Hierarchical Method for Automated Text Documents Classification

Mohamed H. Mousa
University of Jeddah, College of Computer Science and Engineering, Department of Computer Science and Artificial Intelligence
Jeddah, Saudi Arabia
mhmousa@uj.edu.sa

Ayman E. Khedr
University of Jeddah, College of Computing and Information Technology at Khulais
Department of Information Systems
Jeddah, Saudi Arabia
aeelsayed@uj.edu.sa

Amira M. Idrees
Faculty of Computers and Information Technology
Future University in Egypt
Cairo, Egypt
amira.mohamed@fue.edu.eg

Abstract: *Digitalization is currently not a concept the world seeks to apply; rather, it is a fact this world lives in. The transformation for the green world has strongly introduced the principle of eliminating hard copy resources while maintaining their digital versions. The immense amount of information that resides in electronic documents opened a wide road for research a long time ago. On the other hand, information extraction, text mining, and Natural Language Processing (NLP) are three concatenated fields that have gained their unique place in the digital world through time. This research aims to introduce a novel method for Arabic document classification. The research provides multi-tagging to the document according to a set of criteria, one of these tags is the hierarchical classification for the document that could play an efficient role in its related field. For example, documents in healthcare systems beehive could lead to exploring a new symptom of a disease, as it is known that symptoms could continuously mutate over time. The proposed method succeeds through the generated schema to relate between old and new symptoms, which makes it no surprise when evolving and gives a chance for pre-preparation and success to containment. The technical challenges of this study include the ability to successfully apply text mining techniques and machine learning. Additionally, the higher level of challenges that arise in this study is the fact that the processing is applied to Arabic text documents. Arabic has been known to be a complex language as it has its unique nature. The proposed method has been applied, compared with known methods, and its effectiveness has been confirmed by applying a classification task with an Accuracy equal to 99.5%.*

Keywords: *Classification, knowledge discovery, text mining, feature selection, TF-IDF, natural language processing, hierarchical based.*

Received March 10, 2024; accepted November 5, 2024
https://doi.org/10.34028/iajit/22/1/2

1. Introduction

Many online repositories have been developed due to the immense amount of data that emerges in a daily basis on the Internet. As most of the online data is text (unstructured), this situation highlighted the strong demand for classification methods for reliable text classification. This need has been highlighted with the fact that most of the unstructured online data is valuable and could lead to rich information. Machine learning (ML) field has contributed with a high level of success towards a timely reliable and accurate processing of text data with putting a glance that most of the machine learning algorithms could deal with massive amount of data [15].

Text documents classification is one of the Natural Language Processing (NLP) field tasks. Text classification focuses on investigating the documents text and apply a representative label for this document based on its content. As this task is clearly exhausting, many automated methods have been proposed to perform this task [17]. Automated methods lead to more simplification for the classification task with maintaining the high level of accuracy [20]. In addition, automated text documents classification methods provide a standardized model for the task which leads to the possibility to process multiples of data in size with specific requirements and trusted steps. On the other hand, manual text documents classification by the interference of experts is not as efficient when dealing with the continuously increasing amount of documents. Different objectives have been targeted such as language detection [33], credibility detection, sentimental analysis, and others. These objectives confirmed the immense need for automated text processing methods in general and text documents classification in specific [34].

Data transformation from unstructured to structured data is currently an essential objective to all business fields. It enable business leaders to explore the interesting patterns in text. Machine learning techniques successfully contribute to change the perspective of unstructured apparently unrelated data into high structures related and interesting patterns of data. While text processing is challenging, the challenge is more sophisticated for Arabic language. Focusing on Arabic language, it has been reported that it is the fourth

popular language on the Internet with more than twenty two thousand Internet users which represented about five percent of the total Internet users in 2019 while the growth rate is reported to be about nine percent over the last twenty years [34]. Although many research have focused on Arabic as Arabic morphological derivation is one of the reasons for the high language complexity, however, there is still a need for more work in the field targeting to reach a reliable dependable result. Arabic text language processing highlights many challenges [38, 40]. One challenge is the large size of the vocabulary set members. The different form of the same letter is another challenge in Arabic. For example the letter ALEF "أ" could be written as "ا, ٱ, آ, أ, …etc.,". Moreover, the grammar rules of Arabic is related to different conditions including the gender, singular and plural forms, and others. Additionally, the Arabic sentence structure has different forms to be verbal form or nominal [39].

The current research proposed a novel model for text classification. The proposed model is based on a hierarchical approach that applies natural language techniques and machine learning algorithms. The proposed research proposes an adapted method for NLP that targets more accurate exploration to the text keywords which are considered the pillar of the classification task. The text features are extracted by adapting the Term Frequency (TF) and the Inverse Document Frequency (IDF) for each term to determine the TF/IDF method by exploring the semantic relationships of the text terms and adapt the weighting scheme with considering the joint weight of the semantic relation of the term with TF/IDF weighting. Each document is represented in a vector format, features are detected. Synonyms, repetition, common features, irrelevant features, and other vital information are explored to be able to reach the required classification results with the highest performance. Different machine learning algorithms contributes in the current research targeting to determine the most suitable algorithm and ensure the highest performance. The following contribution could be highlighted for the current research.

- An adapted method is proposed following TF/IDF for higher accurate features determination.
- Highlighting the positive contribution of the interdependence of the semantic relationships between the text terms and weighting measures, specifically TF/IDF, in more accurate exploration.
- Proposing a novel method for extracting semantic relationships between the text terms which depends on the siblings' communication nature.
- The contribution of a set of nine classical and deep learning classification algorithms that vary in nature with performing extensive evaluation.

The remaining of the research demonstrates the related work in section 2, the proposed method in section 3, the experiment and evaluation in section 4, and finally the conclusion.

2. Related Work

Classification task has been previously applied for different tasks [8, 44] while text classification in specific has also been applied for many targets besides document classification in [13] such as Query answering [14, 30]. Text classification has proved its effect in different fields such as the educational system adaptation [25, 36], credibility [19, 21, 26, 42, 43], and even contributing in rules based systems [18].

Different research have discussed the previously proposed models for English text categorization [2]. More work has been demonstrated in [11, 43] for Arabic language. On the other hand, different languages have been highlighted in other research [24]. The research [10] applied text mining techniques on Portuguese language dataset with a precision 72.7%. The research highlighted the need for more computational to the Portuguese language to identify the semantic aspects as well as the syntactic aspects of the language in order to be able for processing. Focusing on Arabic, different research have been performed for Arabic corpus enrichment [6]. This focus has been highlighted as a vital research and was always one of the priority branches [21]. Othman *et al*. [32] highlighted the lack of Arabic text data sources that could be utilized for Text classification. The researchers reached this conclusion as most of the datasets either have no defined classes [29] or the defined classes are misleading. Accordingly, the work in the research [35] focused on preparing a dataset of about thirteen thousand documents but it could be considered small for classification tasks. This focus is also tackled in [16] but with a need for more classes' extension.

Sabri *et al*. [37] presented a comparative study of six classification algorithms over an Arabic News dataset collected from Aljazeera website. The authors used the same parameters and concluded that Naïve Bayes (NB) has the highest performance score. This conclusion is also reached Wang *et al*. [41] with a considerably low performance equal to 80.41%. Other researchers such as Jasti *et al*. [22] highlighted the positive contribution of the Feature Selection (FS) task. The research [1] applied the n-gram selection with the combination of K-Nearest Neighbor (KNN) algorithm while the research [5] applied four classification algorithms with two FS methods on an Arabic text dataset that is extracted from BBC website. Although the effort for these research is clear, however, the reached performance highlighted the need to focus for raising the classification accuracy results. Different research applied different classification techniques either traditional techniques [3, 31] or deep learning [9, 12]. It is a common objective for all researchers to seek for higher performance in the applied task, therefore, the researchers seek to shed the

light for the better technique. Boukil *et al*. [7] presented the advancement of Neural Networks (NN) over Logistic Regression (LR) while Alzubi *et al*. [4] presented the advancement of applying FS before NN. Highlighting the contribution of TF/IDF [27, 28], as it is the focus of the current research, a research [27] utilized it for tagging the questions of the students' exams. In addition, NB is applied for the questions classification task. The research reached a precision and recall in the eighties which is considerably low performance. Another research [28] proposed a modification for TF/IDF with a performance no more than 89.7% which is less than the performance reached by the current research based on the proposed adaptation.

Although the previously presented research reached reasonable performance, however, to the best of our knowledge, there was no work that highlighted the benefit of merging different perspectives until this research. This current research does not only perform the highest performance well known FS technique, it also performed a study to highlight the most suitable classification algorithms. Moreover, it highlighted the strong contribution of merging the semantic relationship with the FS technique for higher performance. Moreover, the current research argues that the proposed merge does not need any additional resources, the semantic relationship is extracted through applying one of the most natural relation concept which is the siblings' relation. In the field of text processing, one of the most vital aspects is the data. Therefore, the current research argues that the proposed semantic relation provides higher quality on data which consequently raise the performance of the classification algorithms.

3. The Proposed Hierarchical Method for Automated Text Documents Classification

Utilizing machine learning techniques in NLP requires preparing the data to a uniform that these techniques could process. Therefore, Vectorizers are used to transform the text which is considered categorical features into numerical vectors which could be processed by the machine learning techniques. First, the words conforming the text are considered the categorical features, it is transformed into a corresponding numeric value. Then the phrases in text which are a sequence of words is then transformed into a vector of numbers. Each phrase has a corresponding vector.

3.1. Text Vectorization

Although there are different techniques that contribute in the process of the numeric vectors retrieval [28]. The most well-known methods for this task are Word-Count (WC) and TF/IDF methods. Both methods succeed in representing text in a numeric format as a vector of a sequence of numbers. The first method, WC, targets to count the words in the document with highlighting the highest counted words as features. This perspective has proved to be not accurate in many situations [28]. On the other hand, TF/IDF adopts the concept of the word weight in the documents' set which is proved to be more accurate [27]. Therefore, the current research aims to utilize the TF/IDF method for the required task. TF/IDF effectiveness has been proven in different research to positively enhance the mining performance as per its accurate text-to-numeric representation.

While different methods identify keywords as the most frequent terms in the document, TF/IDF performs in a different vision. It identify keywords according to the term's relativeness with respect to the whole context rather than a single document. An experiment has been conducted targeting to evaluate both methods. The dataset was retrieved from Kaggle website [23], it included Arabic News dataset. The dataset consisted of six news categories with a total of one hundred and eighty news records. The experiment targeted applying each experiment as a FS method for a set of classification algorithms and evaluate the classification results accordingly. Four classification algorithms contributed to this experiment, they are Decision Tree (DT), KNN, Support Vector Machine (SVM), and Random Forest (RF). For each classification algorithm, the algorithm has been applied before and after applying the selection methods. This means that each algorithms has been applied three times, before FS, after FS using the WC method, and after FS using TF/IDF method. The results are illustrated in Table 1. As shown in Table 1, the results confirms the highest accuracy for the classification algorithms using TF/IDF method.

Table 1. Classification results before and after FS.

Algorithm	Before FS	FS using WC	FS using TF/IDF
DT	91.4	91.6	92.2
KNN	91.7	70.2	94.6
SVM	92.5	92.7	96.1
RF	94.3	93.6	95.2

According to the enhancement that is illustrated in Table 1, TF/IDF succeeds with the highest performance results. However, as TF/IDF does not detect the semantics of the terms and depends only on the relativeness of the term in the whole context, the current research argues that considering the semantic relations between terms provides more accuracy in the exploring the keywords in the context. Therefore, the next step proposed an adaptation to the TF/IDF method which explains how the semantic relations could be explored and results in a higher performance

3.2. The Proposed (STF/IDF) Method by Adapting TF/IDF Based on Siblings Relationship

In this section, an adaptation is proposed for the TF/IDF

method, namely Square root Term Frequency/Inverse Document Frequency (STF/IDF), targeting for higher key terms determination accuracy. Figure 1 illustrates the main steps for the proposed method.

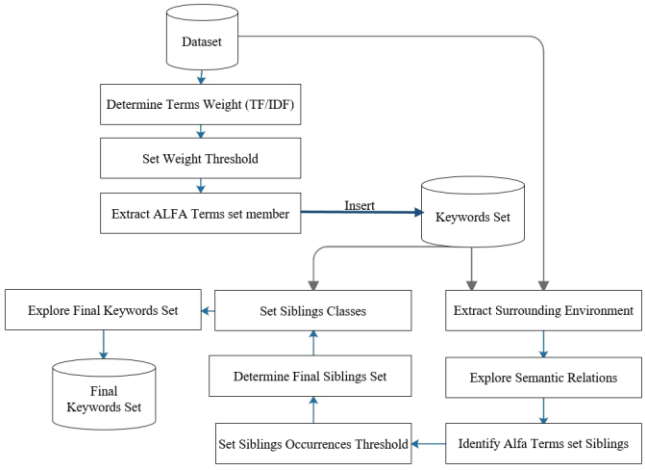

Figure 1. Main steps for the proposed method.

STF/IDF captures the semantics of the terms rather than relying on only the term appearance count in the document in specific and in the context in general. The keywords are explored by the collaboration of two approaches, TF/IDF and the semantic relationship between terms. A term is considered in the keywords' final set for a class when either the term is in the AlfaTerms Set which means that it is above the TF/IDF threshold or the term has a sibling relationship with one of the terms that belong to the AlfaTerms Set. In case that the term and its sibling are members in the AlfaTerms Set and belong to different classes, then they are considered keywords for both classes (multi-inheritance).

The proposed method includes three phases namely determine weight by TF/IDF, Detect terms' siblings, and explore final keywords' set. The first phase deliverable is the first set of key terms, namely AlfaTerms, which are considered subset of the final set of documents' features. The members of this set are determined using TF/IDF measure. In this phase, the TF and the IDF for each term are determined for each term in the documents. According to the identified weights, the terms that weights are above the threshold are extracted and highlighted as members in the AlfaTerms Set. The members of this set are confirmed as keywords to their assigned class. They are then utilized to extend the keywords final set by examining their semantic relationship with other terms in the context.

The second phase deliverable is the first set of key terms, namely SibTerms, which are also considered subset of the final set of documents' features. The key step in this phase is identifying the key terms semantic relationships [38] which are identified by the siblings. A term has a semantic relationship with one of the keywords when they are considered siblings. Siblings are detected when both terms are extracted from similar environments. The environment is identified to be the surrounding text of the term as prefix and postfix in the document. To accomplish the deliverable of this phase, a set of steps are conducted as follows:

The terms in the AlfaTerms set are considered the seeds of this phase. Then, the prefixes and postfixes of all the members in the AlfaTerms set are extracted from the training documents' set. So, at this stage, a developed environment set is built which includes all the prefixes and postfixes with their associated keyword. The next step is to build pairs of the environment set with respect to the associated keyword. Then, using these pairs in the documents' training set, the terms that are embedded in each pair is extracted and considered a sibling to the associated keyword. Each extracted term is included in the primary_SibTerms set. The extracted terms are then weighted by a weighting measure with including the number of extraction for this term with respect to the associated keyword. This step is critical as this is the step that determine whether the sibling will be considered one of the keywords. A sibling is considered one of the keywords when the calculation of the following formula is above the threshold. Finally, the final set, namely Final_Features set of the features are the union set of both SibTerms and AlfaTerms. After detecting the set of features, a comparison between the classification algorithms evaluation results with the AlfaTerms features set that are extracted by TF/IDF and the Final_Features set that are extracted by both TF/IDF and the semantic relations.

3.3. Classification Phase Setup

There are different classification algorithms that positively contributed in text classification. The main deliverable of these algorithms is the labeling of the documents with the main representative class. In this experiment, nine classical classifiers contributed for the three experiments (word count, TF/IDF, and STF/IDF). These classifiers are LR, NB, DT, SVM, RF, KNN, Nearest Centroid Classifier (NCC), Voting Classifier (VC), and NN. Figure 2 illustrates the main steps for the classification phase setup.

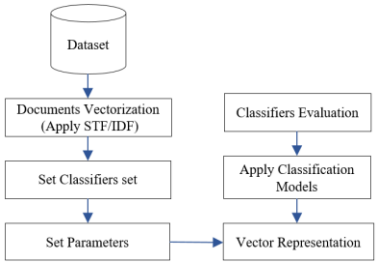

Figure 2. Main steps for the classification phase setup.

The selected classifiers are prepared with the parameters setup. STF/IDF method is used to prepare the documents' vectors. In addition to the nine classical classifiers, two of the deep learning classifiers contributed in this phase. They are Recurrent Neural

Networks (RNN) and Long Short-Term Memory (LSTM). Each of these algorithms includes three layers, input layer, hidden layer, and output layer. The input layer contributes to exploring the semantic relationship between the terms. The input layer converts the input sparse matrix of words to a vector space representation. This is considered a vital step for raising the algorithm performance due to the less training time consumption and the computational complexity reduction. It is worth mentioning that the text processing tasks are applied including tokenization and stemming. The output layer contributed to both experiments, single and multi-labeling. Different designs for deep neural networks [40] are proposed. In this research Bidirectional Recurrent Neural Networks (BiRNN) to contribute to the experiment targeting the confirmation of the proposed STF/IDF applicability and high performance.

4. Experimental Study

The experiment includes twelve classifiers. Three main experiments are conducted. The first includes applying the classifiers with no contribution of any of the FS techniques. The second experiment includes applying TF/IDF as a FS method before classification. Finally, the third experiment includes applying the proposed STF/IDF method before classification. Moreover, the experiment followed the three-fold approach. In each fold, the training phase is conducted over 70% of the data while the testing phase is conducted over 30%. Accordingly, a total of one hundred and eighty sub-experiments are conducted. The dataset is retrieved from Kaggle [23]. It included forty five thousand of the news documents divided into seven categories. The main target of applying the classification algorithms is to determine the performance level of these classifiers with the three different situations and confirms that the proposed adaptation method provides higher performance results. The evaluation of the classification is performed through calculating the accuracy results. The score is determined by the percentage of the number of the classified news in their correct class with the total number of the news dataset members.

Primary experiment has been conducted to confirm neglecting the step of normalizing the terms. This task has been discussed in many research that considered text processing especially those that are conducted on Arabic datasets. The task of normalization is usually conducted to minimize the features size, however, it has been proven in many research that this task leads to a loss in some of the features that could be considered as key attributes. For example, before normalization, the word "فأر" and "فار" are considered two different terms. However, after normalization when replacing the Alef letter "ا ،أ" will be normalized to become one representation, then the two words will become the same which means a loss of the other term in the context. Therefore, this research follows the approach of maintaining the key attributes and not conducting the normalization task.

In the training phase, the number of extracted features varied according to the followed approach, the total average extracted features for each of the three approaches to all categories (word count, TF/IDF, and STF/IDF) are approximately 2500, 1998, and 1575 features respectively.

4.1. Applying Classification Algorithms by the Proposed STF/IDF Features Selection Method

The next step is applying the classical classifiers, the accuracy has been calculated (see Figure 3).

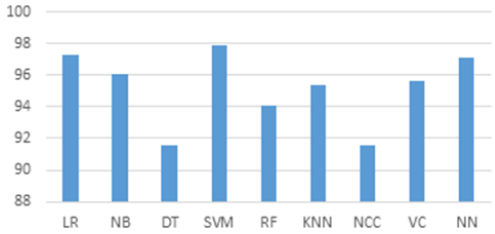

Figure 3. Accuracy of the classical classification algorithms.

The results demonstrates that the highest average accuracy of the classification task is for the proposed method with minimum of 91.6 %. Moreover, the best classifier accuracy lies also in the proposed method experiments by SVM algorithm with 97.9 %. The worst result was for the NCC with the word count method by 86.9%. Moreover, the accuracy of the other classifiers using the proposed method ranges from 96% and 97% while the range of the other methods was from 86.9% and 96.2%.

4.2. Comparing the three Feature Extraction Methods by Applying SVM Classification Algorithm

Figure 4 demonstrates the average accuracy results of SVM with respect to the applied FS method. The accuracy is calculated by the percentage of the correctly classified news to the original classification.

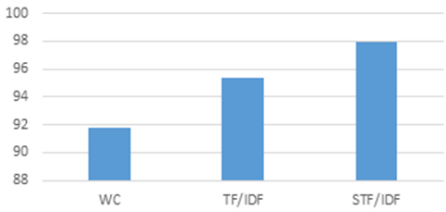

Figure 4. Accuracy of the SVM algorithm.

4.3. Extending Evaluation for the Classification Algorithms by the Proposed STF/IDF Features Selection Method

Moving forward to other performance measures including F-score, AUC, precision, and recall. Table 2

illustrates the results for all classifiers with respect to the applied method. These results are the average of the three folds that are conducted for each classifier.

Table 2. Results for the contributing classifiers.

Classifier	Accuracy	F1-score	Precision	Recall	AUC
LR	97.3	97.4	97.4	97.4	97.9
NB	96.1	96.5	96.4	96.7	97.9
DT	91.6	91.4	91.5	91.4	95.4
SVM	97.9	97.8	97.9	97.8	98.5
RF	94.1	94.9	94.9	94.9	97.8
KNN	95.4	95.4	95.4	95.5	97.8
NCC	91.6	92.2	92.0	92.5	92.5
VC	95.6	95.3	95.3	95.3	97.2
NN	97.1	97.3	97.3	97.3	97.6

Table 3. Confusion matrices for highest SVM algorithm.

	Tech	Sports	Religion	Politics	Medical	Finance	Culture
Tech	6120	10	105	16	0	209	40
Sports	32	6380	36	12	0	40	0
Religion	120	23	6257	21	0	67	12
Politics	29	9	5	6405	7	42	3
Medical	7	5	22	0	6450	14	2
Finance	164	64	65	42	37	6075	53
Culture	28	9	10	4	6	53	6390

For more details in the experiment, Table 3 illustrates the confusion matrix for the best classical classifiers namely SVM. Figure 5 presents an example of one of the News documents that belongs to the Sports class and has been correctly classified to its category by SVM. Figure 6 presents the distribution of the extracted features by the proposed method. A following phase is applied with the deep learning models RNN and LSTM and the results are illustrated in Table 4 including SVM and while Figure 7 confirms that the highest performance belongs to LSTM algorithm.

فاز الإسباني رافايل نادال المصنف ثانياً وحامل اللقب على الأوزبكستاني
دينيس ايستومين 6-2 و6-2 و6-صفر في الدور الثاني لبطولة رولان غاروس
للتنس .كما فاز البريطاني اندي موراي الرابع على الفنلندي ياركو نيمينن 6-1
و6-4 و6-1 و6-،2 والفرنسي جو ويلفريد تسونغا الخامس على الألماني
سيدريك مارسيل ستيبه 6-2 و6-4 و6-2 و6-،1 والإسباني دافيد فيرر السادس
على الفرنسي بينوا بير 6-3 و6-3 و6-،2 والأرجنتيني خوان موناكو الثالث
عشر على التشيكي لوكاس روسول 6-7 (4-7) و6-صفر و7-6 (5-7)،
والروسي ميخائيل يوجني السابع والعشرون على الهولندي روبن هازه 6-3
و7-6 (5-7) و6-،4 والكولومبي سانتياغو جيرالدو على الأسترالي برنارد
طوميتش الخامس والعشرين 6-4 و6-1 و6-،3 والبلجيكي دافيد غوفان على
الفرنسي ارنو كليمان 6-3 و7-6 (2-7) وصفر-6 و6-1 و6-1 .وفازت
التشيكية بترا كفيتوفا المصنفة رابعة على البولندية أورسولا رادفانسكا 6-1
و6-،3 والدنماركية كارولين فوزنياكي التاسعة على الأسترالية ياروميلا
غايدوسوفا 6-1 و6-،4 والألمانية انجيليك كيربر العاشرة على البيلاروسية
أولغا غوفورتسوفا 6-3 و6-،2 والإيطالية فرانشيسكا سكيافوني الرابعة عشرة
على البلغارية تسفيتانا بيرونكوفا 6-2 و6-3 و6-1.

Figure 5. An example of sports document (correctly classified by SVM).

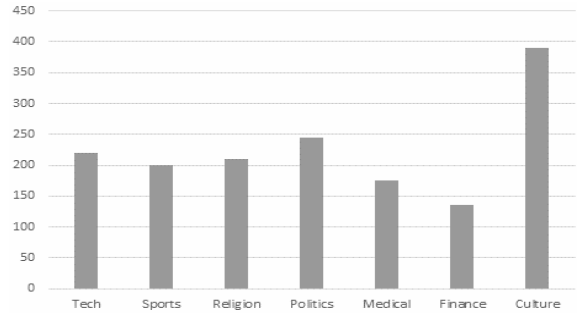

Figure 6. Count of the labels extracted by STF/IDF.

Table 4. Results for the RNN and LSTM Classifiers.

Classifier	Accuracy	F1-score	Precision	Recall	AUC
SVM	97.9	97.8	97.9	97.8	98.5
RNN	98.1	98.0	98.2	97.9	98.9
LSTM	99.5	98.9	98.9	99.1	99.0

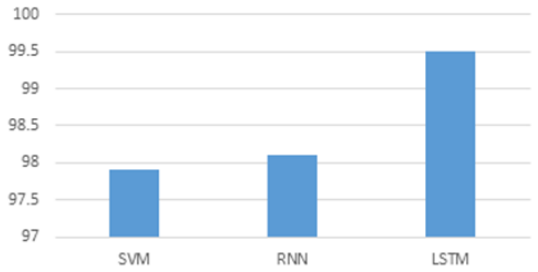

Figure 7. Accuracy of the deep learning classification algorithms.

5. Conclusions

The research focused on processing text documents which could be considered the most rich data source. This source includes data such as books, lecture notes, social networks data, emails, surveys' comments, and many other sources. The immense need for processing these sources targeting for vital information has lead the researchers to closely focus on the challenges of the NLP techniques and text mining methods. In this research, a proposed feature extraction method is proposed as an adaptation to one of the most common methods namely TF/IDF. The current research proposed an adaptation for TF/IDF to consider the semantic relationships between terms. The research proved the argument that considering the semantic relationships enhances the extraction accuracy from text. The research proposed the concept of the siblings' relationship as siblings normally live in the same environment, and hence, the sibling keywords of the main seeds are extracted from text in case they are surrounded by the same environment. Moreover, the Arabic text which was the focus of the current research is known to be a sophisticated language with a high level of ambiguity which raises the complexity processing. The proposed model succeeded in extracting the features following the hierarchy extraction of the siblings from Arabic text with minimum need to the required pre-resources which is also one of the main contribution of the research. The proposed model proved its advancement and raised the extraction accuracy. Moreover, a set of classification algorithms have been applied and evaluated. The algorithms belonged to the traditional set of classification algorithms in addition to two deep learning algorithms. The classification results have been evaluated and compared which highlighted the advancement of the RNN techniques over other algorithms. The highest classification accuracy reached 98.9%. The future research could follow the current research for more enhancements to the extraction methods. Embedding sampling techniques could save the effort of for the compulsory need to a balanced dataset. Working with

multi-labeling for the features is also one of the future research points

Acknowledgment

This work was funded by the University of Jeddah, Jeddah, Saudi Arabia, under grant no. (UJ-23-DR-12). Therefore, the authors thank the University of Jeddah for its technical and financial support

References

[1] Afify E., Sharaf Eldin A., Khedr A., and Alsheref F., "User-Generated Content (UGC) Credibility on Social Media Using Sentiment Classification," *FCI-H Informatics Bulletin*, vol. 1, no. 1, pp. 1-19, 2019. file:///C:/Users/user/Downloads/User-GeneratedContentUGCCredibilityonSocialMedia.pdf

[2] Akcapinar G., "How Automated Feedback through Text Mining Changes Plagiaristic Behavior in Online Assignments," *Computers and Education*, vol. 87, pp. 123-130, 2015. https://doi.org/10.1016/j.compedu.2015.04.007

[3] AlMazroi A., Khedr A., and Idrees A., "A Proposed Customer Relationship Framework Based on Information Retrieval for Effective Firms' Competitiveness," *Expert Systems with Applications*, vol. 176, pp. 114882, 2021. https://doi.org/10.1016/j.eswa.2021.114882

[4] Alzubi J., Nayyar A., and Kumar A., "Machine Learning from Theory to Algorithms: An Overview," *Journal of Physics: Conference Series*, Bangalore, pp. 1-16, 2018. DOI:10.1088/1742-6596/1142/1/012012

[5] Attia M., Abdel-Fattah M., and Khedr A., "A Proposed Multi Criteria Indexing and Ranking Model for Documents and Web Pages on Large Scale Data," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 10, 2022. https://doi.org/10.1016/j.jksuci.2021.10.009

[6] Benabdallah A., Alaeddine M., and Abderrahim M., "Extraction of Terms and Semantic Relationships from Arabic Texts for Automatic Construction of an Ontology," *International Journal of Speech Technology*, vol. 20, no. 2, pp. 289-96, 2017. https://link.springer.com/article/10.1007/s10772-017-9405-5

[7] Boukil S., Biniz M., El Adnani F., Cherrat L., and El Moutaouakkil A., "Arabic Text Classification Using Deep Learning Technics," *International Journal of Grid and Computational Computing*, vol. 11, no. 9, pp.103-114, 2018. https://article.nadiapub.com/IJGDC/vol11_no9/9.pdf

[8] Bourahouat G., Abourezq M., and Daoudi N., "Word Embedding as a Semantic Feature Extraction Technique in Arabic Natural Language Processing: An Overview," *The International Arab Journal of Information Technology*, vol. 21, no. 2, pp. 313-325, 2024. https://doi.org/10.34028/iajit/21/2/13

[9] Chandra P., Ahammed M., Ghosh S., Emon R., Billah M., Ahamad M., and Balaji P., "Contextual Emotion Detection in Text using Deep Learning and Big Data," *in Proceedings of the 2nd International Conference on Computer Science, Engineering and Application*, Gunupur, pp. 1-5, 2022. DOI:10.1109/ICCSEA54677.2022.9936154

[10] Da Rocha N., Barbosa A., Schnr Y., Machado-Rugolo J., De Andrade L., Corrente J., and Silveira L., "Natural Language Processing to Extract Information from Portuguese-Language Medical Records," *Data*, vol. 8, no. 1, pp. 1-15, 2023. https://www.mdpi.com/2306-5729/8/1/11

[11] Dahab M., Idrees A., Hassan H., and Rafea A., "Pattern Based Concept Extraction for Arabic Documents," *The International Journal of Intelligent Computing and Information Sciences*, vol. 10, no. 2, pp. 1-14, 2010. https://scholar.cu.edu.eg/?q=hesham/publications/pattern-based-concept-extraction-arabic-documents-0

[12] Dziadek J., Henriksson A., and Duneld M., "Improving Terminology Mapping in Clinical Text with Context-Sensitive Spelling Correction," *Informatics for Health: Connected Citizen-Led Wellness and Population Health*, vol. 235, pp. 241-245, 2017. https://pubmed.ncbi.nlm.nih.gov/28423790/

[13] Hassan H., Dahab M., Bahnassy K., Idrees A., and Gamal F., "Arabic Documents Classification Method a Step towards Efficient Documents Summarization," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 3, no. 1, pp. 351-359, 2015. file:///C:/Users/user/Downloads/1423537026.pdf

[14] Hassan H., Dahab M., Bahnasy K., Idrees A., and Gamal F., "Query Answering Approach Based on Document Summarization," *International Open Access Journal of Modern Engineering Research*, vol. 4, no. 12, pp. 50-55, 2014. file:///C:/Users/user/Downloads/IJMER.pdf

[15] Hassouna D., Khedr A., Idrees A., and ElSeddawy A., "Intelligent Personalized System for Enhancing the Quality of Learning," *Journal of Theoretical and Applied Information Technology*, vol. 98, no. 13, pp. 2199-2213, 2020. https://www.jatit.org/volumes/Vol98No13/1Vol98No13.pdf

[16] Hawashin B., Mansour A., and Aljawarneh S., "An Efficient Feature Selection Method for Arabic Text Classification," *International Journal of Computer Applications*, vol. 83, no. 17, pp. 1-6, 2013.

https://www.ijcaonline.org/archives/volume83/number17/14666-2588/

[17] Helmy Y., Emam O., Khedr A., and Bahloul M., "A Survey on Effect of KPIs in Higher Education Based on Text Mining Techniques," *International Journal of Scientific and Engineering Research*, vol. 11, no. 3, pp. 1408-1414, 2020. https://www.ijser.org/researchpaper/A-Survey-on-Effect-of-KPIs-in-Higher-Education-based-on-Text-Mining-Techniques.pdf

[18] Idrees A. and Shabaan E., "Building a Knowledge Base Shell Based on Exploring Text Semantic Relations from Arabic Text," *International Journal of Intelligent Engineering and Systems*, vol. 13, no. 1, pp. 324-333, 2020. https://inass.org/publications/contents/?rp=contents2020-1

[19] Idrees A., Alsheref F., and ElSeddawy A., "A Proposed Model for Detecting Facebook News' Credibility," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 7, pp. 311-316, 2019. DOI:10.14569/IJACSA.2019.0100743

[20] Idrees A., ElSeddawy A., and Zeidan M., "Knowledge Discovery Based Framework for Enhancing the House of Quality," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 7, pp. 324-332, 2019. DOI:10.14569/IJACSA.2019.0100745

[21] Idrees A., Helmy Y., and Khedr A., "Credibility Aspects' Perceptions of Social Networks, A Survey," *Social Network Analysis and Mining*, vol. 12, no. 1, 2022. https://link.springer.com/article/10.1007/s13278-022-00924-6

[22] Jasti V., Kumar G., Kumar M., Maheshwari V., Jayagopal P., Pant B., Karthick A., and Muhibbullah M., "Relevant-Based Feature Ranking (RBFR) Method for Text Classification Based on Machine Learning Algorithm," *Functional Nanomaterial-based Flexible Electronics*, vol. 2022, pp. 1-12, 2022. https://doi.org/10.1155/2022/9238968

[23] Kaggle, News Dataset, https://www.kaggle.com/datasets/rmisra/news-category-dataset/, Last Visited, 2024.

[24] Khan M., Rafa S., Abir A., and Das A., "Sentiment Analysis on Bengali Facebook Comments to Predict Fan's Emotions towards a Celebrity," *Journal of Engineering Advancements*, vol. 2, no. 3, pp. 118-124, 2021. https://doi.org/10.38032/jea.2021.03.001

[25] Khedr A., Idrees A., and Alsheref F., "A Proposed Framework to Explore Semantic Relations for Learning Process Management," *International Journal of e-Collaboration*, vol. 15, no. 4, pp. 46-50, 2019. https://doi.org/10.4018/IJeC.2019100104

[26] Khedr A., Idrees A., and Shabaan E., "Automated Ham-Spam Lexicon Generation Based on Semantic Relations Extraction," *International Journal of e-Collaboration*, vol. 16, no. 2, pp. 45-64, 2020. DOI:10.4018/IJeC.2020040104

[27] Mohammed M. and Omar N., "Question Classification Based on Bloom's Taxonomy Cognitive Domain Using Modified TF-IDF and Word2Vec," *PloS One*, vol. 15, no. 3, pp. 1-21, 2020. https://doi.org/10.1371/journal.pone.0230442

[28] Mohsen A., Hassan H., and Idrees A., "Documents Emotions Classification Model Based on TF-IDF Weighting," *International Journal of Computer Electrical Automation Control and Information Engineering*, vol. 10, no. 1, pp. 252-258, 2016. https://zenodo.org/records/1126597

[29] Mohsen A., Idrees A., and Hassan H., "Emotion Analysis for Opinion Mining from Text: A Comparative Study," *International Journal of e-Collaboration*, vol. 15, no. 1, pp. 1-21, 2019. https://doi.org/10.4018/IJeC.2019010103

[30] Mostafa A., Idrees A., Khedr A., and Helmy Y., "A Proposed Architectural Framework for Generating Personalized Users' Query Response," *Journal of Southwest Jiaotong University*, vol. 55, no. 5, pp. 1-13, 2020. http://www.jsju.org/index.php/journal/article/view/714/708

[31] Mouri K., Ren Z., Uosaki N., and Yin C., "Analyzing Learning Patterns Based on Log Data from Digital Textbooks," *International Journal of Distance Education Technologies*, vol. 17, no. 1, pp. 1-14, 2019. DOI:10.4018/IJDET.2019010101

[32] Othman M., Hassan H., Moawad R., and Idrees A., "A Linguistic Approach for Opinionated Documents Summary," *Future Computing and Informatics Journal*, vol. 3, no. 2, pp. 152-158, 2018. https://doi.org/10.1016/j.fcij.2017.10.004

[33] Othman M., Hassan H., Moawad R., and Idrees A., "Using NLP Approach for Opinion Types Classifier," *Journal of Computers*, vol. 11, no. 5, pp. 400-410, 2016. DOI:10.17706/jcp.11.5.400-410

[34] Peng D. and Zhao H., "Seq2Emoji: A Hybrid Sequence Generation Model for Short Text Emoji Prediction," *Knowledge-Based Systems*, vol. 214, pp. 106727, 2021. https://doi.org/10.1016/j.knosys.2020.106727

[35] Pohl H., Domin C., and Rohs M., "Beyond Just Text: Semantic Emoji Similarity Modeling to Support Expressive Communication," *ACM Transactions on Computer-Human Interaction*, vol. 24, no. 1, pp. 1-42, 2017. https://doi.org/10.1145/3039685

[36] Qaffas A., Idrees A., Khedr A., and Kholeif S., "A Smart Testing Model Based on Mining Semantic Relations," *IEEE Access*, vol. 11, pp. 30237-30246, 2023.

DOI:10.1109/ACCESS.2023.3260406
[37] Sabri T., El Beggar O., and Kissi M., "Comparative Study of Arabic Text Classification Using Feature Vectorization Methods," *Procedia Computer Science*, vol. 198, pp. 269-275, 2022. https://doi.org/10.1016/j.procs.2021.12.239
[38] Sarker I., Colman A., Han J., and Watters P., *Context-Aware Machine Learning and Mobile Data Analytics: Automated Rule-Based Services with Intelligent Decision-Making*, Springer, 2021. https://link.springer.com/book/10.1007/978-3-030-88530-4
[39] Sayed M., Salem R., and Khedr A., "A Survey of Arabic Text Classification Approaches," *International Journal of Computer Applications in Technology*, vol. 95, no. 3, pp. 236-251, 2019. https://doi.org/10.1504/IJCAT.2019.098601
[40] Singh M., Sahu H., and Sharma N., *Data Management, Analytics and Innovation*, Springer, 2019. https://link.springer.com/chapter/10.1007/978-981-13-1274-8_28
[41] Wang K., Cao K., Chen M., Yan Z., Zhong L., Yang H., and Cai S., "Front-Page News Classification Model Based on the Stacking of Textual Context and Attribute Information," *Scientific Programming*, vol. 2022, pp. 1-9, 2022. https://doi.org/10.1155/2022/3031195
[42] Yasser F., AbdelMawgoud S., and Idrees A., "A Survey for News Credibility in Social Networks," *International Journal of e-Collaboration*, vol. 18, no. 1, pp. 1-20, 2022. https://doi.org/10.4018/IJeC.304378
[43] Yasser F., AbdelMawgoud S., and Idrees A., *Handbook of Research on Technologies and Systems for e-Collaboration during Global Crises*, IGI Global, 2022. https://www.igi-global.com/chapter/mining-perspectives-for-news-credibility/301832
[44] Zaki S., Ghali N., Abo-Elfetooh A., and Idrees A., "Comparison of Four Ml Predictive Models Predictive Analysis of Big Data," *Journal of Theoretical and Applied Information Technology*, vol. 101, no. 1, pp. 282-289, 2023. https://www.jatit.org/volumes/Vol101No1/24Vol101No1.pdf

Mohamed H. Mousa is a professor received the B.Sc. Degree in Computer Science and Pure Mathematics from Ain-Shams University, Egypt, in 1996, the M.Sc. Degree from Helwan University, Egypt, in 2001, and the Ph.D. Degree from Claude Bernard University Lyon1, France, in 2007. He is Full Professor at the Department of Computer Science, Suez Canal University, Egypt, and is currently Professor at the Department of Artificial Intelligence and Computer Science, University of Jeddah, Saudi Arabia. His research interests include High Performance Computing and GPU Computing.

Ayman E. Khedr currently working as a Professor at the University of Jeddah. He has been the Vice Dean of Post-Graduation and Research and the Head of the Information Systems Department in the Faculty of Computers and Information Technology, at Future University in Egypt. He is a Professor in the Faculty of Computers and Information, at Helwan University in Egypt. He has previously worked as the General Manager of the Helwan e-Learning Center. His research is focused on the Themes (Scientific) Data and Model Management, Data Science, Big Data, IoT, E-learning, Data Mining, Bioinformatics, and Cloud Computing.

Amira M. Idrees is a Professor in Information Systems. She has been the head of Scientific Departments and the Vice Dean of Community Services and Environmental Development, at the Faculty of Computers and Information, at Fayoum University. She is a Professor in the Faculty of Computers and Information Technology at Future University, the head of IS Department, and the head of the University Requirements Unit. My research interests include Knowledge Discovery, Text Mining, Opinion Mining, Cloud Computing, E-Learning, Software Engineering, Data Science, and Data Warehousing.