IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 25, Issue 3, Ser. I (May. – June. 2023), PP 38-49
www.iosrjournals.org

Situation Dependent Uprooting Of Information From Unstructured Data Using Nlp 2 Stage Process

[1]Shankarayya Shastri, [2]Dr.Veeragangadharaswamy T.M
[1]*Research Scholar, Dept. of CSE, RYMCE, Ballari*
[2]*Professor, Dept. of CSE, RYMCE, Ballari*

ABSTRACT: *Now a days information is generating enormously using smartphones, laptops ,sensors, surveillance cameras, etc in various places such as Universities, Digital libraries, Data centres, Emails, Legal documents, social media, chatbots, Surveys, Medical fields, Sensors, Shopping malls, public and private sectors, etc .In digital world, It is estimated that around 80% of all information is unstructured. Unstructured means no format for stored data or information is in sentence format.*

Hence classification of unstructured text is becoming very important task of all type of organizations as it allows to easily do data analysis from unstructured data and automate real time on going processes.

Text classification is the process of categorizing text into different categories based on their subjects, by using Natural Language Processing (NLP).

Data analysis is a crucial process in the field of data science that extracts useful information from any form of data. The ease of access and maintenance makes structured data the most popular choice among many organizations even today. On the other hand, with the rapid growth of technology, more and more unstructured data, such as text and image, are being produced in large amounts. Unstructured data is data that does not have any pre-defined model associated with it. Due to the availability of a huge number of electronic text documents from a variety of sources representing unstructured and semi-structured information, the document classification task becomes an interesting area for controlling data behaviour. Hence in this paper we are presenting situation-based extraction of concepts from unstructured data using NLP techniques. The performance of presented approach is evaluated in terms of Precision, Recall and F1-score.

KEYWORDS: *Unstructured data, Keywords, Key phrases, concept, unigram, Context Relevance score, Central metrics.*

Date of Submission: 06-05-2023 Date of Acceptance: 16-05-2023

I. INTRODUCTION

With the increasing expansion and technological advancement, a vast volume of text data is generated every day in the form of social media platform, websites, company data, healthcare data, eucational fields, news, etc. Indeed, it is a difficult task to extract intriguing patterns from the text data, such as opinions, summaries, and facts, having varying length [6]. The exponential growth of textual content, as well as the ongoing expansion of the information age, makes handling such a massive amount of data much more challenging. Online textual content is either semi-structured or unstructured; examples include academic papers, online journals, news sources, and books. Prior to the development of technology, people could only process this large amount of data, which took a long time [1].

Data analysis and its complexity vary according to the type of data. The complexity of analysis is associated with several aspects such as data resources, the accuracy of analysis, and domain dependence. Structured data is akin to machine-language and makes operation and management of information much straight-forward; whereas unstructured data is usually natural language text with no strict semantic structure or database format. Evidently; if it was viable to instantly transform unstructured data to structured data, then comprehending intelligence from unstructured data would be simpler.

Unstructured data are irregular information with no predefined data model. Streaming data which constantly arrives over time is unstructured, and classifying these data is a tedious task as they lack class labels and get accumulated over time. As the data keeps growing, it becomes difficult to train and create a model from scratch each time. Due to the wide variety of the types of the documents circulating over the internet used in large scale of different applications, identifying the type of document is a critical task for the classification models in order to simplify further operations. Textual semi-structured and unstructured documents have many differences related to their nature which include the structure of the textual representation, degree of ambiguity, degree of redundancy, degree of using punctuation symbols and use of idioms and metaphors [9].

Key-phrases are crucial for searching and systematizing scholarly documents. A list of key-phrases is an important attribute of a scientific text. Key-phrases contain a brief representation of the contents of a text. They help search engines find and systematize the papers [4]. A qualitative selection of key-phrases positively affects a paper's visibility and its number of citations. Text classifiers can organize, arrange, and categorize almost any type of text, including documents, medical research, files, and text found on the internet. Unstructured data accounts for over 80% of all data, with text being one of the most common categories. Because analysing, comprehending, organizing, and sifting through text data is difficult and time-consuming due to its messy nature, most businesses do not exploit it to its full potential [2].

Text categorization is a fundamental task in the field of natural language processing (NLP) and is frequently utilized in the retrieval of information, unreliable analysis and identification, analysis of emotions, identification of spam emails, etc. Text mining is a term used to describe the method of extracting patterns or knowledge from unstructured texts. Text classification is a technique in which we have to extract useful information from text. This is where machine learning and text classification come into play. Text classifiers are used to classify quickly and cost-effectively arrange all relevant text types, including emails, legal documents, social media, surveys, and more. In the data learning and the prediction process, there are several models of classification techniques are to validate the data based on the similarity metrics. For scholars and other researchers needs to search for references that are related to the work and for the documentation process. For the purpose of document classification, most of the existing algorithms consider only the distribution of the content words of the document.

Hence in this work, Context aware extraction of concepts from unstructured data using NLP Techniques are described. The rest of the work is organized as follows: The section II describes the literature survey. The section III demonstrates situation based extraction of concepts from unstructured data using NLP Techniques. The section IV describes the result analysis of presented approach. Finally, the work is concluded in section V.

II. LITERATURE SURVEY

Dan Zhang et. al., [10] describes Text Complexity Classification Data Mining Model Based on Dynamic Quantitative Relationship between Modality and English Context. *is article first starts with the theoretical research of text complexity analysis and analyzes the source of text complexity and its five characteristics of dynamic, complexity, concealment, sentiment, and ambiguity, combined with the expression of user needs in the network environment. Secondly, based on the specific process of text mining, namely, data collection, data processing, and data visualization, it is proposed to subdivide the user demand analysis into three stages of text complexity acquisition, recognition, and expression, to obtain a text complexity analysis based on text mining technology. Experimental results show that the collected quantitative relationship information is identified and expressed in order to realize the conversion of quantitative relationship information into product features.

Dongdong Zhang, Changchang Yin, Jucheng Zeng, Xiaohui Yuan and Ping Zhang et. Al., [11] describes Combining structured and unstructured data for predictive models: a deep learning approach. In this research, authors presented 2 general-purpose multi-modal neural network architectures to enhance patient representation learning by combining sequential unstructured notes with structured data. Described fusion models leverage document embeddings for the representation of long clinical note documents and either convolutional neural network or long short-term memory networks to model the sequential clinical notes and temporal signals, and one-hot encoding for static information.

Fouad Zablith, Ibrahim H. Osman et. al., [14] presents Review Modus: Text Classification and Sentiment Prediction of Unstructured Reviews using a Hybrid Combination of Machine Learning and Evaluation Models. Review Modus, a text mining and processing framework that (1) relies on the model structure and its corresponding assessment questions to train a machine learning algorithm to predict the classification of reviews around the model dimensions; (2) predicts the sentiments within the reviews based on external review training datasets; and (3) transforms the extracted measures from the reviews for further analysis. approach is evaluated in the context of 11 e-Government services where the performance of the framework is compared to the manual processing of unstructured reviews cross-checked by three independent evaluators.

Sathya Madhusudhanan, Suresh Jaganathan and Jayashree L S et. Al., [15] presents Incremental Learning for Classification of Unstructured Data Using Extreme Learning Machine. Describes a framework CUIL (Classification of Unstructured data using Incremental Learning) which clusters the metadata, assigns a label for each cluster and then creates a model using Extreme Learning Machine (ELM), a feed-forward neural network, incrementally for each batch of data arrived. Based on the tabulated results, this work proves to show greater accuracy and efficiency. However this work has certain limitations: (1) difficulty in fixing the random weights by trial and error method until the desired accuracy is achieved for the training dataset and (2) difficulty in choosing the number of hidden neurons i.e., higher accuracy is achieved when the number of hidden neurons increases.

Lipika Dey, Hardik Meisheri and Ishan Verma et. al., [17] presents Predictive Analytics with Structured and Unstructured data - A Deep Learning based Approach. A generic deep learning framework is presented for predictive analytics utilizing both structured and unstructured data. They also present a case-study to validate the functionality and applicability of the proposed framework where we use LSTM for prediction of structured data movement direction using events extracted from news articles.

Shucheng Gong and Hongyan Liu et. al., [19] Constructs Decision Trees for Unstructured Data. a decision tree construction algorithm called CUST was described, which can directly tackle unstructured data. CUST introduces the use of splitting criteria formed by unstructured attribute values, and reduces the number of scans on datasets by designing appropriate data structures. Experiments on real-world datasets show that CUST improves the efficiency of building classifiers for unstructured data.

III. SITUATION DEPENDENT UPROOTING OF INFORMATION FROM UNSTRUCTURED DATA USING NLP 2 STAGE PROCESS

Situation dependent uprooting of information from unstructured data using using NLP 2 stage process is presented in this section. The flow diagram of presented approach is shown in Fig. 1.
The main aim is to satisfy the objectives that are listed as follows:

1. To Extract situation-based information from unstructured data
2. To identify and extract similar or overlapping words from unstructured data

Unstructured data, typically text are data that does not have a predefined format (e.g., e-mail, word processing documents or presentations). The unstructured text is generated and collected in wide range of forms including word documents, email messages, power point presentations, survey responses, transcripts of call center interactions, posts from blogs and social media sites. Other types of unstructured data include images, audio and video files. The unstructured documents are just that documents that can be free from and don't have set structure but are still able to be scanned, captured and imported.

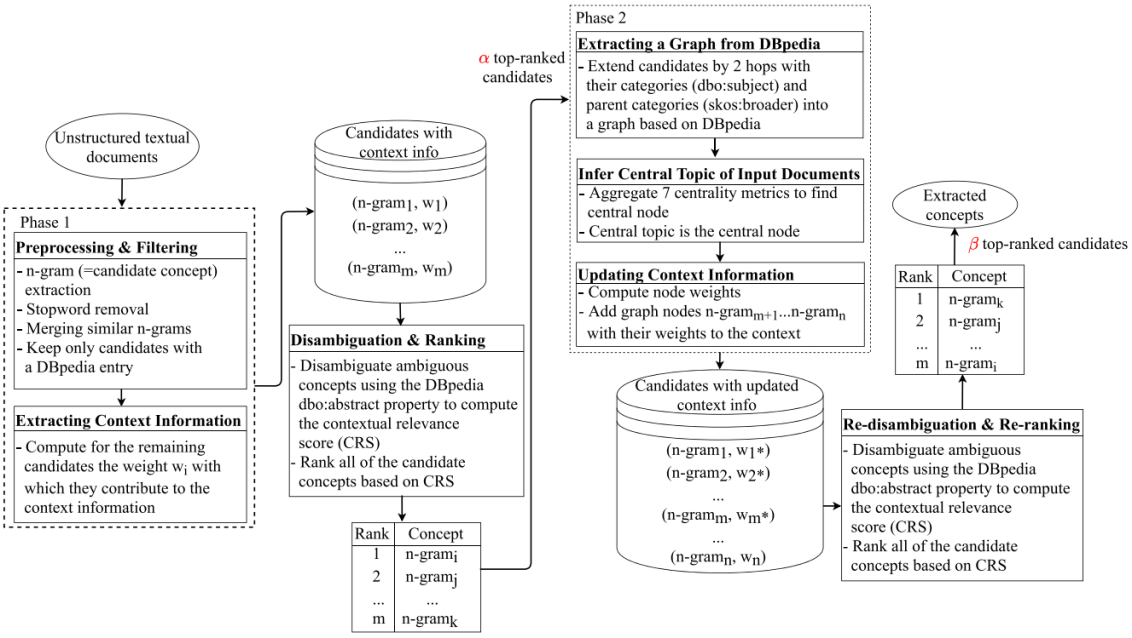


Figure 1: Workflow diagram for Situation dependent uprooting of information from unstructured data

In this paper, for experimental purpose we are taking sample input data set as computer science related Database, Operating systems, Data Mining unstructured text documents (.txt format more than 300 text files files).

The input unstructured documents are need to be pre-processed by applying tokenization, stop word removal, infrequent word removal functions. The pre-processing stage removes missing or inconsistent data values occurred due to human or computer errors, Pre-processing stage can improve the accuracy and quality of dataset more accurate, reliable and consistent.

The work of Situation dependent uprooting of information from unstructured data is divided into two stages. Stage 1 and 2 step by step operations are given below

Stage 1: Pre-Processing and Filtering candidate Concepts

Given a set of unstructured text documents D={d1,d2,d3……dn}that describes a set of concepts C={c1,c2,c3,….cm} where n>1 and m>1
Goal is to identify and classify most relevant concepts from C.

Step 1:unstructured textual documents diϵD are tokenized into n-grams representing the initial candidate concepts ciϵC.

diЄ D => Tokenization(c1,c2,c3….cm) to form n-grams

Step 2: In n-grams, stopwords are removed using a stopword list comprising common terms and Count the occurrences of the remaining candidates ci in D yields a set of tuples s τ={(ci,fi),….} comprising a candidate ci and its respective frequency fi.

Step 3: Remove short, infrequent n-grams from T (and therefore also from C) according to a frequency threshold ft to reduce noise.

Step 4: Retain only meaningful unigrams (n>=1)
if and only if it occurs at least ft times more often than any larger n-gram which contains this unigram somewhere. (Refer Figure 2)

Step 5: If there are multiple n-grams cj of higher order that contain ci,fi refers to most frequently occurring cj. The remaining n-grams from C are merged according to two rules, s.t. only a single n-gram is present for semantically similar n-grams.
a)First, a plural token is filtered out if it is also present in singular form. (Refer Figure 2)
b)Second, the present participle of a regular verb is discarded, if a version without it exists. (Refer Figure 2)

Step 6: From the remaining candidates in C, filter out those with no matching DBpedia entry.

Step 7: Generate initial context information Cinfo,

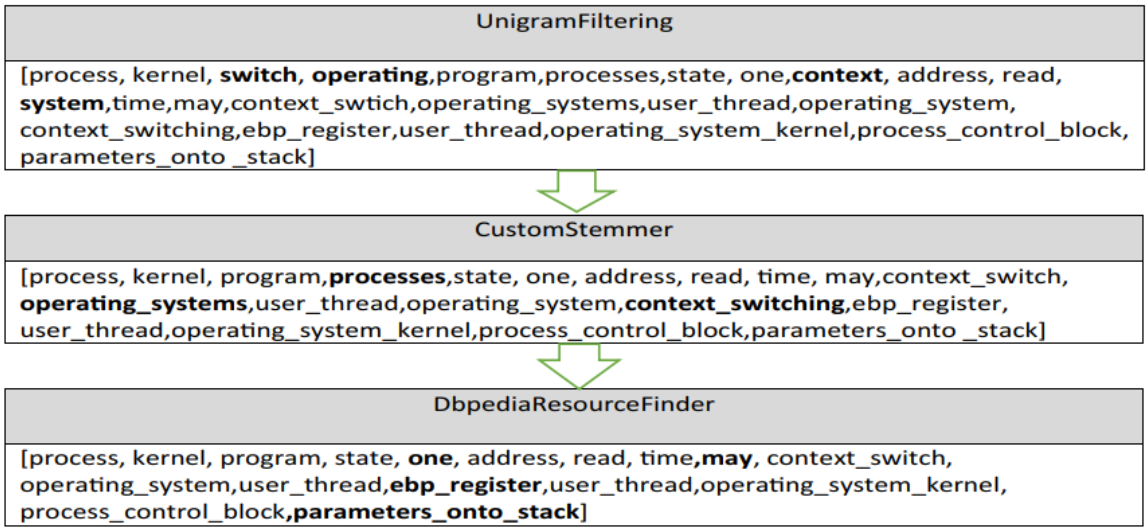


Figure 2: Filtering process for a set of candidate concepts sampled from a tokenized input document about ‘‘Operating Systems”. Candidates in bold font are filtered out at their respective stage: first unigrams are removed if they are part of longer frequent n-grams, then semantically similar n-grams are merged after data cleaning and lastly n-grams with no DBpedia entry are discarded.

Stage 2: Ranking and disambiguating process

Step 1: Assume that initial context information Cinfo generated from stage 1 describes the underlying topic T

Step 2: Calculate and prepare a list of candidate concepts such that they describe the context information. Cinfo=(ci,wi), where ci represents a candidate concept from C and weight wi its contribution to the context information. wi is calculated as:

wi = |ci|

Step 3: Disambiguates candidate concepts by computing the Context Relalevance Score (CRS) as follows:

$$CRS_i = \frac{\sum_{c_j \in C_{info} \cap DBP_i} w_j}{|DBP_i|}$$

Step 4: All candidates are mapped to exactly one DBpedia entry, they are ranked w.r.t. the context information according to CRSi

Step 5: More abstract n-grams from DBpedia are added to the context information

Step 6: Inferring the Central Node

$$CS_c = \sum_{i \in CM} \frac{CM_i(c)}{rank_{ci}}$$

Step 7: Updating Context Information

Centrality metrics we adopt

1. Degree centrality,
2. Katz Centrality,
3. Eigenvector Centrality,
4. PageRank Centrality,
5. Betweenness Centrality,
6. Closeness Centrality and
7. Information Centrality

IV. RESULT ANALYSIS

In this section, Situation dependent uprooting of information from unstructured data using NLP technique is implemented. In this approach three different datasets namely Database, Data Mining and Operating System datasets are used for experiment purpose. The performance of presented approach is investigated with different existing available approaches.

The result analysis of presented approach is evaluated using the parameters like, true positive rate (sensitivity), precision, recall and F1-score with the reference of Ground truth of database. Precision.

Precision: The precision is employed to calculate the positive patterns which are predicted correctly from the total predicted patterns in a positive class.

$$\text{Precision}=\text{Key}_{\text{Corrected}}/\text{Key}_{\text{Predicted}} \quad (1)$$

where $\text{Key}_{\text{corrected}}$ is the total correctly predicted key-phrases that are matched with standard key-phrases and $\text{Key}_{\text{predicted}}$ is the total predicted key-phrases from a document.

Recall: It can also known as Sensitivity. It is the ratio of accurately expected positive values with respect to the actual positive values; and can be calculated as

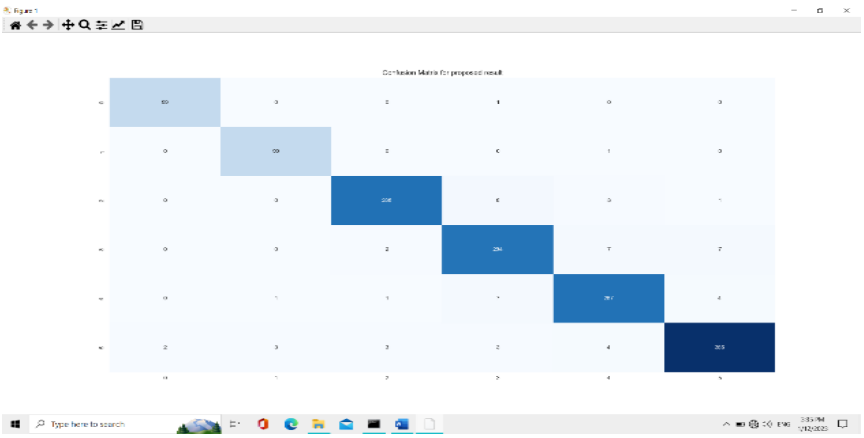
$$\text{Recall}=\text{Key}_{\text{Corrected}}/\text{Key}_{\text{Predicted}} \quad (2)$$

F1-Score: F1-score is one of the most important evaluation metrics in machine learning. It elegantly sums up the predictive performance of a model by combining two otherwise competing metrics precision and recall.

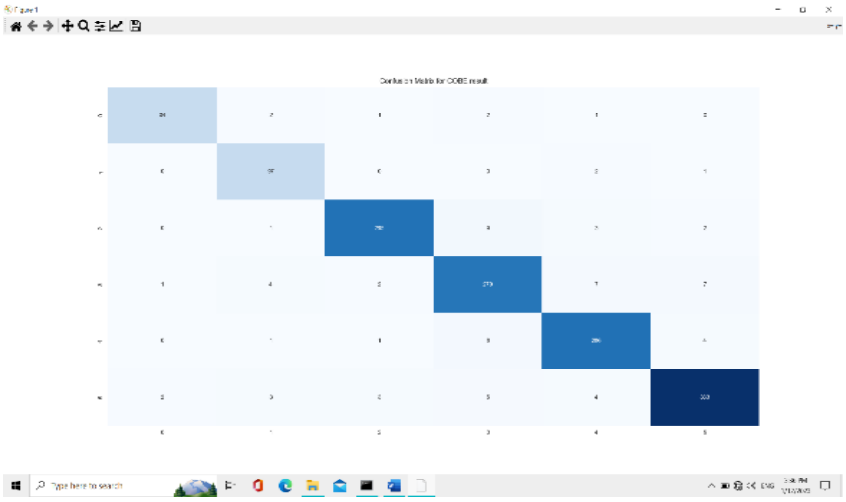
$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

.

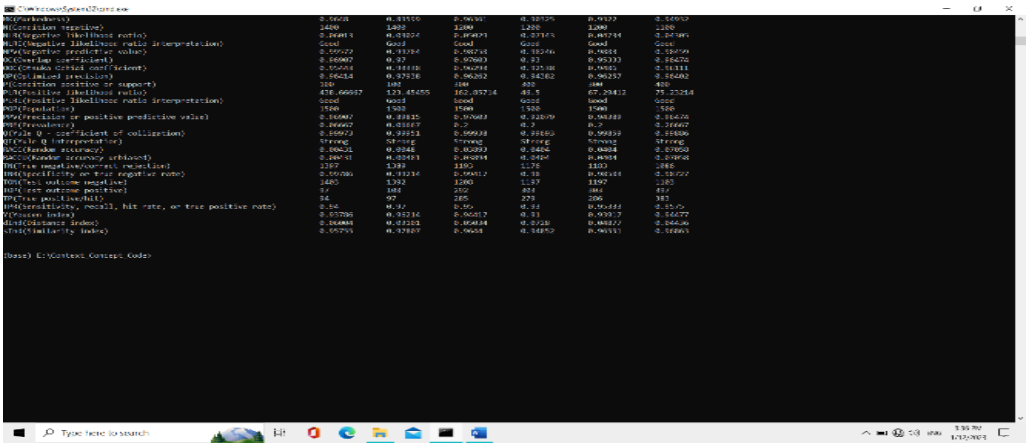
The Fig. 3 (a) shows the confusion matrix for presented approach and 3(b) shows the confusion matrix for COBEC approach. The x-axis indicates different methods and y-axis indicates the performance rate. Presented approach using SemEval 2010 dataset has high performance than KCFA approach



3.(a)



3.(b) **Fig. 3: Confusion Matrix for (a) Presented Matrix and (b) COBEC Approach**



The Fig. 4 shows the output screen of implemented Situation dependent uprooting of information from unstructured data.

The table 1 shows the performance evaluation of SemEval2010 dataset.

Methods	F1-score	Precision	Recall
KCFA	0.9051	1	0.8627
Presented Context aware extraction of concepts from unstructured data using ML algorithms	0.9248	1	0.8863

Table 1: Performance Evaluation on SemEval2010 dataset

Compared to KCFA, presented approach has better results in terms of precision, recall and F1-score.

The Fig. 4 shows the performance metrics comparison.

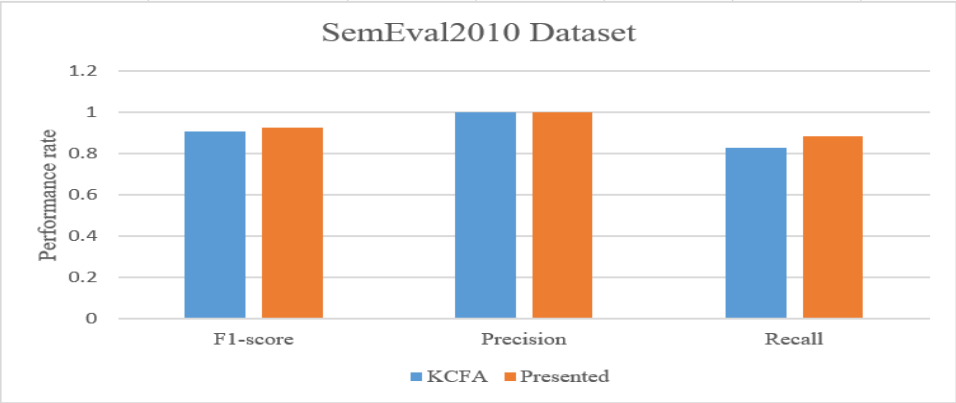


Fig. 5: SemEval2010 dataset performance metrics comparison

The table 2 represents the performance metrics evaluation on Wiki20 dataset.

Table 2: Performance Evaluation on Wiki20 dataset

Methods	Precision	Recall	F1-score
KCFA	1	0.4437	0.6146
Presented approach using Wiki20 dataset	1	0.8652	0.8347

Presented approach using Wiki20 dataset has better results in terms of precision, recall and F1-score than KCFA. The Fig. 5 shows the performance metrics comparison using Wiki20 dataset.

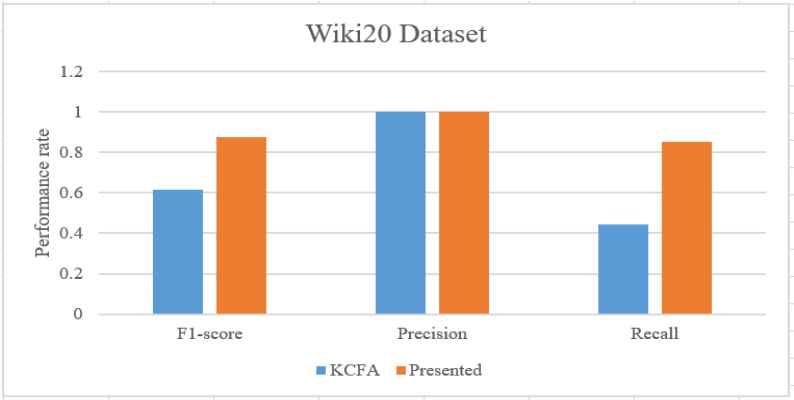
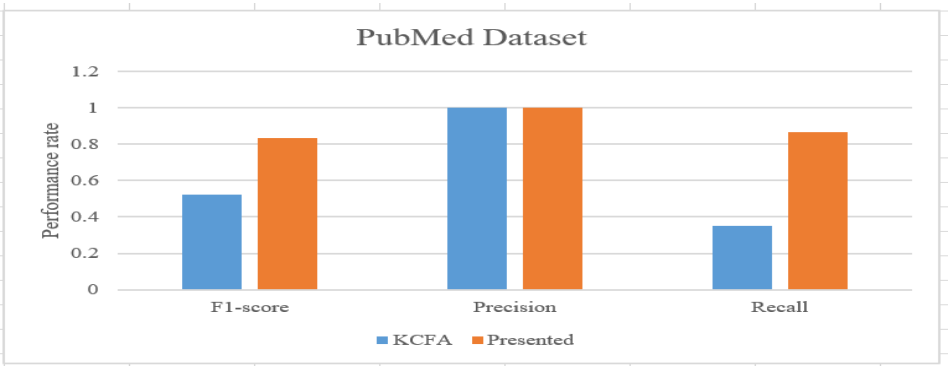


Fig. 6: Wiki20 dataset performance metrics comparison

Compared to KCFA, presented approach using Wiki20 dataset has better recall, better F1-score and equal precision. The Table 3 represents the performance evaluation on PubMed dataset.

Methods	Precision	Recall	F1-score
KCFA	1	0.3529	0.5217
Presented Context aware extraction of concepts from unstructured data using ML algorithms	1	0.8652	0.8347

Table 3: Performance Evaluation on PubMed dataset



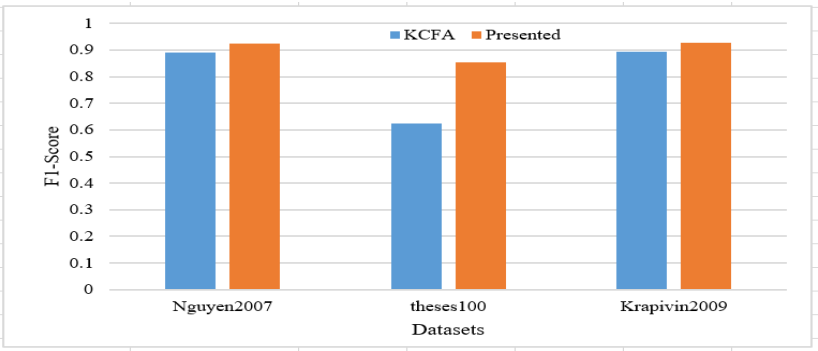
The performance evaluation of presented approach on PubMed dataset has high recall, high f1-score than KCFA. The Fig. 6 shows the graphical representation.

The table 4 indicates the performance of KCFA and Presented approach using different datasets like Nguyen2007, Theses100 and Krapivin2009.

Methods	Nguyen2007	Theses100	Krapivin2009
KCFA	0.8895	0.6233	0.8937
Presented	0.9247	0.8546	0.9262

Table 4: F1-Score comparison

Presented approach using Krapivin2009 dataset has high F1-score. The performance comparison of three different datasets is shown in below Fig. 7.



The table 5 represents the recall performance of KCFA and presented approaches using Nguyen2007, Theses100 and Krapivin2009 datasets.

Table 5: Recall Performance Comparison

Methods	Nguyen2007	Theses100	Krapivin2009
KCFA	0.801	0.4528	0.8078
Presented	0.9126	0.8612	0.8463

Presented approach has high recall using Nguyen2007 dataset while the KCFA approach has high recall using krapivin2009. The fig. 8 shows the graphical representation of recall performance comparison.

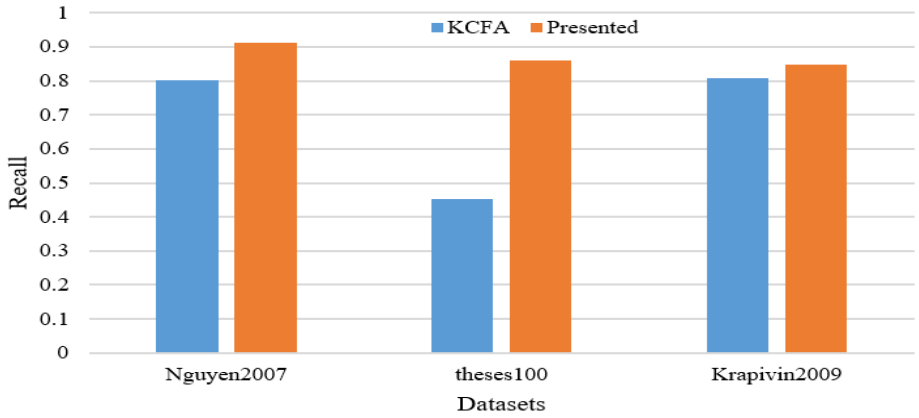


Fig. 9: Comparison Graph for Recall

Presented approach with Nguyen2007 dataset has high recall than KCFA with Nguyen2007 dataset. The theses100 and karipivin2009 datasets also has better recall for presented approach. The table 6 shows the Macro-averaged recall comparison between different datasets using different approaches.

Datasets	**COBEC (1)**	**COBEC (2)**	**COBEC -T(1)**	**COBEC -T(2)**	**Presented**
SemEval	0.164	0.127	0.14	0.098	0.6964
DM	0.232	0.249	0.198	0.21	0.6722
Wiki20	0.125	0.138	0.122	0.135	0.706
OS	0.302	0.348	0.209	0.255	0.6444
DB	0.272	0.342	0.253	0.313	0.68888
Theses100	0.097	0.192	0.082	0.166	0.7325
Nguyen2007	0.204	0.257	0.191	0.238	0.6718

Table 6: Comparison of Macro-averaged Recall

The theses100 dataset has high macro-averaged recall among other datasets (SemEval, DM, Wike20, OS, DB and Nguyen2007) for presented approach. The Fig. 9 shows the recall performance comparison of different approaches using different datasets.

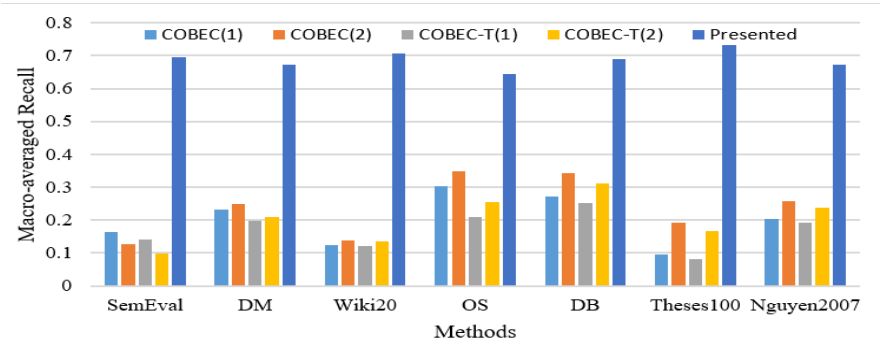


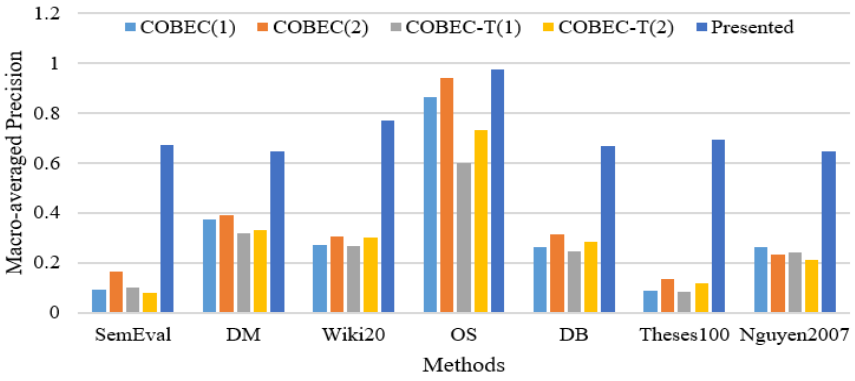
Fig. 10: Comparison graph for Macro-averaged Recall

The table 7 represents the macro-averaged precision performance evaluation.

Datasets	COBEC (1)	COBEC (2)	COBEC -T(1)	COBEC -T(2)	Presented
SemEval	0.094	0.164	0.102	0.078	0.673
DM	0.373	0.393	0.32	0.333	0.645
Wiki20	0.273	0.306	0.266	0.3	0.772
OS	0.866	0.943	0.6	0.733	0.976
DB	0.265	0.314	0.248	0.285	0.667
Theses100	0.087	0.133	0.083	0.12	0.692
Nguyen2007	0.0264	0.234	0.243	0.213	0.647

Table 7: Comparison of Macro-averaged Precision

Compared to different datasets, The OS dataset has high precision for presented approach. The Fig. 10 shows the graphical representation of Macro-averaged precision.

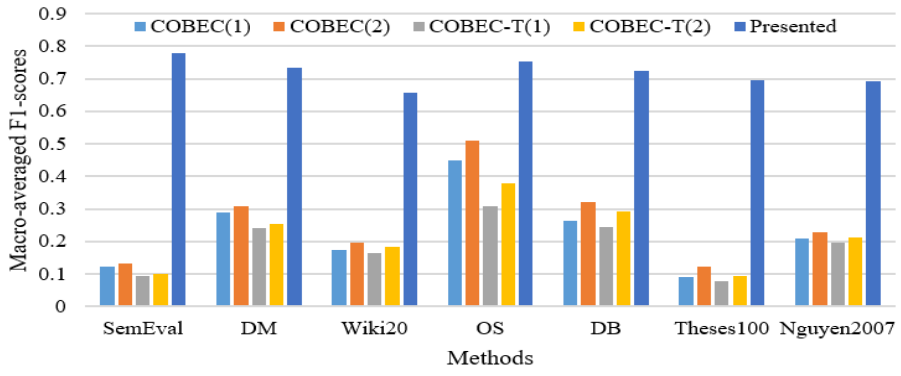


The OS dataset has better precision for all approaches and in addition it has high precision for presented approach. The Table 8 compares the Macro-averaged F1-score of different datasets for different approaches.

Datasets	COBEC (1)	COBEC (2)	COBEC-T (1)	COBEC-T (2)	Presented
SemEval	0.124	0.132	0.094	0.1	0.7786
DM	0.288	0.307	0.242	0.254	0.734
Wiki20	0.174	0.197	0.166	0.184	0.6578
OS	0.448	0.51	0.31	0.38	0.7534
DB	0.263	0.321	0.244	0.293	0.7243
Theses100	0.0906	0.123	0.078	0.093	0.695
Nguyen2007	0.209	0.227	0.196	0.212	0.693

Table 8: Comparison of Macro-averaged F1-score

Among these datasets, the SemEval dataset has high macro-averaged F1-score for presented approach. The Fig. 11 shows the comparison graph for macro-averaged F1-score.



The SemEval dataset has high Macro-averaged F1-score than other datasets. The Fig. 12 shows the ROC (Receiver Operating Characteristics) curve comparison between COBEC and Presented Context aware extraction of concepts from unstructured data using Machine Learning algorithms.

In fig. 12, the red color curve line indicates the ROC of COBEC approach whereas Blue colour line indicates the ROC of presented approach. Compared to COBEC approach, presented Context aware extraction of concepts from unstructured data using Machine Learning algorithms has better ROC. The Fig. 3 shows the performance comparison in terms of Accuracy, Kappa Coefficient, Sensitivity, Specificity and Macro F1-score.

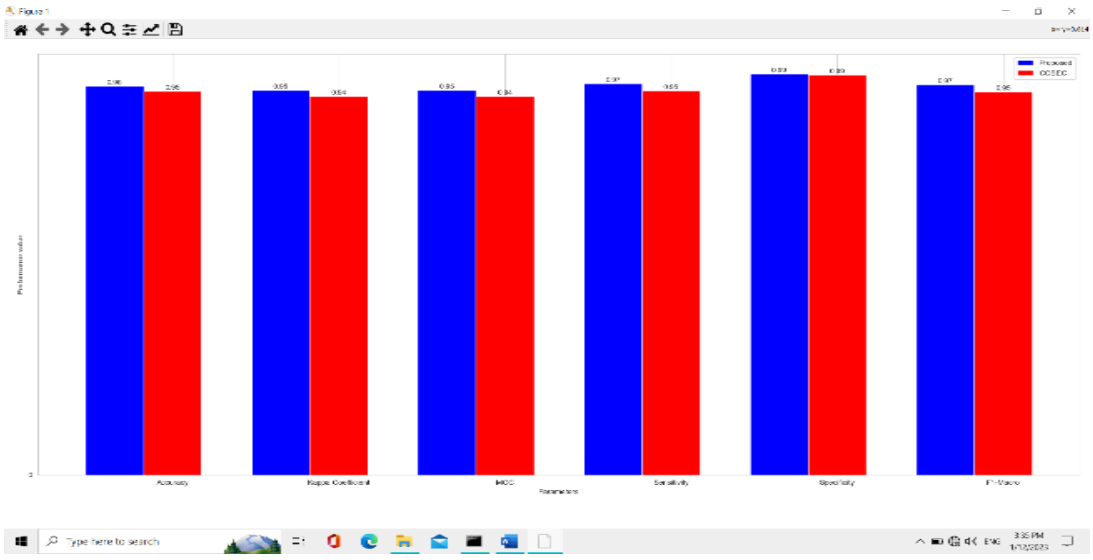


Fig. 14: Performance Comparison between COBEC and Presented Approach

In fig. 13 the blue colour indicates presented Context aware extraction of concepts from unstructured data using Machine Learning algorithms approach and red colour indicates COBEC Approach. The Context aware extraction of concepts from unstructured data using Machine Learning algorithms has high accuracy, Kappa Coefficient, Sensitivity, Specificity and Macro F1-score than COBEC Approach.

V. CONCLUSION

In this work, Situation dependent uprooting of information from unstructured data using NLP technique is presented. In this analysis, different datasets namely Database, Data mining and Operating systems datasets are used. In this approach, .txt files are taken as the input. The unstructured text documents are pre-processed to remove the unnecessary data and to clean the data. Feature database creates the subfolder to store the classification results. This approach has classified different documents and their domains where they belong to. The performance of presented approach is measured in terms of Precision, Recall and F1-score. Different datasets are

used to investigate the performance of presented approach. The performance of presented approach is compared with three datasets namely SemEval2010, Wiki20 and Pubmed datasets, however better results are achieved through SemEval 2010 dataset. In addition Macro-averaged F1-Score, Macro-averaged recall and Macro-averaged precision are investigated with different methods. Nguyen2007 dataset has high recall, theses100 dataset has high macro-averaged recall and OS dataset has high macro-averaged precision.

REFERENCES

[1]. Mohammad Badrul Alam Miah, Suryanti Awang, Md Mustafizur Rahman, A. S. M. Sanwar Hosen and In-Ho Ra, "A New Unsupervised Technique to Analyze the Centroid and Frequency of Key phrases from Academic Articles", Electronics 2022, 11, 2773, doi:10.3390/electronics11172773

[2]. Abdullah Alqahtani, Habib Ullah Khan2, Shtwai Alsubai, Mohemmed Sha, Ahmad Almadhor, Tayyab Iqbal4 and Sidra Abbas, "An efficient approach for textual data classification using deep learning", Frontiers in Computational Science, 2022, DOI 10.3389/fncom.2022.992296

[3]. Dmitriy Dligach, Timothy Miller, Steven Bethard and Guergana Savova, "Exploring Text Representations for Generative Temporal Relation Extraction", Proceedings of the 4th Clinical Natural Language Processing Workshop, pages 109 - 113 July 14, 2022, 2022 Association for Computational Linguistics

[4]. Anna Glazkova and Dmitry Morozov, "Applying Transformer-based Text Summarization for Keyphrase Generation", 2022, DOI:10.48550/arXiv.2209.03791

[5]. Miah, M.B.A.; Awang, S.; Azad, M.S.; Rahman, M.M, "Keyphrases Concentrated Area Identification from Academic Articles as Feature of Keyphrase Extraction: A New Unsupervised Approach", Int. J. Adv. Comput. Sci. Appl. 2022, 13, pp. 788-796.

[6]. Menghan Zhang, "Applications of Deep Learning in News Text Classification", Hindawi Scientific Programming Volume 2021, Article ID 6095354, 9 pages, doi:10.1155/2021/6095354

[7]. Shubham Jain, Amy de Buitléir, Enda Fallon, "A Framework for Adaptive Deep Reinforcement Semantic Parsing of Unstructured Data", 2021 International Conference on Information and Communication Technology Convergence (ICTC),**DOI:** 10.1109/ICTC52510.2021.9620904

[8]. Anja Wilhelm and Wolfgang Ziegler, "Extending semantic context analysis using machine learning services to process unstructured data", SHS Web of Conferences 102, 02001 (2021), doi:10.1051/shsconf/202110202001ELTC2021

[9]. Nany Katamesh, Osama Abu-Elnasr and Samir Elmougy, "Deep Learning Multimodal for Unstructured and Semi-Structured Textual Documents Classification", Computers, Materials & Continua,DOI:10.32604/cmc.2021.015761

[10]. Dan Zhang, "Text Complexity Classification Data Mining Model Based on Dynamic Quantitative Relationship between Modality and English Context", Hindawi Mathematical Problems in Engineering Volume 2021, Article ID 4805537, 10 pages, doi:10.1155/2021/4805537

[11]. Dongdong Zhang, Changchang Yin, Jucheng Zeng, Xiaohui Yuan and Ping Zhang, "Combining structured and unstructured data for predictive models: a deep learning approach", BMC Medical Informatics and Decision Making", (2020) 20:280, doi:10.1186/s12911-020-01297-6

[12]. Ahmed Ghozia, Gamal Attiya, Emad Adly and Nawal El-Fishawy, "Intelligence Is beyond Learning: A Context-Aware Artificial Intelligent System for Video Understanding", Hindawi Computational Intelligence and Neuroscience Volume 2020, Article ID 8813089, 15 pages, doi:10.1155/2020/8813089

[13]. Tushar Ghorpade, Bhavika Tuteja, Vaibhav Dholam, Gauri Patil, Ashutosh Bhujbal, "Learning of Unstructured Data Using Machine Learning Algorithm", International Journal Of Information And Computing Science, 2019, ISSN NO: 0972-1347, Volume 6, Issue 4, April 2019

[14]. Fouad Zablith, Ibrahim H. Osman, "Review Modus: Text Classification and Sentiment Prediction of Unstructured Reviews using a Hybrid Combination of Machine Learning and Evaluation Models", Applied Mathematical Modelling (2019), doi:10.1016/j.apm.2019.02.032

[15]. Sathya Madhusudhanan, Suresh Jaganathan and Jayashree L S, "Incremental Learning for Classification of Unstructured Data Using Extreme Learning Machine", MDPI Journal Algorithms 2018, 11, 158; doi:10.3390/a11100158

[16]. Mona Mowafy, A. Rezk, H. M. El-bakry, "Building Unstructured Crime Data Prediction Model", International Journal of Computer Application (2250-1797) Issue 8 Volume 4, July-August 2018, doi:10.26808/rs.ca.i8v4.01

[17]. Lipika Dey, Hardik Meisheri and Ishan Verma, "Predictive Analytics with Structured and Unstructured data - A Deep Learning based Approach", IEEE Intelligent Informatics Bulletin December 2017 Vol.18 No.2

[18]. Li Guo a, Feng Shi, Jun Tu, "Textual analysis and machine leaning: Crack unstructured data in finance and accounting", The Journal of Finance and Data Science 2 (2016) 153e170, doi.org/10.1016/j.jfds.2017.02.001

[19]. Shucheng Gong and Hongyan Liu, "Constructing Decision Trees for Unstructured Data", International Conference on Advanced Data Mining and Applications, ADMA 2014: Advanced Data Mining and Applications pp 475–487

[20]. Hassanin M. Al-Barhamtoshy and Fathy E. Eassa, "A data Analytics for unstructured Text", Life Science Journal 2014;11(10).