IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology,
WI-IAT 2022, Niagara Falls, Canada

Accurate Context Extraction from Unstructured Text Based on Deep Learning

Maha Mallek
LARIA, ENSI - LIS UMR CNRS 7020
University of Manouba - AMU
Manouba, Tunisia - Marseille, France
maha.mallek@lis-lab.fr

Ramzi Guetari
SAMOVAR Laboratory - SERCOM Laboratory
Telecom SudParis - University of Carthage
Paris, France - La Marsa, Tunisie
ramzi.guetari@ept.ucar.tn

Sébastien Fournier
LIS UMR CNRS 7020
Aix Marseille University (AMU)
Marseille, France
sebastien.fournier@lis-lab.fr

Wided Lejouad Chaari
Research Unit LARIA, ENSI
University of Manouba
Manouba, Tunisia
wided.chaari@ensi-uma.tn

Bernard Espinasse
LIS UMR CNRS 7020
Aix Marseille University (AMU)
Marseille, France
bernard.espinasse@lis-lab.fr

***Abstract*—Information today is the pillar of the development and the economic growth. A reliable information comes from a flawless exploitation of the appropriate data. The semantic analysis of data is one of the key factors that have a direct impact on the quality of the information that is extracted from it. The subject matter of a given document is closely related to the context of the information conveyed by that document. The reliable identification of the context of a document is, in our opinion, an important added value for an efficient semantic exploitation of the data. In this paper we propose a new approach for the accurate identification of the context associated with a document, given a precise definition of the context. The system has been evaluated on corpora from the "New York Times" and on a context dataset, called "WikiContext", that we have built ourselves for this purpose. Our experiments have shown that our system outperforms the most advanced context extraction systems.**

***Index Terms*—Information Extraction, Context Extraction, Deep Learning, Data Management.**

I. Introduction

Modern information media manage quantities of data that are increasingly voluminous, heterogeneous and complex. This data generally carries vital information whose exploitation requires an in depth analysis which represents a tedious, if not impossible, task for human beings. Traditional processing methods are now obsolete because they are inappropriate, totally inefficient and unable to meet today's needs in terms of data analysis and exploitation. The quantity is not the only difficulty to overcome. The nature of the data is a challenge that requires new, more efficient and more intelligent methods to respond to this lack or total absence of data structuring. It also requires automated means to meet the challenge of volume. In order to produce these automated systems capable of exploiting this data and extracting relevant knowledge, a number of problems must be solved, including a thorough semantic analysis of the content of unstructured documents.

The context (or subject) of a document is one of the most important pieces of information for understanding its content. The results of a clinical report in a medical context are used to establish a diagnosis, while in a judicial context it could be used to establish the seriousness of a defendant's actions. Extracting the context of a document and classifying it are among the tasks to be solved in order to improve the performance of a modern data processing system.

Context is thematic information that captures "the subject of a document" [1]. The title of a document can be seen as its context except that it may be poorly chosen and therefore inadequate or may be singularly lacking in precision which is a source of error. Context is often used to describe a group of texts/documents that all deal with the same subject. Once the context has been accurately established, it becomes easy to classify the text according to its content, which then allows for better extraction of useful information.

Several works have been devoted to the identification of the context of a discourse with more or less convincing results. The most significant results come from methods based on keywords [2] [3] [4], topic modeling [5] [6] [7], metadata [8] [9] [10] and text summarization [11] [12].

In keyword-based approaches, the subject of a document is extracted based on the frequency and co-occurrence of different words in the text. By exploring and organizing the content of a textual document,Topic modeling-based approaches, which include a subject, aim to find subjects. In fact, they seek to organize information into manageable clusters, each of which represents a different subject. Metadata-based context extraction methods use formatting data, like font size, to extract a document's title. Finally, approaches exploiting automatic text summarization make it possible to generate the most important sentence which can represent the context of a document. These approaches will be detailed in the next section.

However, all these methods have the same weakness: they fail to determine the context of a document in a satisfactory way, which significantly affects all tasks related to context, including information retrieval, classification, understanding the content of a document, etc. By relying on keywords, context depends on the choice of keywords and the quality of this choice is not guaranteed. If we rely on the text summary method, the context will be composed of a key phrase. It is almost impossible to find such a significant sentence in a document that can accurately represent its context. Moreover, even if such a sentence exists, the quality of the result is intimately linked to the quality of the discourse. The goal of our work is to produce as a result of an intelligent process, the most precise and relevant context possible, representing a given document.

The remainder of the paper is organized as follows: Related work on context extraction from unstructured texts is presented in the Section II. Section III presents the proposed approach, based on a specific definition of context, as well as the various processes performed by the different components of our implemented system. Using a purpose built "Wikicontext" dataset and the New York Time corpora, the experimental results of our system are presented in section IV, compared with results obtained by state of the art context extraction systems. Finally, Section V concludes this paper and outlines future work.

II. Related Work

There are many research works dedicated to document context extraction. In this section we present different approaches that have been used for context extraction. We mainly focus on the following methods: Topic modelingbased method, keywordsbased methods, metadatabased methods and text summarizationbased methods.

A. Topic modeling-based approaches

In these approaches, the topic represents either a set of keywords that describe the context, or a single word that categorizes the topic of the text, such as (economy, education, culture, etc...). Guo et al. [13] proposed an approach to extract the general theme of a document that lacks much precision. Let's consider the following two sentences:

"*The 2022 Indian Presidential Election will be the 16th presidential election to be held in India*." (S1)

"*Polling agencies are projecting that President Emmanuel Macron has won the 2022 French election*." (S2)

According to the approach presented in [13], these two sentences are classified in the same predefined context "political". However, the identified topic is not precise enough to best characterize the actual context of each sentence. Indeed, the first one concerns the context "2022 Indian presidential election", whereas the second refers to the "French presidential election of 2022". We can conclude that it obviously seen that a single keyword is not enough to characterize the context of a text in a concrete manner.

B. Keyword-based approaches

There are many works based keywordbased models for the extraction of hot topics from news and blogs. These approaches analyze the news on the web and return the words with the highest frequency during a given period of time [14] [15] [16]. This is a perfectly suitable approach, but is dedicated exclusively to short text data from the web and microblogs. Additionally, the authors of [17] [18] have suggested an approach to automate the process of extracting the topic and title from a single document using keywordsbased techniques.

Nevertheless, the extracted topic can be inappropriate. Taking as an example a document with the title " Britain Accuses Ghana Lawmakers of Visa Fraud k", the result of the topic extraction is "Visa Ghana Parliament Britain". This shows that the set of words extracted from the document does not allow to build a sentence or that the sentence obtained is meaningless.

C. Metadata based approaches

Considering that titles typically summarise a document's main idea, additional effort was invested into to the study of title extraction. These studies can be divided into two groups: techniques for HTML pages and approaches for PDF documents. These methods neglect the semantics of the content and instead rely on the style that was applied to the document (font size, alignment, margin, etc.) and some metadata to identify this key phrase.

a) Approaches for PDF documents: [19] and [20] have proposed a concise rulebased heuristic that identifies a PDF's title by taking into account style information (font size). To do so, they used simple empirical rules that reflect the usual practices when presenting a text. We can list the most popular rules as follows: "Titles are frequently located in the largest font sizes," "Titles are typically positioned on the upper parts of the initial pages," etc.

b) Approaches for HTML documents: These solutions depend on elements (tags) in the document's header and body to extract the title. For instance, "Hn" (where n = 1, 2, 3, 4, 5, or 6) and "title" are among the most used tags in this context. Authors of [21] have proposed a general schema enables to learn text titles using style information. The methods used for PDF and HTML documents include a number of weaknesses and ambiguities. Indeed, since the authors of the documents generally enter the metadata, it is therefore subjective. The styles and rules that these methods of title extraction are based on are not always accurate, especially since the authors can change the styles. While PDF documents lack structure, HTML documents also suffer from a lack of reliability. In fact, nothing in the HTML document's authoring rules requires the use of H1 before H2 and H2 before H3, when it comes to the tags H1, H2 ... and H6. Furthermore, these techniques are completely useless if the title is not included in the text.

D. Text summarization-based approaches

Very few publications can be found in the literature that discuss the issue of extracting context based on automatic text summarization. [22] presents an approach that uses the occurrence of words in order to summarize and extract the main topic of a textual document. This approach is effective for automatically extracting the summary of a text but not necessarily its context. A summary gives a general idea of the content of a document and does not necessarily offer its context.

In conclusion, making the context extraction task automatic, accurate, and relevant is the challenge of our work. In fact, our goal is to identify an accurate context that precisely represents the content of a given document by applying text mining techniques.

III. Proposed approach

In this section, we present our new approach for context extraction. We, first, present an history of our previous work and define preliminary concepts. We then present a review of the different components of our process.

A. History and Preliminary concepts

a) History of our approaches: In the last few years, our research team has worked extensively on the context extraction task as well, to solve the issues listed above. In 2020, we proposed a new approach that is composed of two main steps [25]: In the first step, we extract five keywords to identify the final context of the document using FPgrowth algorithm. The second step consists in identifying the sentence that provides a brief idea of the text content. To do this, the document is scanned to identify this one which contains the maximum of the final keywords extracted in the previous process.

Although this approach is effective compared with other systems, results also leave room for further improvement, mostly in the quality of the sentence generated. In [25], we have realized an extractive based approach that cannot generate very high quality context. Indeed, their performance depends strongly on the different sentences included in the document. This latter may not include a sentence with the majority of candidate labels extracted. There is scope for neural network that can enhance the sentence generation phase and thus the quality of the final context.

In this paper, we aim to develop a far well-understanding context which covers what a document is about. Our goal is to make compact contexts in the sense of conveying a semantic title of the document.

b) Preliminary concepts:

- **Context:** The context is very important since it provides an overall idea of what a document or a relation is about. A context "ctx", in our approach, is defined as the minimum information that provides a brief idea of a document "D". It is defined by a label "L" and a set of key words "Kw". A context according to [25] is formalized as:

$$ctx =< Idc, L, Kwi = 1..n > \quad (1)$$

- **Contexts'DataBase** A "Contexts'DataBase" will be used to store contexts that have been etracted from the different documents This database is composed of two tables: The first contains the set of different labels that describe each context. The second table includes a list of contexts labeled with their identifiers, the appropriate set of keywords as well as the collection of documents associated to each context.

B. Context Extraction Process Overview

Our approach to context extraction may be summarized by the process presented in Fig 1. This approach is based on three main tasks: The process begins with the identification of the different Keywords (process I) which consists in identifying the main topics of a document. Among the different keywords extracted, we consider only five that can represent the minimal information that gives a brief idea of the text content. The process continues to extract the label of the document (process II). In this process, we aim to produce a well understanding sentence from the extracted keywords using an LSTM model which allows, each time, to predict the next word. Our approach focuses, finally, on classifying the document in the Contexts 'Database (process III).

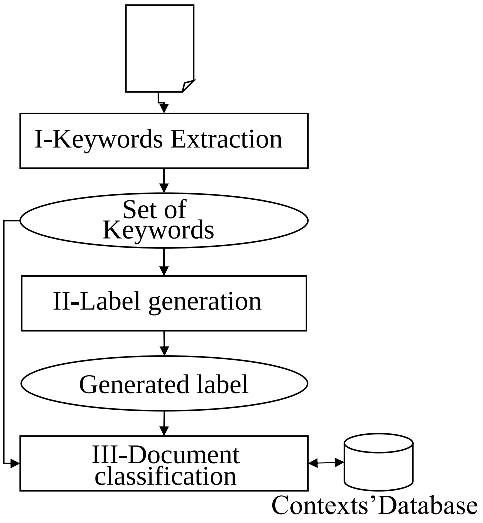

Fig. 1. Overview of the components in our approach for context extraction

C. Keywords Extraction Task

The Keywords extraction is the first step that prepares the document in order to extract the context. This step aims to identify important words that are considered as the main topics of the document. The keywords extraction task requires three main steps as shown in Fig 2: (a) Topics extraction, (b) Candidate label extraction, and finally (c) Final Keywords extraction.

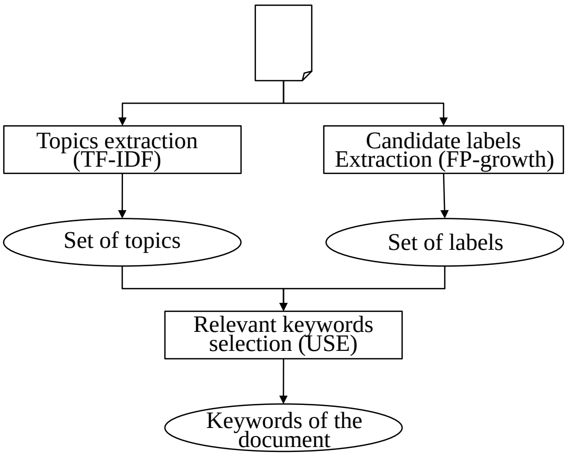
Topics extraction (TF-IDF)
Candidate labels Extraction (FP-growth)
Set of topics
Set of labels
Relevant keywords selection (USE)
Keywords of the document

Fig. 2. Keywords Extraction process

a) Topics extraction: More precisely, this step includes determining the most important words that best describe the document's content. To do this, we have evaluated five different techniques: TF-IDF [26], TextRank [27], KeyBert [28], Yake [29] and Rake [30]. Experiments allowed us to use the TF-IDF algorithm which achieved the best results compared to the other methods. TF-IDF stands for term frequencyinverse document frequency. It is a statistical measure used to quantify the importance of words in a document.

b) Candidate labels extraction: Candidate labels represent the different units that appear frequently in a document.

These units take the form of unigram words, bigrams words, or trigrams words [31]. In this step, we use the FP-growth algorithm [32] to discover these units from the document.

c) Final Keywords extraction: In this step, the Universal Sentence Encoder (USE) [33] model is used to calculate the semantic similarity between each candidate label with the different topics that have already been identified in the topics extraction phase. The USE is trained with a deep averaging network (DAN) encoder and it is based on transfer learning. The five candidate labels that have the highest similarity rate with the different topics are defined as the final keywords of the document and can represent the minimal information that gives a brief idea of the text content.

D. Label generation Task

In order to generate the final label from the extracted keywords, we propose a LSTM model which, from the different identified keywords, allows to model a coherent sentence that represents the main idea of the document. The label generation task requires two main phases as shown in Fig 3.: (1) Training phase and (2) Modeling phase.

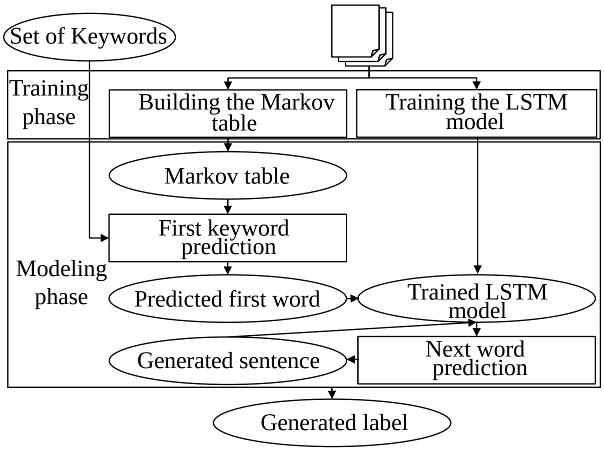

Fig. 3. Label generation process

a) Training phase: Training is the essential step in any machine learning process. It is the first step that prepares our model in order to make accurate predictions and perform the label generation task. There are two main tasks:

- **Training the LSTM model:** Starting from a set of corpora containing sentences expressed in different contexts, we trained an LSTM network. Our LSTM model consists of six layers. The embedding layer enables to convert keywords into a low dimensional dense word vectors. The two LSTM layers are separated with a dropout one to avoid over–fitting. Another dropout layer is applied after the second LSTM layer. Finally, to compute the score of each generated words predicted by the model, a “dense” layer is used with Softmax as the activation function.
- **Building the markov table:** In our approach, the markov table is used to predict the first word of the label and check the sentence generated, for each iteration, by our LSTM model. In this paper, for each context, our Markov table is builded and trained using all the documents of the same dataset used to train our Model.

b) Modeling phase: Among the 5 selected keywords, we consider the one that, according to our Markov table, has the highest probability of being the first in the sentence. Then, we use our LSTM model to predict the next word and use it as part of the prefix for the next input of the model. At this stage, we only consider the generated words that belong to the keyword list or the stop word list. The newly generated sentence is checked by the Markov table. This process is repeated until the list of extracted keywords is complete.

E. Document Classification Task

Our approach to classifying a document in the Contexts’Database may be summarized by the following process (Fig 4):

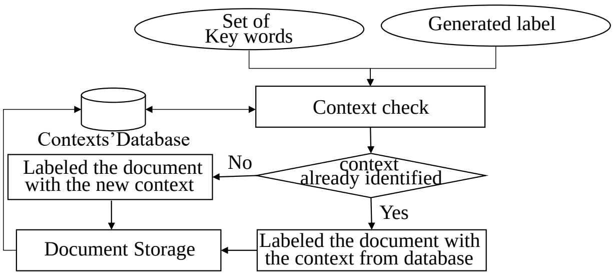
Set of Key words
Generated label
Context check
Contexts’Database
No
Labeled the document with the new context
context already identified
Yes
Document Storage
Labeled the document with the context from database

Fig. 4. Document classification process

The document is classified using both the geneated label and various extracted keywords that represent our context. To do this, two cases can occur: either the context has already listed in the contexts’ Database or if it is a context representing a new document theme. When the context is well identified in the first case, the Contexts’Database is updated by adding the locator of the document to the corresponding context. In the second case, the context database is updated by inserting the new identfied item.

To check if the context has already been identified before or it is not, we compare it with those of the Contexts’ Database. To do this, three main stages are applied: (a) Classification based on the extracted keywords, (b) Classification based the generated label and finally (c) Final classification based on both keywords and generated label.

a) Classification based on the extracted keywords and classification based the generated label: To perform both the classification based on the extracted keyords (a) and the one based on the generated label (b), we have used Bert model. Our Bert model is composed of five layers: The first one ensures that the input to the model respects the required format. The second layer transforms the text into an appropriate input to Bert. The third layer represent the Bert model which transforms its inputs into a vector of fixed size. To avoid overfitting, we use a dropout layer. Finally, a dense layer is included to predict the class of each vector using the softmax function. To classify the generated label or the extracted keywords in our Contexts’Database, we propose a transfer learning approach using the pretrained model of BERT learned on our dataset WikiContext.

Another contribution of this paper is that we have applied data augmentation methods to increase our WikiContext dataset size and to improve the accuracy of classification. To do this, we used the ”TextAttack” framework which allows the augmentation of textual data. Indeed, from each document in the dataset, this model can generate up to 12 additional texts by performing some transformations on the original text such as rephrasing or replacing some words by these synonyms.

b) Final classification based on both keywords and generated label: The final classification model is used to combine the two Bert models respectively for the generated label and the extracted keyword to acheive best results. This combination is performed by adding a new layer that calculates the average between each classification value of the two Bert models described above

IV. Experiment and Evaluation

In order to prove the practical interest of our approach, a system has been developed implementing our approach. This system allows to obtain a more relevant and precise context representing a given document.

A. Datasets

In order to evaluate our approach, we use two different datasets. The first one is our dataset “WikiContext” created manually by collecting 600 English texts published in Wikipedia. Our corpus is composed of 30 contexts and is manually annotated with keywords and title for each text. This dataset is used to perform quantitative evaluation of three process: keywords extraction, label generation and document classification. We use F1score metric, to evaluate the performance of these process.

To perform qualitative evaluation, of the label generation process, we consider NewYork Times dataset as a reference. In this step, we’ll present some generated contexts from various models and compare them qualitatively with our results using ROUGE measure.

B. Results

a) Keywords extraction process results: We compare our method for keywords extraction, against results by multiple methods, including TFIDF, TextRank, Rake, Yake and KeyBert. These methods are tested on our datase WikiContext. The output keywords extracted from each tool is evaluated against reference keywords annotated in our dataset. The results of this evaluation are shown in Table I.

TABLE I
Keywords extraction performance in terms of F1meure on our WikiContext dataset

Algorithm used for keywords extraction	**Precision(%)**	**Recall(%)**	**F1score**
TFIDF	46.89	47.71	47.30
TextRank	15.48	20.94	17.80
Yake	27.17	27.25	27.21
Rake	37.25	37.25	37.26
KeyBert	14.80	14.86	14.83
TFIDF+Jiang and conrath	49.20	65.25	56.09
TFIDF+USE	**65.80**	**73.20**	**69.30**

We can see that our approach significantly outperforms all the baseline methods. The MACRO F1 value of our approach, which is much better than the previous solution on our dataset. For our approach, the best result is based on combining TF-IDF, FPgrowth and USE, and it achieves 69,30%, which is higher than the use of Jiang and conrath similarity 56,09%.

b) Label generation results: The context extraction process is evaluated qualitatively. Table II summarized the performance in terms of ROUGE–Measure on all the text included in the “New York Times” dataset.

TABLE II
Generated label performance in terms of ROUGE measure on New York Times dataset

Algorithm used for generated label process	**ROUGE1**	**ROUGE2**	**ROUGEL**
label extracted by [34]	’f’:0.3, ’p’:0.5, ’r’:0.22	’f’:0.18, ’p’:0.33, ’r’:0.12	’f’:0.32, ’p’:0.60, ’r’:0.22
label extracted by [25]	’f’:0.48, ’p’:0.75, ’r’:0.35	’f’:0.21, ’p’:0.55, ’r’:0.13	’f’:0.44, ’p’:0.70, ’r’:0.33
label extracted by our approach	**’f’:0.63, ’p’:0.65, ’r’:0.62**	**’f’:0.27, ’p’:0.58, ’r’:0.18**	**’f’:0.52, ’p’:0.55, ’r’:0.50**

We can see that our approach significantly outperforms all the baseline methods. Indeed, the ROUGE1, ROUGE2 and ROUGEL values of our approach are much better than the previous solution on the New York Times dataset.

c) Document classification results: In order to evaluate this process, we have performed a transfer learning approach using the pre-trained model of BERT [35] learned on our WikiContext dataset. To get better performance, we have applied data augmentation methods to increase our WikiContext dataset size. Table III summarized the obtained results.

TABLE III
Results provided by the document classification process

Model used for the classification process	**Data augmentation (DA)**	**Precision**	**Recall**	**F1-score**
Bert Model based on extracted keywords	without DA	0.61	0.57	0.54
Bert Model based on extracted keywords	with DA	0.82	0.71	0.76
Bert Model based on generated label	without DA	0.83	0.81	0.82
Bert Model based on generated label	with DA	0.95	0.90	0.91
Bert Model based on both keywords and generated label	without DA	0.98	0.88	0.88
Bert Model based on both keywords and generated label	with DA	**0.95**	**0.91**	**0.92**

According to III, the pretrained BERT model based on the extracted keywords and the generated label and with applying a data augmentation method outperforms the other models with F1-score of 0.92%.

Adding precision to context extraction could be integrated to different applications as decision-making. Particularly, when dealing with search engines, a more accurate context improves document classification process and then information search effectiveness.

V. Conclusion

Unstructured data is by definition difficult to exploit by automated means and remains dedicated to manual processing guided by the power of the human spirit. New intelligent algorithms and applications as well as specialized computer systems must be invented to overcome this impotence. To be efficient and relevant, such systems have to ”understand” the content of these unstructured documents and this understanding is only effective if the context or the theme treated are well identified. In this paper, we have presented a new approach that automatically extract context of a given document. This approach enabled us to produce a more accurate and relevant context compared with other systems . The results show that the extracted contexts are quantitatively and qualitatively much more precise than those identified by these others systems.

Results also leave room for further improvement: (i) The proposed solution assumes that a document is only associated to one context. In the future, we will concentrate on improving the effectiveness of this approach by taking into account the possibility that a document could be attributed to several contexts or distinct subcontexts; (ii) At the present time, we only consider the English language. In our perspective, it is crucial to extend this study to include complex languages as Arabic and Chinese;

References

[1] M.R. Brett, “Topic Modeling: Basic Introduction,” Available: http://journalofdigitalhumanities.org/2-1/topicmodeling-a-basicintroduction-by-megan-r-brett/.

[2] M. Yinghua, S. Guiyang, L. Jianhua and L. Shenghong, “A novel text subject extraction method,” IEEE International Conference on Natural Language Processing and Knowledge Engineering, 2003.

[3] F.Z. Lahlou, A. Mountassir, H. Benbrahim and Ismail Kassou, “A Text Classification based method for context extraction from online reviews,” Intelligent Systems: Theories and Applications (SITA), 2013.

[4] Z. Wang, K. Hahn, Y. Kim, S. Song and J.M Seo, “A news-topic recommender system based on keywords extraction,” Multimedia Tools and Applications, vol. 77, pp. 4339–4353, 2018.

[5] F. Viegas, W. Cunha and Ch. Gomes, “Semantically-Enhanced Topic Modeling,” Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pp. 893-902, 2018.

[6] X. Zhang and R. He, “Topic Extraction of Events on Social Media Using Reinforced Knowledge,” International Conference onKnowledge Science, Engineering and Management, pp. 465-476, 2018.

[7] S. Yang, Q. Sun, H. Zhou, Z. Gong, Y. Zhou and J. Huang, “A Topic Detection Method Based on KeyGraph and Community Partition,” Proceedings of the 2018 International Conference on Computing and Artificial Intelligence, pp. 30-34, March 2018.

[8] Y. Hu, H. Li, Y. Cao, D. Meyerzon and Q. Zheng, “Automatic Extraction of Titles from General Documents using Machine Learning,” Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries, pp. 145-154, June 2005.

[9] Y. Wu, X.J. Zhang, Q. Li and J. Chen, “Title extraction from Loosely Structured Data Records,” Machine Learning and Cybernetics, Vol. 5, 2008.

[10] S. Changuel, N. Labroche and B. Bouchon-Meunier, “A General Learning Method for Automatic Title Extraction from HTML Pages,” International Workshop on Machine Learning and Data Mining in Pattern Recognition, pp. 704-718.

[11] X. Ji and H. Zha, “Extracting Shared Topics of Multiple Documents,” Advances in Knowledge Discovery and Data Mining, , Seoul, Korea, 2003.

[12] J. Silva, J. Mexia, A. Coelho and G. Lopes, “Multilingual Document Clustering, Topic Extraction and Data Transformations,” Progress in Artificial Intelligence, Knowledge Extraction, Multi-agent Systems, Logic Programming and Constraint Solving, Porto, Portugal, 2001.

[13] N. Guo, Y. He ; C. Yan ; L. Liu and C. Wang, “Multi-Level Topical Text Categorization with Wikipedia,” IEEE/ACM 9th International Conference on Utility and Cloud Computing (UCC), Shanghai, China 2016

[14] Y. Jahnavi and R. Yalavarthi, “Hot topic extraction based on frequency, position, scattering and topical weight for time sliced news documents,” 15th International Conference on Advanced Computing Technologies (ICACT), 2013.

[15] H. Ma, “Hot topic extraction using time window,” International Conference on Machine Learning and Cybernetics, China, July 10-13, 2011.

[16] Y.P. Zhang and H. Zhang, “Social Topic Detection for Web Forum,” International Conference on Computer Science and Service System, 2012.

[17] A. Sajid, S. Jan and I.A. Shah, “Automatic Topic Modeling for Single Document Short Texts,” International Conference on Frontiers of Information Technology (FIT), 2017.

[18] J. Yun, L. Jing and Y. Zhang, “Document Topic Extraction Based on Wikipedia Category,” Fourth International Joint Conference on Computational Sciences and Optimization, 2011.

[19] J. Beel, B. Gipp, A. Shaker and N. Friedrich, “Extracting Titles from Scientific PDF Documents by Analyzing Style Information,” International Conference on Theory and Practice of Digital Libraries, pp 413-416.

[20] J. Beel, S. Langer, M. Genzmehr and C. Mueller, “Docear’s PDF inspector: Title extraction from PDF files,” Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries.

[21] S. Gupta and K.K Bhatia, “Domain Identification and Classification of Web Pages Using Artificial Neural Network,” International Conference on Advances in Computing, Communication and Control, pp. 215-226, 2013

[22] M. Shoaib Jameel, Anubhav, N. Singh, N.k. Singh, C. Singh and M. K. Ghose, “An Intelligent Automatic Text Summarizer,” International Conference on Intelligent Human Computer Interaction, pp 223-230.

[23] M. Mallek, R. Guetari, N. Ettayeb, and W. Ghariani “Graphical representation of statistics hidden in unstructured data: a software application,” 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 5-8 Oct, Banff, Canada, 2017.

[24] J. Jiang, and D. Conrath, “Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy,” International Conference Research on Computational Linguistics (ROCLING X), pp. 19–33, 1997.

[25] M. Mallek, S. Fournier, R. Guetari, B. Espinasse and W. Chaari, “An Unsupervised Approach for Precise Context Identification from Unstructured Text Documents,” International Conference on Tools for Artificial Intelligence (ICTAI), 2020

[26] G. Salton G and C. Buckley, “Term-weighing approaches in automatic text retrieval,” In Information Processing and Management, 1988

[27] R. Mihalcea R and P. Tarau, “TextRank: Bringing Order into Texts,” Conference on Empirical Methods in Natural Language Processing, 2004

[28] P. sharma, “Self-supervised Contextual Keyword and Keyphrase Retrieval with Self-Labelling,” 2019

[29] Campos R., Mangaravite V., Pasquali A., Jorge A.M., Nunes C., and Jatowt A., “YAKE! Collection-independent Automatic Keyword Extractor,” Advances in Information Retrieval. ECIR, France, pp. 806-810, 2018.

[30] S. Rose, D. Engel, N. Cramer and W. Cowley, “Automatic keyword extraction from individual documents,” Text Mining: Applications and Theory, pp.1 - 20

[31] KA. Dhand, JS. Umale and PA. Kulkarni, “Context Based Text Document Sharing System Using Association Rule Mining,” Annual IEEE India Conference (INDICON), Pune, India, pp. 11-13, Dec 2014.

[32] C. Borgelt, “An Implementation of the FP-growth Algorithm,” the 1st international workshop on open source data mining, Chicago, Illinois, pp. 21 - 21, 2005.

[33] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. Abrego , S. Yuan, Ch.Tar and R. Kurzweil, “Multilingual Universal Sentence Encoder for Semantic Retrieval,” July 2019.

[34] A. Sajid, S. Jan and I.A. Shah, “Automatic Topic Modeling for Single Document Short Texts,” International Conference on Frontiers of Information Technology (FIT), 2017

[35] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2019