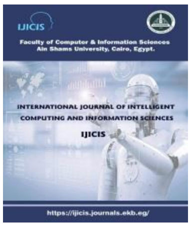
IJICIS, Vol.21, No.1, 68-83

DOI: 10 IJICIS-2103-1070



International Journal of Intelligent Computing and
Information Sciences

https://ijicis.journals.ekb.eg/

Mining publication papers via text miningEvaluation and Results

Ahmed S. Ibrahim
Faculty of Computer and Information Sciences,Ain Shams University, Cairo,11566, Egypt,
a.saeedibr@cis.asu.edu.eg

Sally Saad
Faculty of Computer and Information Sciences,Ain Shams University, Cairo,11566, Egypt,
sallysaad@gmail.com

MostafaAref
Faculty of Computer and Information Sciences,Ain Shams University, Cairo,11566, Egypt,
aref_99@yahoo.com

Received 2021-3-9; Accepted 2021-4-5

Abstract: *Data nowadays is the language of technologies as every process needs a data to be processed the input is data and the output also is data. Analyzing the data is a significant task especially with the increasing production of the data particularly data as a text, it would be difficult to manually analyze the data, extract information and detect the hidden patterns from unstructured text. Datamining is automated technique for gathering or deriving a new high-quality information and uncover the relations among the data. Text mining is one of main branches of the data mining however data mining is more comprehensive this paper, an overview for mining the publication papers via text mining techniques and their results and evaluation would be presentedas following: the first approachis keywords extraction using natural language processing (NLP) approach, the second approach named entity recognition and the last approach is document clustering where machine learning techniques are applied to the both of them.*

Keywords*: Publication Papers, Machine Learning, Text Mining, Named Entity Recognition, Clustering*

1. Introduction

In the era of evolution, the amount of data is increased continuously in a tremendous way as data might be represented in a different form as following numbers and text. data is the main unit that would be used to produce meaningful information, so mining the data considered a substantial process to discover

* Corresponding author: Ahmed S. Ibrahim
Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
E-mail address: a.saeedibr@cis.asu.edu.eg

the patterns or hidden relation they might be found in random data. Most of the data is generated in a form of text where it might be in a structured format (row and column) or unstructured (document).

Mining an unstructured text is considered a difficult task in a comparison of the structured text. Text mining may be a valuable process as it is used for analyzing data without prior knowledge of the precise words or terms to apprehend main concepts, themes and uncover hidden relationships and trends that the authors have used to express those concepts. Most of the Publication papers documents consists of the following paragraphs, charts, images, and tables, so mining the text part of the documents should help in finding and discovering the relationships among the words or among the documents, where this process would help researches and viewers to find the most similar papers related to the main paper, on the other hand, saving them time and efforts [9].

Mining publication papersconsists of five modules [12]. Thefirst moduleis the information Retrieval (IR), which is the process of gathering information resources where it is particularly text (Publication Papers) from a collectionof unstructured data thatfulfills the required data. The second module used for pre-processing is Natural Language Processing (NLP).The third module is retrieving information that relates to specific topic this referred to information extraction.The fourth module is a named entity recognition using the convolution neural network (CNN)a supervised learning technique.last module is clustering using the k-mean unsupervised leaning technique, so the implementation of the system isbased onthe machine learning algorithm, and NLP techniques.This paper is organized in the following manner: Section II for Related Work, Section III Proposed Technique, Section IV Discussion, Section V analysis and results, then followed by the Conclusion and References as the last section.

2. Related Work

Jai Sharma et. al presented the contributions and novel techniques used for mining the text. Several techniques could be used for mining the text and discover the knowledge from a text, these techniques use either machine learning (ML) or natural language processing algorithms. The first technique is text categorization (text classification) is supervised learning technique where the created model is trained using predefined input/output examples to acquire the ability to classify after the training and testing phase. The second technique is Clustering is an unsupervised technique which is an unsupervised learning technique used to group similar documents where no pre-defined (labeled) input is used. The third technique is Information Extractions (IE) where data mining is applied to extract the interesting patterns or knowledge however, a preprocessing step is required which is converting the documents into structured databases. The fourth technique is information visualization (IV) using visual hierarchy or map for a large textual source which provides browsing and searching capabilities easily. The fifth technique is Natural Language Processing.

S.-H. Liao [1] presents the methodology of mining the text in the following steps: gathering documents, pre-processing, extracting data, text diversion, feature extraction, pattern selection, analysis, and evaluation. This paper introduced the widely used text mining approaches, for example clustering, categorization, information extraction, topic tracking, text summarization, and their application in diverse fields are surveyed.

Sumathy et. al [7] discussed different kinds of text documents which are three types: structured; semi-structured or unstructured. described the used technique for extracting useful information. Giving an overview of applications, tools and issues faced mining the text. in this paper, a general framework has been presented for concept-based mining which can be visualized as knowledge text refinement and filtration stages.

Chau et. al [11] aimed to extract meaningful entities from police narrative reports, such as a person, location, and organization using machine learning (ML) with feedforward /backpropagation neural network the system Performed well for some entities, and not so well for others, Person names and narcotic drugs score(74.1% - 85.4% precision) respectively, Personal property and addresses precision was not significant.

MohdAriff et.al [16] Introduced a study among three different methods for clustering which are applied and compared to identify the best method which yields the best result for clustering the documents. The used methods in the study are hierarchical clustering, k-means, and k-medoids. All methods are applied to sports articles with four different types of articles with 60 articles as the total number. Finally, the study resulting presented as the following: the best method found is hierarchical clustering which found to be more stable producing meaningful result because the word is less sensitive however, the k-means clustering resulting a high score in one condition if the word is correctly chosen that the reason that makes this method is more sensitive towards general words. The worst method result is medoids clustering.

Tomasz [15] presents the most important concepts related to document clustering. Discussing and presenting different clustering algorithms however these algorithms may vary in complexity and the quality of the results. Three different categories of the clustering algorithms are presented each one has its feature. the first one is hierarchical algorithms where it performs better than others. The second category is flat or partitional algorithms where it is faster than others. the last one is hybrid where combines both mentioned features like bisecting k-means. Concluding that the algorithms could be chosen based on two variables complexity and the quality of the results.

YADAV et.al [17] presented a comparative study for different NER architectures that discovered recently with the pervious approaches and highlighting the improvements of the achieved result. most NER architectures are based on the supervised or semi supervised algorithms. Deep neural network records a significant result for NER systems. The survey covered, both classic machine learning models, and modern feature-inferring neural network models. The neural network models surpass the classical model. Ma et.al (2016) reach the significant results in English language scoring 91.62% for the accuracy of their system.

3. Proposed Methodology:

This paper introduces a methodology for mining the published papers. The main objective of this methodology is to search or query for the published papers that match the similar keywords which exist in a particular paper. The search operationproduces a collection of papers that wouldshare a common terminology that helps in finding the invisible relationships among papers, then detecting the main entities in the document and classifying keywords to its corresponding class (entity) for every paper which helps to discover the link among words, at the end clustering the documents into a set of

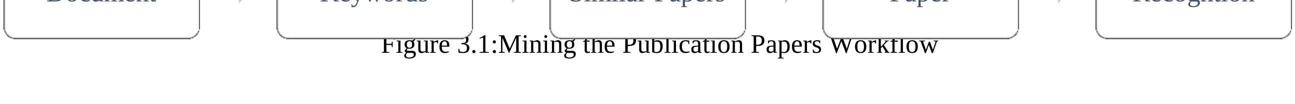


Figure 3.1:Mining the Publication Papers Workflow

groups. This methodology is based on text mining techniquesusingmachine learning and natural language processing.

A. Reading the publication papers (Documents)

Reading documents (publication papers) is the initialstage for mining the documents, most of the documents contain a various type of data however, the text format is the data type commonly exists in the documents. Only text is essential part for this methodology. Papers would be in several file formats for example: (Txt, Pdf, Docx, RFT, LTX …. etc.).

B. Keywords Extraction

Also known as key-phrase extraction, is one of the areas in text mining that is used to distinguish or recognize the most important and useful words, phrases were also called terms. It is the process of identifying the terms automatically that represents optimally the theme of a published paper to the reader. The manual process of keyword extraction is a very difficult and time-consuming task. Therefore, there is in need for an automated process for extracting keywords from the documents [13].

keyword extraction process needs a pre-processing steps to prepare the published paper for the extraction process by removing inconsistences, redundancies words from them,so every document must go through some steps to refinethem from unnecessary words and extract keywords with the highest impact factorwhich are used for querying process to find outthesimilar paper figure 3.2shows resulting from this phase a vector of words [14].

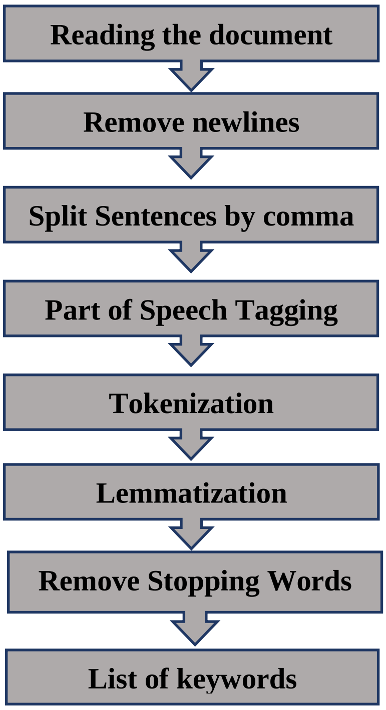
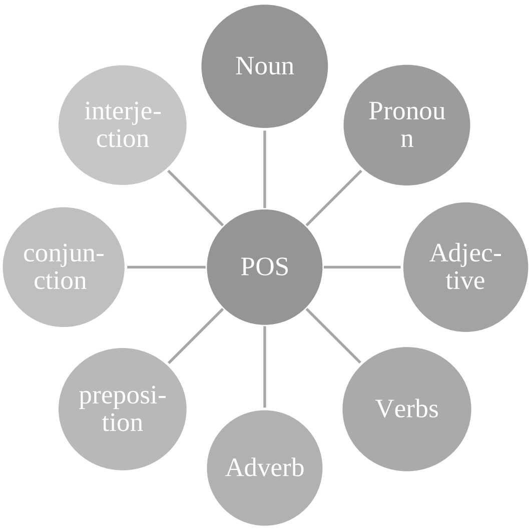


Figure 3.2: keywords extraction Preprocessing steps



- **Parts of speeches (POS)**

Part of speechis an expression that represents the class or category of the wordwhich is considered as the smallest unit in a sentence that has a meaning, based on its context and definition, or how the word is employed in a sentence. This task is not straightforward, as the word may have a various part of speechdepending on the context of the wordwhereit is used [2]. There are eight main classes of the part of speech consists of eight main classes (also known as word classes) see figure 3.3:

3.3 Different classes of part of speech

A word in a sentence can be represented in more than one POS and that should be taken into consideration where it might give different meaning based on the context of the sentence.
For example, with the word play:
a) Playcould be used asa verb: e.g. She played the ball and ran forward.
b) Playcould be used asa noun e.g. The play was interesting.

After eliminating the line separator and split the paragraphs in the documents by a separator, a vector of sentences would be produced ready for processing.Stanford POS techniques is a library used to find the POS for every word. The input is a vector of sentences and the output of this stageis a list representing each word with its corresponding part of speech based on the context of the words in the sentences.

- **Lemmatization**

Lemmatization is the process of combining the words with various forms into the original form (Root). ensuring that the root word belongs to the language, so those wordswould be processed as a single item, identified by the word's lemma, or dictionary form. Words may appear in different inflected forms in many languages.In English,for example the verb 'to go' may appear in a different form as 'go', 'gone', 'goes', ‘going’. the lemma for a word is the base form of the word, that one might look up in a dictionary.

The lexeme of the word is the union of the root form with its part of speech. Lemmatization relies on identifying the meaning of a word and its part of speech in a sentence correctly, as well as within the larger context surrounding that sentence, such as neighboring sentences or even a whole document. Lemmatizing and stemming are closely related to each other however, the main divergence is that Lemmatizing works on multiple words with the knowledge of the context, and therefore it would be able to discriminate between words that have various meanings based on part of speech [4].

Define the part of speech tags is based on the conditional probabilities by surrounding context features using Probabilistic approaches NLP Stanford POS tagger, obtaining from the manually tagged corpus these probability values. Based on figure 3 lemmatization would be applied to every word in the sentence using theirPOS produced previously See Figure 3. the word with its corresponding POS would be the phases'input and the output would produce a list of words that represents each word with its corresponding origins [5].

- **Tokenization**

The task of representing the document as a group of words by portioning or breaking every sentence into a series of tokens or parts (document unit) such as words, keywords, symbols, phrases, or breaking the paragraph into a set of sentences. Sentences are separated by line breaks, line breaks, whitespace. Those token helps to understand the context by analyzing the sequence of the words or sentences which is used for text analysis.

- **Removal of Stop word**

The task of eliminating any unnecessary words from a document that will not change or affect the meaning of the sentence. Those words referredas stopping words for example: a, an, but, and, of, the, etc. Most of the search engineswas implemented to discard or filter stopping words.

C. Searching for Similar Papers

The keyword list produced previously would be used to find the scientific papers containing similar keywords using the search engine such as Google, Yahoo, Bing, etc. See Figure 3.4. The query might outcome many papers therefore, the papers selection criteria would from 7 to 10 recent papers.

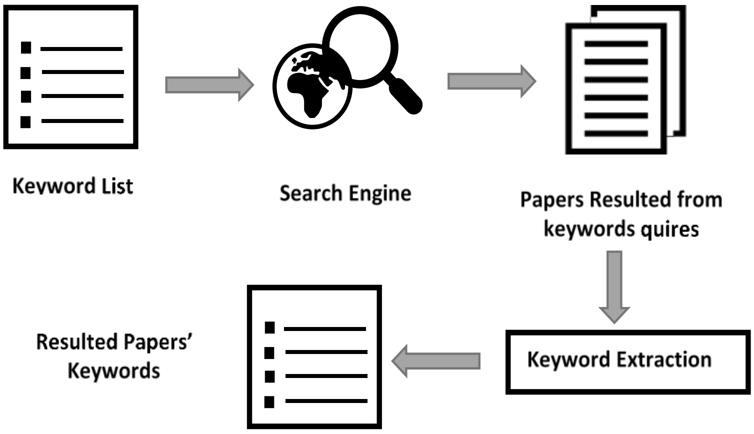

Keyword List
Search Engine
Papers Resulted from keywords quires
Keyword Extraction
Resulted Papers' Keywords

Figure 3.4search for related papers usingthe extracted keywords

The following steps show how this module is built:

- TF-IDF term by a term frequency–inverse document frequency technique is used to determine the frequency of keywords.
- A threshold value is used to refine the list of keywordswhere a word will not be considered as a keyword only ifits frequency is less than the given threshold otherwise the word is considered as a keyword.
- The search engineis used tobrowse for similar published papers usingrefined keywords see Figure.3.5.
- Determine the similaritybetween the main paperand each paper in the resulted list. The next phase is measuring the similarity using the cosine similarity measure.
- Papers with a high similarity value would be considered as similar papers that match the main paper.

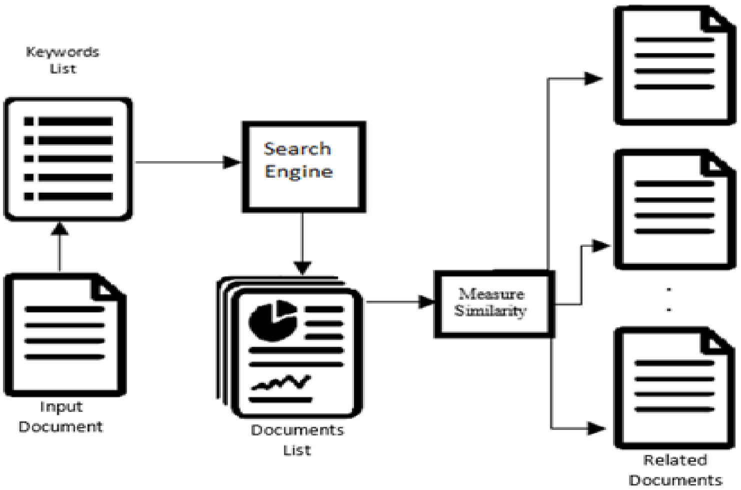


Figure 3.5: Searching for related documents workflow

D. Named Entity Recognition

It recognizes the main entities in the document automatically then classifying them to a predefined class, also referred as Entity identification, entity extraction or name entity recognition.person,locations, organizations etc. considered as a predefined class in the text. The main objective of analyzing text using NER is to recognize the key information of a text. NER is one of the important modules to solve many problems in the research field like answering a question,retrieving information, Machine Translation, inserting annotations into videos, WebCrawling, etc. This task could be implemented using machine leaning and natural language processing techniques

Entity extraction is considered as a subtask of information extraction that deals with structured or unstructured text.Entity extraction consists of two tasks, the firstly task is the identification of proper names in textand the secondly task is the classify these names into a group of predefined categories/entities (classes) see Table 1.

The main entities and their terms are identified and extracted fromall related papers that match the main paper. A preprocessing step must be applied usingthe following steps:apply word tokenization and part-of-speech tagging to the sentence. where those steps are applied inthe keyword's extraction phaseSee Figure 3.6.

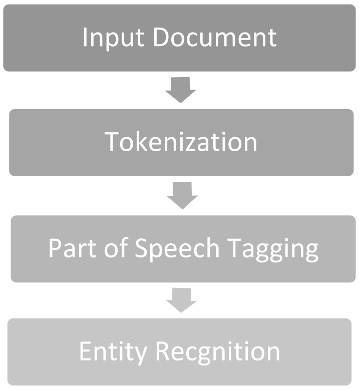


Figure 3.6: Entity Recognition Pre-processing Steps.

Table 1: NER Entities Description

ENTITY	DESCRIPTION
PERSON	People Name, including fictional
NORP	Nationalities, religious or political groups
FACILITY	Buildings, Airports, Highways, Bridges, Squares, Tunnels, Stations
ORG	Companies, Corporation, Agencies, Institutions, Foundation
GPE	Countries. Cities, States, Organization or Nations
LOC	Locations, Mountain Ranges. Bodies of Water
PRODUCT	Objects, Cars, Foods or Drink. etc. (Not services)
DATE	Absolute or relative dates or period
CONCEPT	Scientific, Computer Terms
EVENT	Named Hurricanes, Battles. Wars, Sports Events, Festivals, etc.
WORK_OF_ART	Titles of books, Songs or Paints, etc.
LAW	Name of documents related to laws
QUANTITY	Measurements, i.e.height, mass or distance
LANGUAGE	Name of any language
TIME	Times less than a day
PERCENT	Percentage, including "%".
MONEY	Monetary or financial values, including unit
ORDINAL	"first", "second", etc.

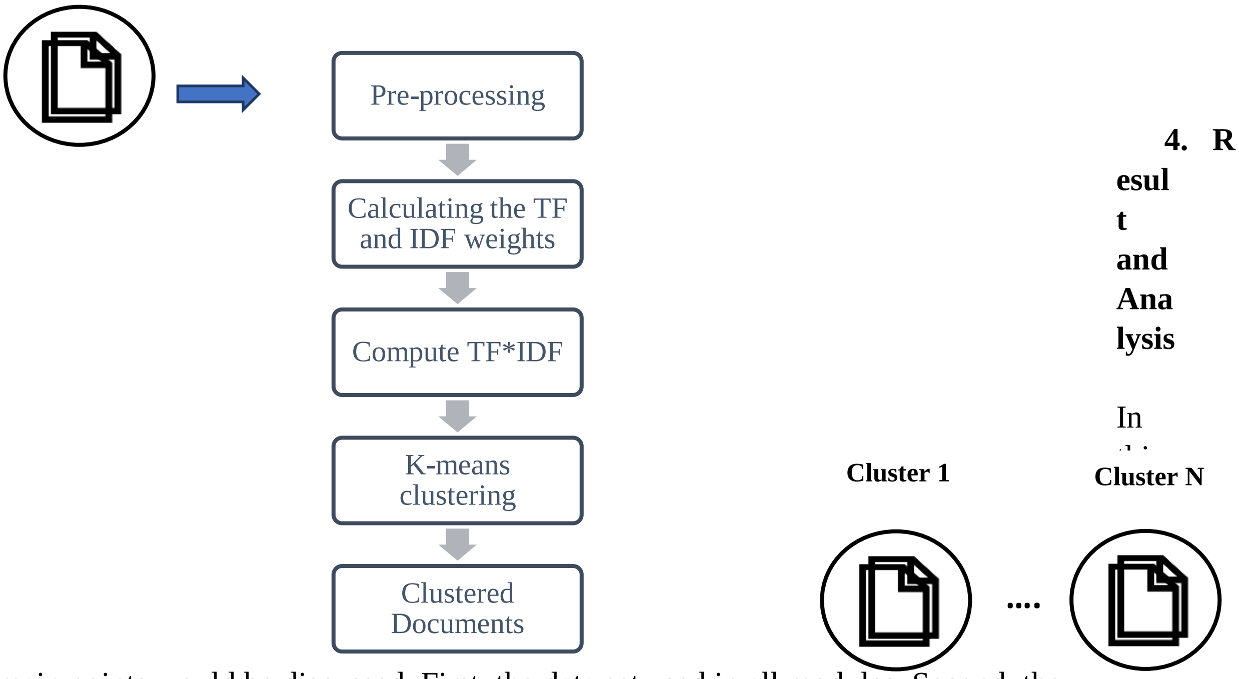
E. Publication Papers Clustering

Clustering documents is the last module of mining the publication papers. Document clustering is an unsupervised technique for organizing a document, also referred as text clustering.it is a special subset of data clustering. it groups the data (document) into several groups based on their features. each group or subset represents a unique cluster. Document clustering algorithms are implemented widely in different fields such as web searching engine and browsing [10].

There are many techniques that involves clustering data like SVD, SSM, BRICH, PCA, and K-means. The goal of using document clustering is to organize resulted papers obtained from a query search into subgroups that each group documents relate to each other.A preprocessing step is required before using any clustering technique. The preprocessing phase consists of multiple steps as following:Filtering, Tokenization, Lemmatization, Remove stopping words.

The technique used for clustering documents in this module is K-means clustering seeFigure 3.7.its characterized by simplicity and good results more than the others. The k-means algorithm used for implementation presented as the following steps: which its popularity is due to

1) Select the number of clusters [n]
2) Select [n] random points as a center of the clusters (documents)
3) Allocate every document to the closest cluster
4) Terminate, only If thestopping criteria is reached.
5) Else, redetermine the new points for newly formed clusters.
6) Return to the step 3



4. R esul t and Ana lysis

N

In

on,three main points would be discussed. First, the data set used in all modules. Second, the evaluati on

Figure 3.7:Proposed clustering Architecture

metrics used to determine accuracy, precision, recall and f1 score. The last point is the evaluation of modules would be shown and analyzing the outcome of the modules.

4.1 Dataset:

The dataset is acquired manually for most of the modules of the system as no such dataset meet the system requirements. Dataset helps in testing the modules and verify the implemented system by determining the accuracy of each module and overallsystem. Each dataset for each module will be explained in terms of how the data acquired, what it looks like (data form), how it going to be used in the system, the size of the data collection, the source of the dataset acquired see Table 2. The data acquired mostly is a set of documents, keywords, or sentences.

Table 2: Comparison among the four main modules of the system dataset

Property / Module	Extracting keywords	Searching for similar paper module	Named Entity Recognition (NER)	Paper Clustering
Dataset Creation	Manually	Manually	Manually	Automatically
Resources	Scientific research gates for example google scholar etc.	Extracted keywords from the gathered published papers.	Representative sentences formed manually /gathered from websites	Acquiring the dataset from the result of second module.
Dataset Form	Published Papers	Set of keywords	Set of Sentences	Published Papers
Size	200 papers	3000-4000 Keywords	350 different Sentences	1000 papers

4.2 Evaluation Metrics:

$$\boldsymbol{Percision = \frac{tp}{tp + fp}} \quad (1)$$

$$\boldsymbol{Recall = \frac{tp}{tp + fn}} \quad (2)$$

$$\boldsymbol{F1 = 2 * \frac{Percision * Recall}{Percision + Recall}} \quad (3)$$

Where: The $\boldsymbol{tp}$ is stands for true positive, $\boldsymbol{fp}$ false positive and $\boldsymbol{fn}$ false negative.

4.3 Experimental Results

A. Name Entity Recognition

Table 3: Precision, Recall and F1 Score of the used entities

Entity	Precision	Recall	F1 Score
GPE	75.00	85.70	80.00
PERSON	57.70	100.0	73.20
CONCEPT	63.30	55.3	59.00
PRODUCT	60.00	53.70	56.70
ORG	82.60	76.00	79.20
DATE	66.50	33.33	44.41
MONEY	65.00	75.71	70.23
LOC	100.00	47.6	64.50
NORP	100.00	90.00	94.74
LANGUAGE	72.30	60.00	65.68

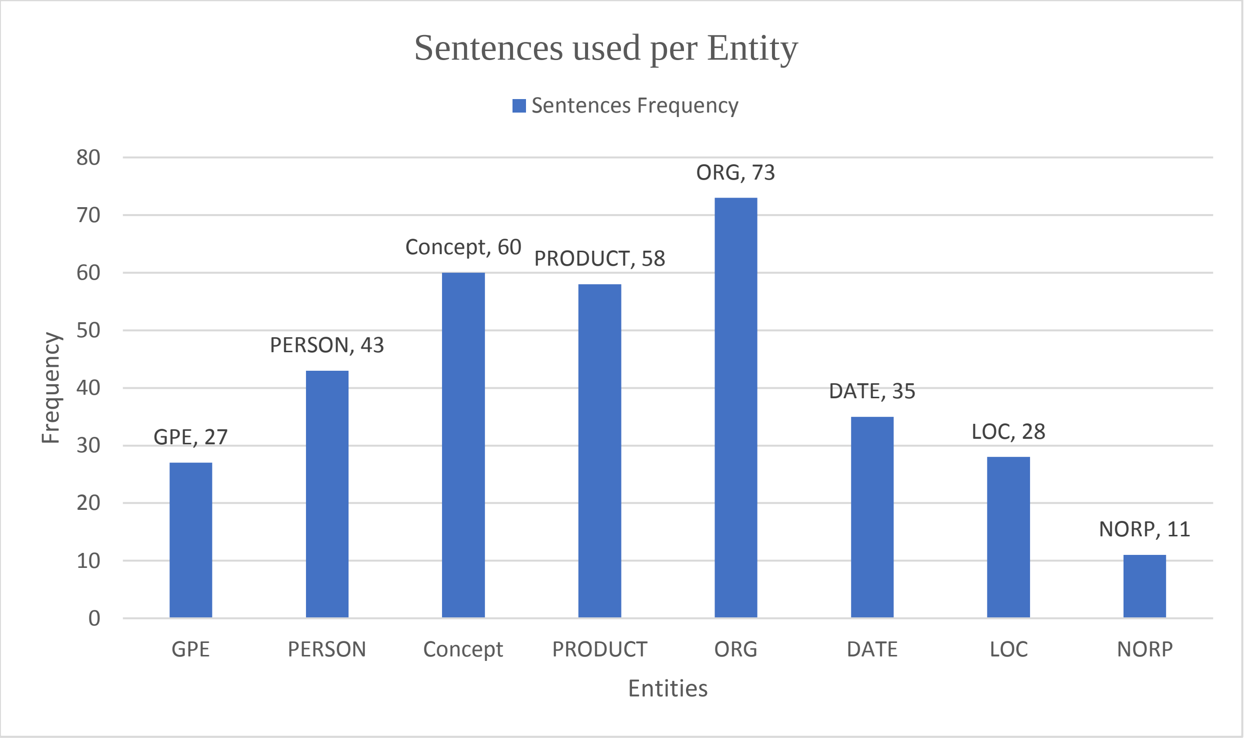


Figure 5.1: Shows Number of sentences used in each entity

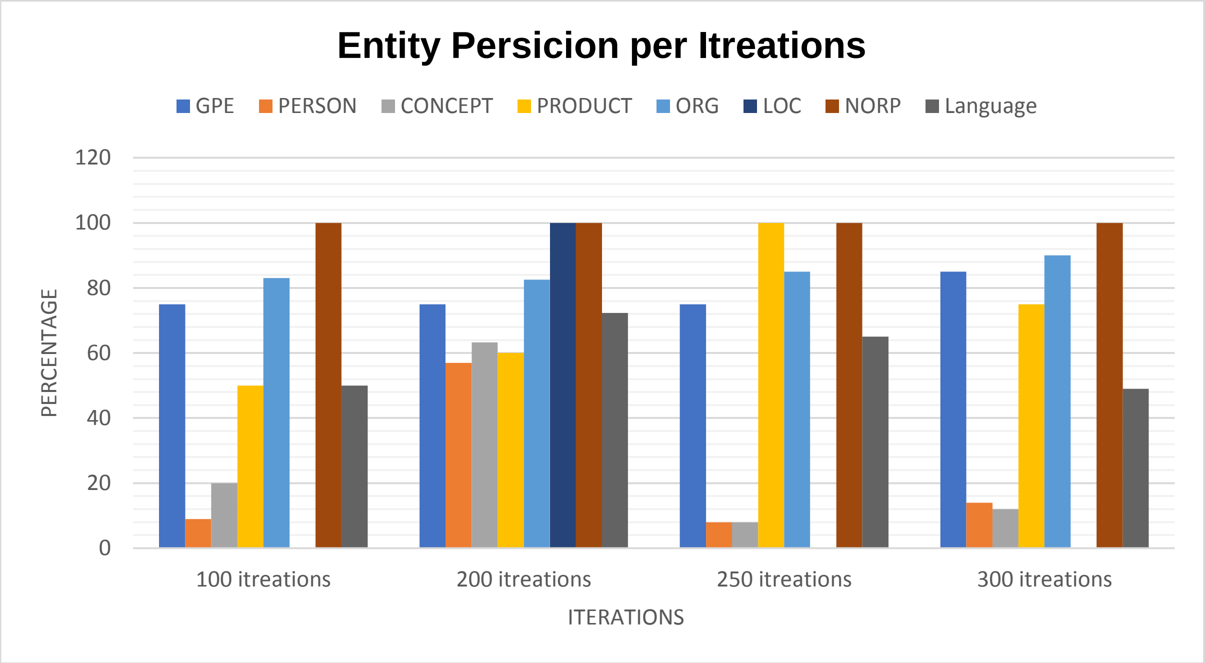


Figure 5.2: How number of iterations effect on the precision of the module

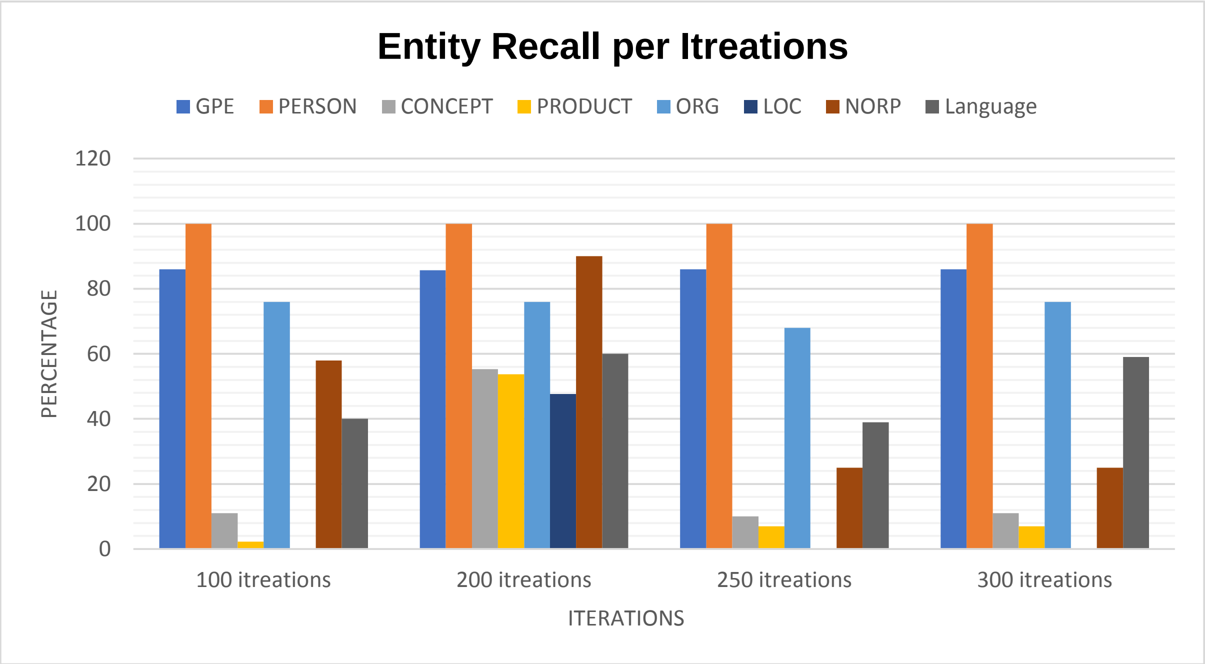


Figure 5.3: How number of iterations effect on the recall of the module

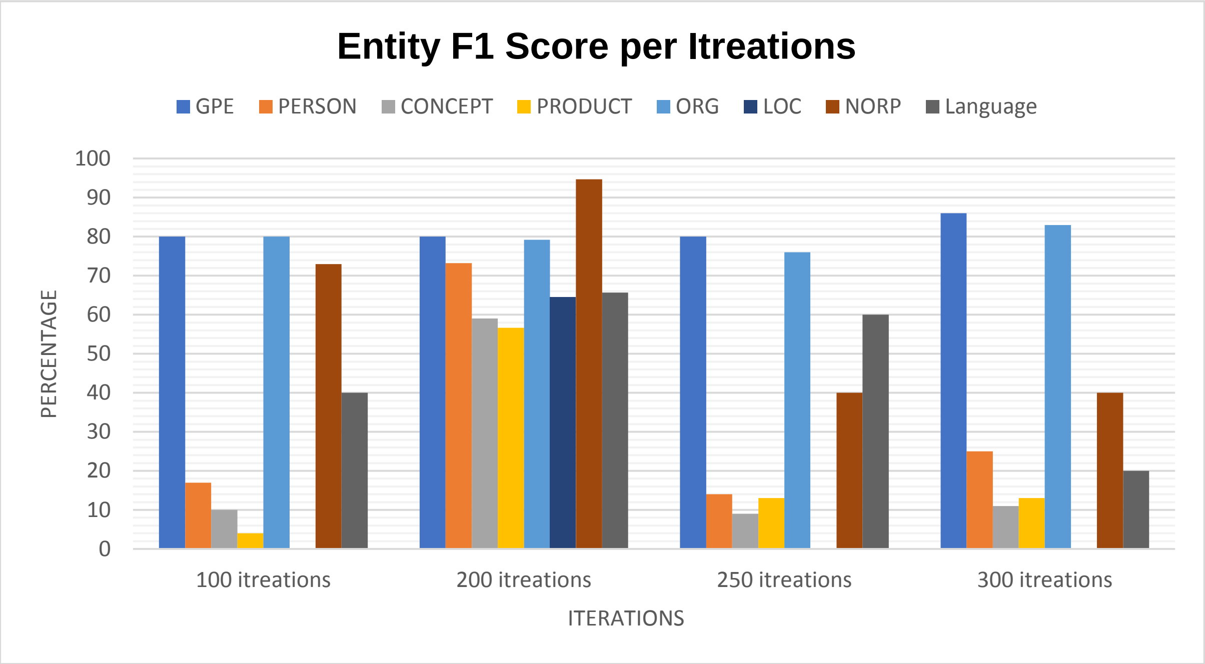


Figure 5.4: How number of iterations effect on the F1 score of the module

As shown above, the figures illustratehow the precision,recall, and f1 score change and get affected by increasing the number of iterations. When the number of iterations is increased that means the model is trained and might reach to overfitting level where the model is trained too much where it memories only the trained data so, the outcome results become unsatisfied. Concluding that increasing the number of iterations doesn’t always ensure better system outcomes.

What if the model is not trained enough meaning the number of iterations applied is not enough less than normal that leads the model to be underfitting so, the model is not trained well and might not iterate on all examples(trained data) where the model will not be able to learn or memories the data.

After many trails, the number of iterations which tend to better results is 200 iterations.put in the consideration that the results would get better when the number of training data set is increased. Afew hundreds is a good start however, more different sentences would make this model robust and more accurate.

B. Document Clustering

As the document clustering falls under unsupervised learning the accuracy or goodness of the module is determined using the silhouette coefficientor the silhouette score represented in Equation 4. It is a metric used to determine the goodness and how accurate the system.

$$\textbf{Silhouette Score} = (\mathbf{x} - \mathbf{y})/\mathbf{max}(\mathbf{x}, \mathbf{y}) \qquad \textbf{(4)}$$

Where:

a) **x:**The intermediate distance between each point within a cluster.

b) **y:**The intermediatedistance between all clusters.

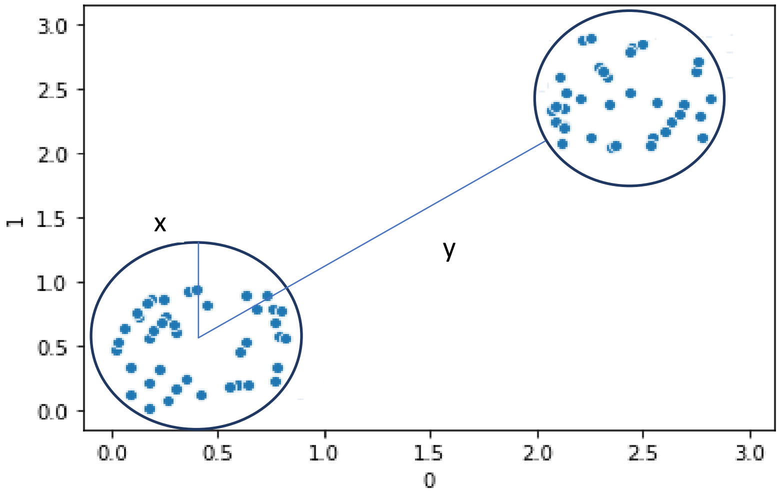


Figure 5.5: demonstrate what x and y refers between two clusters

The coefficient results always range from -1 to 1. The following interpretation clarify what each range stands for:

a) **1:** Means clusters are well apart from each other and clearly distinguished.
b) **0:** Means clusters are indifferent, the distance between clusters is not significant.
c) **-1:** Means clusters are assigned in the wrong way.

The silhouette coefficient or the silhouette score of this module is 0.75 which is close to 1 means that the clusters are well apart from each other.

5. Conclusion

This paper introduced an evaluation, result, and analysis of the architecture that concerns mining the publication papers using natural language processing and machine learning techniques. The introduced system consists of the following modules. First extracting keywords from the documents using natural language processing techniques. Second, a named entity recognition identifies the main theme for the documents using also machine learning algorithm. Third clustering documents using a machine learning algorithm. A plenty number of researches applying the text mining techniques for business and financial purposes, it used to predict the changes in various sectors through given data. However, the limitation of the researches that concerns mining the publication papers is high. Therefore, mining the publication papers is important where it has a great impact especially for researchers to keep them updated with the most recent researches that are related to a specific domain. This process would save the efforts and time for them.

References

1. S.-H. Liao, P.-H. Chu, and P.-Y. Hsiao, “Data mining techniques and applications–a decade review from 2000 to 2011,” Expert Systems with Applications, vol. 39, no. 12, pp. 11 303–11 311, 2012.

2. Avinesh, et.al, G. “Part of Speech Tagging and Chunking using Conditional Random Fields and Transformation Based Learning”. Proceedings of IJ-CAI Workshop on” Shallow Parsing for South Asian Languages. 2007.

3. Cutting, D., Kupiec, J., Pederson, J., and Sibun, P. “A Practical Part of Speech Tagger”. Proceedings of the Third Conference on Applied Natural Language Processing, Vol, 1992.

4. Berger, Adam L., Stephen A. Della Pietra Y and Vincent J. Della Pietra Y. A Maximum Entropy Approach to Natural Language Processing. Computational Linguistics (22): p. 39- 71, 1996.

5. Lafferty, et.al. “Conditional Random Field: Probabilistic Model for Segmenting and Labeling Sequence Data”.Proceedings of the Eighteenth International Conference on Machine Learning. 2001.

6. A. Henriksson, H. Moen, M. Skeppstedt and et.al, “Synonym extraction and abbreviation expansion with ensembles of semantic spaces,” Journal of biomedical semantics, vol. 5, no. 1, p. 1, 2014.

7. K. Sumathy and M. Chidambaram, “Text mining: Concepts, applications, tools and issues-an overview,” International Journal of Computer Applications, vol. 80, no. 4, 2013.

8. R. Rajendra and V. Saransh, “A Novel Modified Apriori Approach for Web Document Clustering,” International Journal of Computer Applications, pp. 159–171, vol.3, 2013.

9. N. Zhong, Y. Li, and S.-T. Wu, “Effective pattern discovery for text mining,” IEEE transactions on knowledge and data engineering, vol. 24, no. 1, pp. 30–44, 2012.

10. Roul and et.al . "A novel modified apriori approach for web document clustering." In Computational Intelligence in Data Mining-Volume 3, pp. 159-171. Springer, New Delhi, 2015.

11. Chau et.al.: Extracting meaningful entities from police narrative reports. In: Proceedings of the 2002 Annual National Conference on Digital Government Research, pp. 1–5. Digital Government Society of North America (2002)

12. K.Thilagavathiand et.al .“A Survey on Text Mining Techniques”, International Journal of Advanced Research in Computer Science and Robotics, ISSN: 2320 7345 Vol2, Issue 10, Oct. 2014 pp41-50

13. SIDDIQI, Sifatullah; SHARAN, Aditi. Keyword and key-phrase extraction techniques: a literature review. International Journal of Computer Applications, 2015, 109.2.

14. ZHANG, Chengzhi. Automatic keyword extraction from documents using conditional random fields. Journal of Computational Information Systems, 2008, 4.3: 1169-1180.

15. TARCZYNSKI, Tomasz. Document clustering-concepts, metrics and algorithms. International journal of Electronics and Telecommunications, 2011.

16. Mohdariff, et.al, M.I. (2018). Comparative Study of Document Clustering Algorithms. International Journal of Engineering and Technology (UAE). 7. 246-251. 10.14419/ijet. v7i4.11.20816.

17. YADAV, Vikas; BETHARD, Steven. A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470*, 2019.