-categorization Arabic Citizens' requests. Majorly, the suggested system consists of 2 major stages, the initial one is the training phase that has the task to learn classifier and provide it with experience as well as to create prior knowledge related to each one of the categories via the analysis of the set related to the labeled Citizens' requests. The other phase is the testing phase, which tests and evaluated the system's performance via input set of un-labeled Complaints of citizens in which the system act on allocating correct categories.

With regard to Training phase, Complaints of citizens will be passé on many steps that work on text and discovering properties, which differentiate between the categories. The set of pre-categorized citizen's requests will be applied based on six classes related to the work of local governments (Water Directorate**,** Directorate of Sewage**,** Municipality Directorate**,** Directorate of Electricity**,** Directorate of Education**,** Directorate of Health**).** Each one of the articles will pass through some steps such as pre-processing, feature extraction, feature selection and the use of training classifiers. Such stages had been define thoroughly as follows: as follows:

1. **Pre-processing :**With regard to this step, the partial stripping regarding texts from all the numbers and symbols as well as to remove words, which does not have meaning, also does unification form for all words through eliminating linguistic additions like prefixes

and suffixes as well as punctuations. Pre-processing consists of 4 sub-steps detailed in the following way:

A. Tokenization

This stage has the aim of identifying as well as distinguishing each one of the words in text via splitting text and then convert it to separate words according to space separators between these words, also the performance of this step has improved by applying important addition of removing everything in text which does not belong to letters or words, and such improvement resulted in eliminating all symbols and numbers from text with no requirement for pre-list consists of symbols that should be removed. The word's list, which is the result of such process, is referred to as (Tokens).

B. Normalization

The is one of the important steps which presented all text's words in a unified format, a feature in the Arabic language must be considered in this task specified via Arabic diacritics (حركات التشكيل) added on letters in a way that they alter the word's form and badly impact the unification process. Some samples related to normalization can be seen in table (1).

Table (1) : Samples of word Normalization

Tokens	Normalization
بَلَدِيَّة	بلدية
مَدرَسَة	مدرسة
مَرِيضُ	مريض
المِيَاهُ	المياه
ألتيار	التيار
طَفح	طفح

C. Stop Words Removing

This stage comes after the normalization step, it has the aim of reducing the text's feature dimensions through removing common words that are frequently used and at the same time not having suitable meaning with regard to text. The Arabic language has various stop words like vowels, prepositions, and 5 names. Names related to seasons, months, and days of the year have been added. The list will be created for all the above-mentioned words, consisting of over (550) Arabic stop words. Samples regarding such stopwords are the use of system as can be seen in the table 2).

Table (2): Samples related to Standard Stop-words in Arabic

لات	ليس	لن	ان	في	على	الى	من
مازال	ما	كأن	كان	ذو	فو	ليت	لعل
هكذا	تلك	أولئك	هؤلاء	الذين	اللاتي	مابرح	مابقي

لما	لم	كيف	امسى	اصبح	صار	كيفما	فيما
عليهما	عليه	عليهم	انتن	انتما	انت	له	لهم
متى	بعد	يكون	كما	كالتالي	كالاتي	هي	هذا
ايهما	ايهم	أي	أينما	لاسيما	مع	نحو	حيثما

D. Suggested Arabic Light Stemmer

This could be consider as a step of high importance in the pre-processing. It odes remove suffixes, prefixes, and linguistic additions, and therefore acquiring word's original root. In such an approach, all the words which had been obtain from single-origin are going to be returned to original and therefore one form will be used for representing them. The use of stemming results in saving memory and time, which will increase the performance. The Arabic language holds extreme privacy with regard to stemming. Since it is widely utilized language with huge synonyms as well as vocabulary, it also consists of various grammar rules, and that require distinctive rules for dealing with additions occurring on word suffixes and prefixes. An approach was suggested for the Arabic Light stemmer (ALS), that is defined as several conditions determining the way of applying to stem on a specific word or not, the rules related to eliminating prefixes and suffices in Arabic Light Stemmer are defined in the following way: (Condition) S1 $\rightarrow$ S2

In the case of achieving a conditions with the suffix S1, S2 will replace S1 as word with no suffix. For instance, the suffixes used in the words of Arabic language like: (" ," نفاياتهم، النفايات، بالنفايات) are going to be eliminated for returning the word's root) .نفايات(Another example regarding prefixes in the Arabic language words is: (العمل، بالعمل، فالعمل، كالعمل) the prefixes which appear in each word's beginning will also be deleted for getting the word's original root (عمل).

There are 2 cases suggested to apply the rule used to delete suffixes and precedents in the following way:

- The first one: Word's length should be over = 4 letters as follows:

  Len(word) >= 4

  The major aim of such case is that the maximal length which is related to the suffixes and prefixes in Arabic language includes three characters as can be seen in the table3.3, so that words which have length not more than four and suffixes or precedents will be eliminated, after that single letter will remain or none, in such condition the world will be with no meaning, thus, there will be lack of word's root.

- The other case: Following applying the initial case, in the case when the rest of the word is considered to be in the list of suffixes and prefixes, the rule is going to be undone.

The major aim of the second case is ensuring that the word that remains following applying the initial case still specifies certain meanings and not in one characters or words in the list of suffixes and precedents in addition to the stop words' list.

Table (3): Samples Prefixes and Suffixes in Arabic

Prefixes	Suffixes
"ال"، "لل"،"وال"،"بال"،"فال"،"الأ"،"فا"،"لأ"،"بأ"،"با"	"ان"،"ون"،"كم"،"كن"،"هم"،"هما"،"ها"،"ين"،"تن"،"تم"،"ات"، "يه"،"ية"،"كم"،"نا"،"هن"

Table 4 show the output related to the suggested Arabic light stemming.

(Table 4): Samples words and their stems that are applied in the suggested Arabic light stemming.

Word	Stem	Word	Stem
ممرضين	ممرض	المركزية	مركز
متجاوزين	متجاوز	بالكابسات	كابسات
قاموا	قام	المريض	مريض
النفايات	نفايات	معلمين	معلم

Figure 3 will show diagram related to steps of pre-processing.

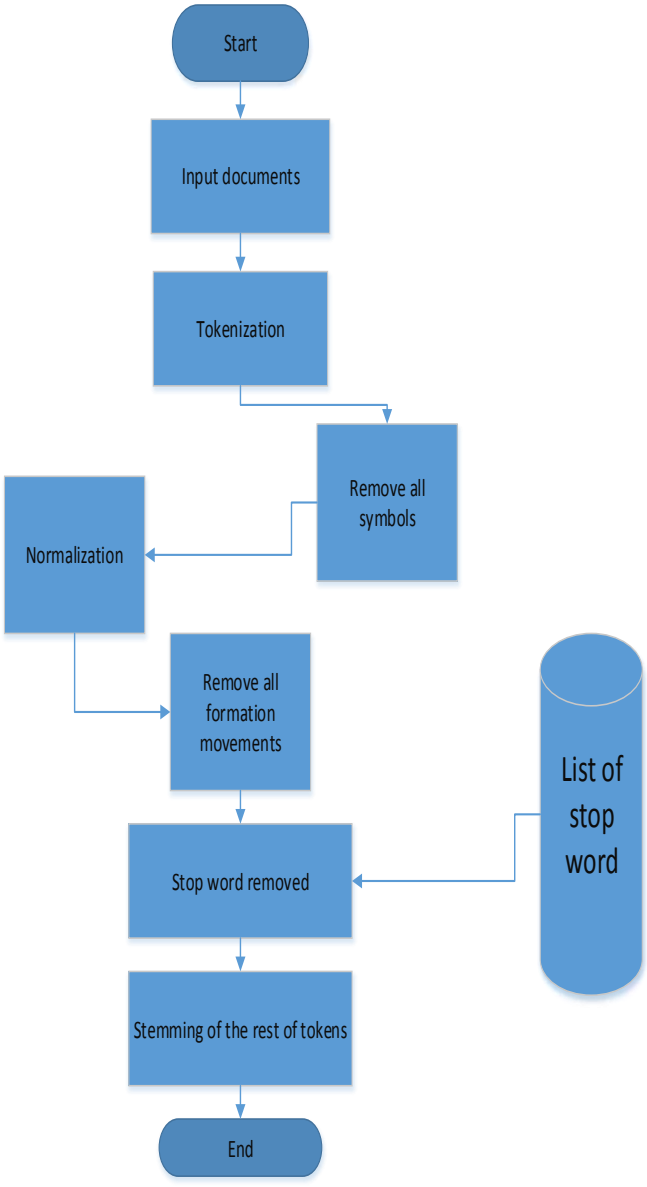

Fig. (3): Block diagram related to pre-processing steps.

2. **Feature Extraction:** this stage is very important in the classification of documents following the pre-processing phase for creating features that can be maintained in the following steps, namely, feature selection with the use of Vector Space Model or BoW Model. These phases are defining the feature extraction algorithm:

A. Creation of Vocabulary

Vocabulary in the presented system is implemented as a list that consists of all the required information related to document's features in training stage, in which each one of the features includes 3 elements (frq, string, key) in document's vocabulary list. The Key indicates the sequence related to feature location in document's features, and the string indicates features, while frq indicates the number of frequency regarding each one of the document's features.

For instance:

D :" عدم مرور الكابسات الخاصة بلنفايات لخمسة ايام ادى الى تراكم النفايات في الحي وتحويلها الى مكب "

Vocabulary related to above document will be constructed in the following way:

(1:الكابسات 1:2، بالنفايات:1، النفايات :1، 3: مكب 1:)

B. Documents Representation

Each one of the documents in the training phase will be converted from full texts to vector documents, specified via certain feature's array. Thus, the set related to the text documents is going to be a matrix with a single row for each one of the documents and a single column for each one of the features, which happens in document. Each one of the features is related to sequences in certain documents also with certain weight features which is represented via the number of the feature's incidences in documents which indicate the significance related to such features.

3. **Features Extraction (FS):** is carried out in the system's training stage with the use of BoW model, such model have 2 major phases: build vocabulary as feature's list and represent each one of the documents as 1D feature's array. Each one of the elements in array have feature with the index, also this feature's weight is obtained from term frequency weighting technique (T-F).

Suggested Features Selection (TISF) Algorithm show the performance related to features selection stage with the use of the suggested (TISF) approach, such suggested approach use the semantic relatedness degree between document's words that product from the opinion model for building a list related to the semantic features which is related to each one of the classes in the semantic basis, such approach makes 2 systems regarding Rough Set classification model as well as opinion model for totally overlapping for creating Proposed System of classification which is on the basis of semantic method. There are 2 thresholds which are allocated to the selection of the features, the first one = maximum feature weight in the model, while the other one is equal to maximum semantic degrees between highest feature weight and all document's features, also with using the model.

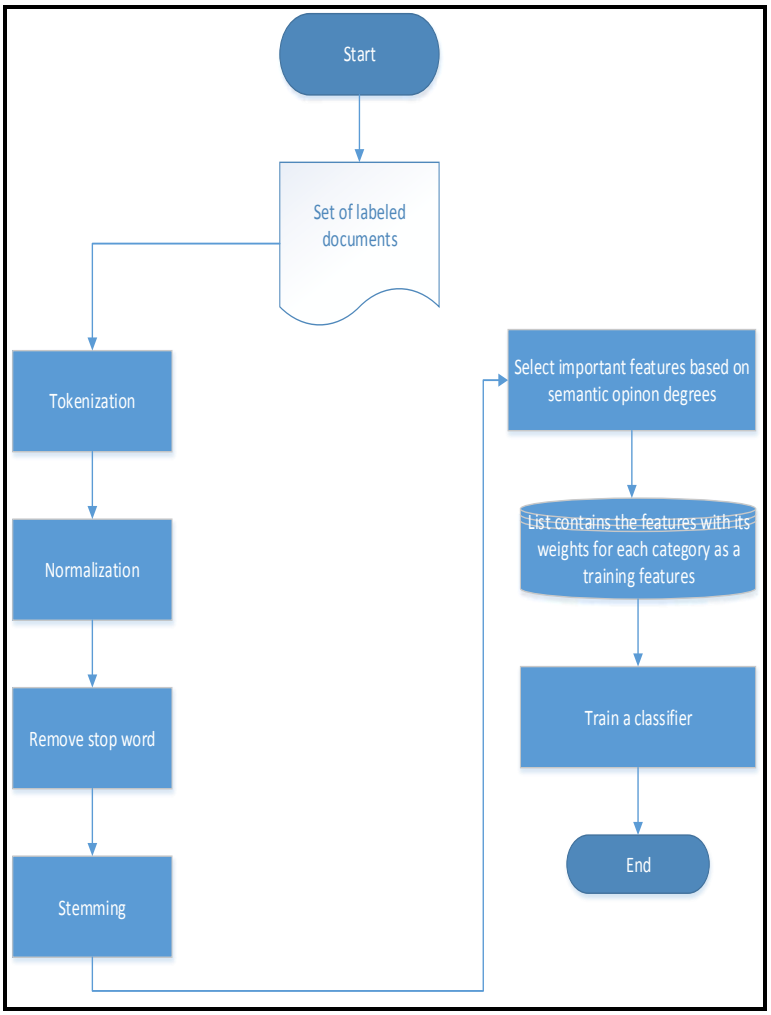

Figure (4): Block diagram related to proposed system training step.

- **The Suggested Rough Set Classifier algorithm**

The suggested rough set classifier is carried out with the use of its 2 stages (testing and training). The initial stage is to illustrate the training phase on the suggested approach of rough set training model, with regard to such stage, list related to selection features for each one of the classes (string with its weights) will be considered as input to classifier for the purpose of training rough set classifier, the major significance related to the rough set in such stage is in the use of rough set attributes reduction for reducing the feature lists' dimensions, following the training regarding features of each one of the certain classes, testing stage will be assigning predict class regarding each one of the input documents, such system's phase is carried out according to dependency degree between certain classes and the features.

With regard to the testing stage, the requests that are un-labeled will be used as alternatives to labeled document articles in the training stage. The major significant factor in the testing aspect related to the system of the Proposed system, is selecting an algorithm of classification that is obtained from major laws regarding RST. It is considered to be effective and straightforward, also it is extremely fast, thus it takes less time and less memory space since such approach resulted in optimum results in articles' classification in real categories to which they really belong.

Proposed system testing stage includes 2 major phases: the first one is the pre-processing stage for preparing the text inadequate form, while the second one is the classifying stage for obtaining the system's ultimate results, the upcoming 2 points will show the training stage's performance steps in the system of the proposed system.

1. Preprocessing Stage

The pre-processing of training stage consists of tokenization process for splitting each one of the words from the others, also it depends on spaces between them, also the tokenization will take just the Arabic words, the normalization process unifies the words through re-writing words with no formation movements, removing stop words in addition to applying the suggested Arabic light stemming. This stage's output is considered a set of tokens regarding each one of the inserted documents.

2. Categorizing Step

In the case when the intentions is dealing with or classifying it one the basis of RST required looking into text from rough set's perspective, which handle documents in the following way:

A. Handling document features as information system:

1. The training stage's labeled set can be defined as a set related to the objects in the information system.
2. With regard to rough sets, each one of the objects has set related to the attributes which are divided t decisions attributes and condition attributes, such object will be labeled document which will be inserted in the training stage, conditions attributes indicates words related to document, and decision attribute specified via class which labeled document belong to.
3. Based on what was indicated above, features in the ultimate features list regarding the train will be classified as conditions attributes in low approximation concepts, also the document's features that are not indicated in the list are defined to be in the upper approximation concept.
4. As stated by point 3, each one of the lists which are providing certain class represents low approximation set related to the conditions attributes which are conducted on class certainly.

B. Suggested Rough Set Document Classifier According to Dependency Degree

1. Applying the pre-processing stage on testing un-labeled documents compute how many features match with features in each one of the lists regarding the specific class. In which training classifier includes all feature lists, each of them specify certain class.
2. On the basis of the rough set's dependency degree estimated with the use of ratio between test document's features and each class's features with the use of these formulations which are created from the equation (1).

$$DX(A) = \frac{FN \times \sum FW}{N} \quad (1)$$

In which DX(A) represents dependency degree between class A and the input document X. N indicates the number regarding common features between class A and document X. FW indicates summation regarding the common feature weights. N will be representing total number related to features in the document X.

3. Following estimating dependency degree between input document and all classes, class's candidacy which hold highest dependency degree with input document representing document's category.

7. Result and discussion

a. Preprocessing results

There are some samples which results from utilizing pre-processing phase on each one of the categories in Arabic data-set, the pre-processing stages are tokenization, normalization, remove stop words as well as stemmer, the tokenization process split the document's features, the normalization process have the task of unifying the form related to the features, then to remove all the Arabic stop words depend on previous arranged list and lastly applying Arabic light stemmer to remove all feature's additions as well as getting original root of words. Sample results related to the pre-processing phase throughout the use on all categories from data-set.

Table (5) : Sample of Words Preprocessing Result

Tokenization Results	Normalization Results	Removing Arabic Stopwords	Results of Arabic light stemmer stage
أولاً	اولا	_	
:	_		
إن	ان	_	
البلدية	البلدية	البلدية	بلدية
بحاجة	بحاجه	بحاجه	بحاجه
إلى	الى	_	
المياه	المياه	المياه	ماء
الجديدة	الجديدة	الجديدة	جديدة
…	_		
الطلب	الطلب	الطلب	طلب
في	في	_	
هذا	هذا	_	
الموضوع	الموضوع	الموضوع	موضوع
التجاوز	التجاوز	التجاوز	تجاوز
على	على	_	
بعِض	بعض	_	

النفايات	النفايات	النفايات	نفايات
المُتعَلِقة	المتعلقه	المتعلقه	متعلقه
بهِ	به	_	

b. Features Extraction Results

The presented section will be provided to clarify the results related to extracting features, all the document's features will be subject to extraction process via using certain model which can be considered as vector space model that is also referred to as (bag of words) model, such model work on creating special features array for each one of the documents, such matrix is considered as 1D array, each one of the extracted features is specified via given item which belong to this document's array, each one of the items consists of significant information regarding such feature as follows: feature from after utilizing pre-processing, as this feature's weight which is estimated depend on term frequency weighting TF, also feature's index which specify feature's order in document, the step's results are of high importance for achieving the following phases in our system, each one of these tables (4.3, 4.4, 4.5, 4.6) will display acquired results following applying feature extraction on specific category from 6 categories in Arabic Citizen request data-set.

Table (6): Samples results related to feature extraction on request of six categories from data-set.

H 423	E 504	M 355	W 233	S 201	E 365
المصاب36	مدرس118	بلدية33	انبوب29	مجاري45	الكهرباء204
المريض81	مدرسة93	متجاوزين12	ماء89	طفح59	التيار21
مستشفى54	مدرسين32	تجاوز21	كسر45	تصريف81	ضعف6
صحية77	معلمين11	عرصه11	مجمع9	اسنه22	انقطاع98
صحي13	معلمات4	مشاع67	اسالة53		الخصخصة4
لقاح52	تربية75	مصفره77			
طبيب59	مدرسة82	ازالة32			
ممرض8	ناجح8	تبليط90			
اسعاف16	راسب66	نفايات140			
علاج79	مكمل75	حاوية20			
مصاب92	معلم54	الازبال63			
راقد5	معلمة21	الطمر27			
صيدلية110	متوسطة13	مساطحة9			
	اعدادية87	انقاض86			
	الابتدائية65				
	روضة2				

c. Proposed of features selection results

The present phase that will be shown in this section has the task of reducing the document's dimensions via just selecting the significant features from a lot of document's features, such selection of features depends on applying two important semantic features

(TISF), such system take feature's list with its weights for each one of the documents that is acquired from features extraction phase and assign 2 thresholds, the first one is considered to be equal to list's highest weight, while the other specify feature which has the largest semantic degree with features within the first one. Such approach add significant characteristics to enhancing the classification process via relying on deep meaning as well as the association between features to candidate significant features for representing documents rather than being totally dependent on feature's frequency in documents. The outputs related to such phase can be considered as list of chosen features related to its weights for each one of the system's categories.

Table (7): Sample results related to the suggested (TISF) phase on the citizen request regarding Health directorate category from data-set.

Features (Words)	Weights
المصاب	2.83600
المريض	2.90631
مستشفى	2.504125
صحية	3.36540
صحي	2.44301
لقاح	2.569782
طبيب	3.95603
ممرض	2.58632
اسعاف	2.64734
علاج	2.61532
مصاب	3.66931
راقد	4.60745
صيدلية	2.90711

Table (8): Sample results related to the suggested (TISF) phase on the citizen request regarding Education directorate category from data-set.

Features (Words)	Weights
مدرس	2.64734
مدرسة	2.61532
مدرسين	3.66931
معلمين	4.60745
معلمات	2.90711
تربية	3.31194
مدرسة	3.74367
ناجح	1.71153
راسب	2.54734
مكمل	1.61532
معلم	2.65431
معلمة	2.80945

متوسطة	2.90711
اعدادية	3.11194
الابتدائية	3.9967
روضة	2.7214

Table (9): Sample results related to the suggested (TISF) phase on the citizen request regarding Municipality directorate category from data-set.

Features (Words)	Weights
بلدية	2.64734
متجاوزين	2.61532
تجاوز	3.66931
عرصه	4.60745
مشاع	2.90711
مصفره	3.31194
ازالة	3.74367
تبليط	3.11153
نفايات	3.54332
حاوية	2.66931
الازبال	4.2345
الطمر	1.90011
مساطحة	1.19194
انقاض	3.99367

Table (10): Sample results related to the suggested (TISF) phase on the citizen request regarding the Water directorate category from data-set.

Features (Words)	Weights
انبوب	4.68926
ماء	2.66931
كسر	3.61532
مجمع	2.61143
اسالة	2.90711

Table (11): Sample results related to the suggested (TISF) phase on the citizen request regarding the Sewerage directorate category from data-set.

Features (Words)	Weights
مجاري	2.31277
طفح	3.03759

تصريف	1.73903
اساله	1.30695

Table (12): Sample results related to the suggested (TISF) phase on the citizen request regarding the Electricity directorate category from data-set.

Selected features	Weights
الكهرباء	4.21326
التيار	2.56931
ضعف	1.61532
انقطاع	4.62221
الخصخصة	1.90711

d. Classification Results

This section shows the final step of the proposed system, this stage represents the results which are related to the testing stage that has the aim of classifying un-labeled citizen requests, the suggested classification in the presented section is on the basis of RST, it deals with texts as information system to find certainly rules which achieve target category, each one of text's features indicates the condition feature that must exist to achieve the decision which specify given class. With regard to the process of classification dependency degree related to each one of the condition features has been computed in the text for determining association degree between such features and specific certain category. Each one of the document inputs to system in test stage is going to be subjected to pre-processing stage prior to the initiation of the suggested rough set classifier procedure. The table below, will show samples related to the results acquired from the phase of classification.

Table (13): indicated the samples that are related to the results that output testing classifier

Documents	Real Category	Proposed Decision	Evaluation
D21	"مد انابيب المياه"	"مد انابيب المياه"	True
D193	"النفايات الصلبة"	"النفايات الصلبة"	True
D45	"انقطاع التيار الكهربائي"	"انقطاع التيار الكهربائي"	True
D201	"المراكز الصحية"	"المراكز الصحية"	True
D15	"التجاوز على اراضي المجاري"	"مديرية البلدية"	False
D220	"العلاقات الدولية"	"العلاقات الدولية"	True
D76	"مديرية البلدية"	"مديرية البلدية"	True

D5	”المركز الامتحاني“	”المركز الامتحاني“	True

8. Conclusions

Complaint system is an electronic complaint system that enables the citizens to present and follow up on their complaints and requests submitted to the government where the focus was on the service directorates in the target provinces such as (water, municipal, sewage, electricity, health, education). The system is a two-way communication tool by which the complainer or the requester can track the process of his/her complaint or application submitted to one or more of directorates. The DSS in this system benefits and assist greatly in accelerating citizens' requests and making their voices heard by the concerned government agencies which participated in, improve the performance and the services. Use RST has been very successful in classifying citizens' requests to the government through using the dual classification we have adopted, which is to determine where the citizens' requests will be directed to any services directorate and the second is the type of citizens' requests. Using Arabic light stemmer in the system was of benefit as it provided speed and efficiency in the results of requests submitted by citizens to the government. The requests of citizens to the government was found by a large difference between the target provinces (Baghdad, Basrah, Kirkuk, Diyala, Najaf and Karbala) the result of differences in expression or difference in the cultural and social level in those provinces and this led us to choose more than one province to represent the full diversity of all Iraqi provinces, and at the time that we found that applications in the province One is somewhat similar. Citizens' request were found with very different in Diyala province due to high different in feature.

Reference

Ahmed T. Sadiq & Sura Mahmood Abdullah, (2012). Hybrid Intelligent Techniques for Text Categorization, International Conference on Advanced Computer Science Applications and Technologies (ACSAT), IEEE, Malaysia.

Ahmed T. Sadiq, Yossra H. Ali & Mohammad Natiq Fadhil, (2013). Text Summarization for Social Network Conversation, International Conference on Advanced Computer Science Applications and Technologies (ACSAT), IEEE, Malaysia.

Alkahtani, M., Choudhary, A., De, A. & Harding, J., (2018). A Decision Support System based on Ontology and Data mining to Improve Design using Warranty Data, Computers & Industrial Engineering, Loughborough's Research Repository.

Aman Dubey, (2018). Smart Underwriting System: An Intelligent Decision Support System for Insurance Approval & Risk Assessment", 3rd International Conference for Convergence in Technology (I2CT), The Gateway Hotel, XION Complex, Wakad Road, Pune, India. Apr 06-08.

Arora A., (2017). Management Information Systems, Himalaya Publishing House, Mumbai, India.

Ayad R. Abbas Dhafer Hamed Abd & Ahmed T. Sadiq, (2019). A New framework for Automatic Extraction Polarity and Target of Articles, Journal AUR Revista, Vol. 26, No. 2, pp:358-367.

Banerjee, U. K., & Sachdeva, R. K. (1995). Management Information System: A new frame work. New Delhi: Vikas Publishing House.

G. Koteswara Rao, (2011). Decision Support For E-Governance: A Text Mining Approach", International Journal of Managing Information Technology (IJMIT) Vol.3, No.3.

Hasanen S. Abdullah, Saif Ali Abd Alradha & Ahmed T. Sadiq (2019). English Poems Categorization using Text Mining and Rough Set Theory, Bulletin of Electrical Engineering and Informatics Journal, Vol. 9, No. 4.

J. Dorre, P. Gerstl & R. Seiffert, (1999). Text mining: Finding nuggets in mountains of textual data", in: KDD-99, San Diego, CA, ACM, pp. 398–401.

Kelwin Fernandes, (2015). A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News, Universidade do Minho, Portugal.

M. Sridevi & Arunkumar B.R., (2016). Information Extraction from Clinical Text using NLP and Machine Learning: Issues and Opportunities, National Conference on "Recent Trends in Information Technology" (NCRTIT), International Journal of Computer Applications (0975 – 8887).

Marta Nave, Paulo Rita & João Guerreiro, (2018). A decision support system framework to track consumer sentiments in social media, Journal of Hospitality Marketing & Management.

Mathias Kraus, (2017). Decision support from financial disclosures with deep neural networks and transfer learning, Decision Support Systems, Vol. 104, pp. 38-48.

Qing Zhu,(2017). Court Judgment Decision Support System Based on Medical Text Mining, Association for Information Systems, AIS Electronic Library.

Ryan Cobb, Sahil Puri & Daisy Wang, (2013). Knowledge Extraction and Outcome Prediction using Medical Notes, Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA.

Samaneh Abbasi, (2013). Aspect-based Opinion Mining in Online Reviews, Doctor Thesis, school of computer science, Simon Fraser University, Canada.