University of Sindh Journal of Information and Communication Technology (USJICT)
Volume 6, Issue 3, September 2022
ISSN-E: 2523-1235, ISSN-P: 2521-5582
Website: http://sujo.usindh.edu.pk/index.php/USJICT/

Towards the Optimal Use of Machine Learning Algorithms in Text Mining: A Quick Review

Syed Zafar Ali Shah, Sadaqat Jan and Ibrar Ali Shah

Department of Computer Software Engineering
University of Engineering &Technology Mardan, Pakistan
zafar@uetmardan.edu.pk, sadaqat@uetmardan.edu.pk, ibrar@uetmardan.edu.pk

Abstract: This paper aims to provide a quick review to jump-start the research in the field of text mining where Machine Learning (ML) algorithms have been used and several accomplishments have been reported by the research community. There are different categories of text mining, and the implementation of ML algorithms and techniques have been supported in the literature to give promising results. However, in this area of study, most of the research activities in terms of time and efforts are consumed during the initial stages where implementations and experiments are carried out to evaluate various combinations. The accomplishments in this field can be further advanced by presenting early investigations concisely and analytically. Thus, the benefits of this paper are threefold: first, it will provide a platform for the new researchers to start quickly with a shorter literature review and knowing more precisely about the combinations of text mining and ML; secondly, clear analysis has been presented about the text mining categories where the performance of ML algorithms have been reported successful; and lastly, the problems have been identified for which the algorithms were used in various studies. This will enable the new researchers to directly target the problem instead of implementing the existing techniques. With the help of well-structured questions, the results are more analytical and present multidimensional views to this research issue. Main findings include that ML has been widely used in document classification and Support Vector Machine (SVM) is the most successful algorithm reported.

Keywords: text mining; machine learning; review; sentiment analysis; support vector machine;

I. INTRODUCTION

Internet usage is exponentially increasing every year and so is its importance. The social media, electronic gadgets, and academic activities share a huge amount of data and has rapidly increased in the last decade. From learning to entertainment, social networking to marketing the easiest and most commonly used medium of communication is text [1]. Due to this large amount of textual data, it is very hard for an individual to process it manually. In order to classify and extract interesting and relevant information from this large pool of data, experts develop several text mining techniques for classification, clustering, feature extraction, information retrieval, natural language processing, text summarization, emotion detection, topic modeling, concept extraction and many more. Text mining is the process of extracting meaningful information and non-trivial patterns from unstructured, semi-structured and structured textual data [2,3].

Machine learning has successfully been used for medical information retrieval systems and clams the limitless capabilities for future systems [4]. It is pertinent to mention that text mining is a vast field that includes different segments like sentiment analysis then its variations like bipolar or multipolar sentiment analysis. On the other hand, there is an enormous number of machine learning algorithms and their combinations. Many ML algorithms and techniques have been used successfully and reported in the field of text mining. However, it has been observed that most of the efforts and time are consumed by such implementations or the combination of these implementations in the early stage of the research.

Using text mining the trends in machine learning research for the last three decades have been presented by [5] and found it useful for the upcoming researchers to get an intuition of trends to their area of interest. [6] covered more than 50 research articles and indicate the importance of input data in the performance of an algorithm in the domain of text mining and natural language processing. Although many systematic literature reviews (SLR) has been conducted to find the impact of machine learning and text mining in diverse fields like software engineering [7,8], medical healthcare [9,10,11], safety and security [12,13], material sciences [14], online discussion forum [15,16], quality enhancement [17,18], social media [19,20,21], legal issues [22,23], academics [24].

The limitation of the previous works is that there is no SLR focused on the effects of machine learning on different areas of text mining i.e. document categorization, document clustering, concept extraction, sentiment analysis, text summarization, entity relation modeling, emotion intensity detection, review or opinion mining etc. This SLR aims to study the impact of machine learning on the above sub-areas of text mining, the ML algorithms used, finding the best

performing algorithm, category of ML algorithms (supervised, semi-supervised, unsupervised, reinforcement) etc.

The rest of this paper is organized as follows. Section 2 discusses the research methodology. Section 3 explains the results and discussion. Section 4 presents the limitations and future work, and Section 5 is the conclusion followed by references.

II. METHODOLOGY

For conducting and reporting this systematic review, our methodology was to first identify the research question and then design search strategy and inclusion-exclusion criteria. The details of these steps are explained in the following sub-sections.

A. Research Questions Identification

Table 1 presents eight research questions addressed through this systematic review. To identify and analyze the deep insights from the selected studies were classified into sub-categories of text mining (RQ1). (RQ2 & RQ3) is to record the impact of the application of ML and its degree respectively. The primary and secondary challenges addressed are recorded through (RQ4 & RQ5). RQ6 is about the category of ML i.e. supervised, semi-supervised or unsupervised ML. In order to record the ML algorithms used and best among them for a given study, RQ7 & RQ8 are added respectively.

TABLE I. RESEARCH QUESTIONS

RQ#	Research Question
1	Which category of text mining is the target of the study?
2	Does ML application produce better results as compared to traditional approaches?
3	What is the degree of improvement, by using ML?
4	What was the primary challenge, for which this study was carried out?
5	What was the second challenge reported and addressed?
6	What category of ML algorithms was used i.e. supervised, unsupervised etc.?
7	Which ML algorithms were used?
8	Which ML algorithms were considered the most suitable?

The purpose of the first three research questions is to know the impact of the machine learning application on the sub-categories of text mining, which sub-category have gain high accuracy in results due to the usage of ML algorithms and which are low in accuracy. This will help us identify better steps about the application of machine learning in these sub-areas. RQ4 and RQ5 can help us to know about the common issues normally faced by the researchers while working in these sub-categories of text mining. RQ6, RQ7, and RQ8 will help us investigate mostly used and high performing algorithms in each category of ML i.e. supervised, semi-supervised, unsupervised, and reinforcement learning.

B. Search Strategy and Inclusion/Exclusion Criteria

For this systematic literature review, we developed a search strategy to identify relevant literature. This search strategy was tailored and the following search terms were used in combination with both 'machine learning' and 'deep learning' separately:

'Search and information retrieval', 'Document clustering', 'Document classification', 'Web mining', 'Information extraction', 'Natural language processing', and 'Concept extraction'.

We restricted the search from 2015 to 2020 and a total of 117 records were extracted.

The study is based on original research articles, and conference papers only. In order to guarantee the quality and relevance of academic literature included in the review process, the abstracts of the articles were thoroughly examined for analysis and purification. At a later stage, each research paper was carefully evaluated, and a criterion was set to exclude papers published in languages other than English. As a result, four articles in non-English language were excluded from the study. Additionally, after removing duplicate records, eleven more articles were eliminated. Finally, 70 articles were selected based on the inclusion and exclusion criteria assessment.

III. RESULTS AND DISCUSSIONS

This section presents the results obtained from primary studies selected for SLR. First, we present the answers to the research questions individually and then the overall results and its interpretation in the light of facts obtained from the study collectively

A. RQ1: Which category of text mining is the target of the study?

In this section we will see the distribution of the selected studies for this SLR into the categories of text mining. Table 2 shows the categories with the number of studies falls in that category. Although we have successfully classify 70% of the studies in to sub-categories of text mining i.e. document classification, document clustering, concept extraction, emotion intensity detection, entity extraction, entity relation modeling, opinion mining, and sentiment analysis as shown in Table 3, but unfortunately couldn't classify 30% of the studies due to its generic nature, so we put them under the label of text mining.

TABLE II. CATEGORIES OF TEXT MINING

Categories of Text Mining	Count of Studies
Concept Extraction	3
Document Classification	32
Document Clustering	3
Emotion Intensity Detection	4
Entity Extraction	1
Entity Relation Modeling	1

Categories of Text Mining	Count of Studies
Opinion Mining	2
Sentimental Analysis	3
Text Mining	21
Grand Total	70

B. *RQ2: Does ML application produce better results as compared to traditional approaches?*

68 out of 70 studies (i.e. 97.14%), reported that ML application produce better results as compared to traditional approaches. [25] reported a tie between ML and non-ML approaches while [26] preferred non-ML approach on ML approach due to its performance.

C. *RQ3: What is the degree of improvement, by using ML?*

In order to categorize the studies for RQ3, we define three classes on the basis of resultant accuracy of the studies. Those studies whose result accuracy was equal to or above 80% they were labeled as 'Best', studies whose accuracy fell in the range of 60% < Accuracy < 80%, were labeled as 'Average', while below 60% accuracy were declared as 'Poor'. This criterion for classification is adopted because majority of papers discussed their results in form of accuracy and F1 score. Figure 1 shows the percentage wise classification of studies according to its category. Studies having accuracy above 80% were considered 'best', because the results were considered promising compare to the alternate methods used, and the authors were satisfied from their own findings. While studies having accuracy between 60% to 80%, majority of the authors found them useful but feel need for improvement, so we put them in 'Average' categories. And studies who accuracy was below 60%, were considered not satisfactory and hence we consider them 'Poor'.

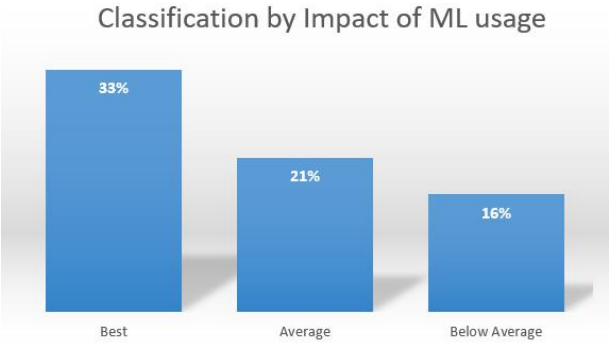

Figure 1. Categorical classification percentage of studies

D. *RQ4: What was the primary challenge, for which this study was carried out?*

The primary challenge means the basic issue which are addressed through that selected study. In order to summarize the response for this research question, we have divided it into five clusters i.e. Academic, Health, Safety & Security, Legal, and Material Science. Percentage distribution of these clusters are shown in Figure 2.

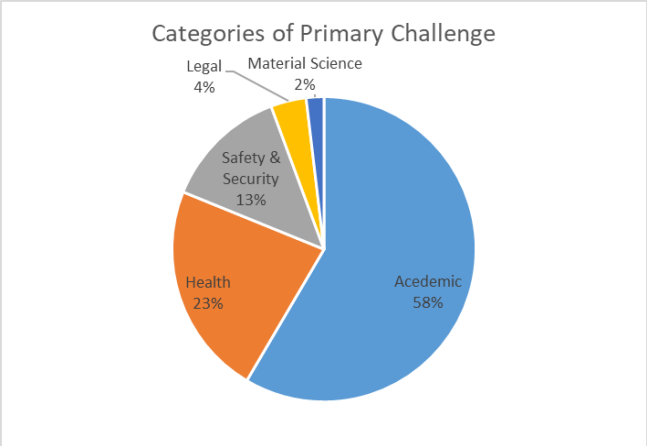

Figure 2. Summery of RQ4 responses for the selected studies.

E. *RQ5: What was the second challenge reported and addressed?*

The secondary challenge is not reported by 54 % of the selected studies. The remaining studies reported and addressed it. The categories of the secondary challenge reported are: lack of proper dataset, biased model, irrelevant result, overfitting, improper terminology, low performance, and curse of dimensionality. Percentage for each category is shown in Figure 3.

F. *RQ6: What category of ML algorithms was used i.e. supervised, semi-supervised, or unsupervised?*

The categories of ML algorithms used in the selected studies are classified into four classes i.e. supervised, semi-supervised, unsupervised, and reinforcement learning. Figure 2 shows the overall distribution into these four classes. The purpose of this question was to investigate the most commonly used ML algorithms in each category and also to compare the performance of these algorithms within each category and present the best performing algorithms for future researches in these categories but due to the limited number of studies in semi-supervised, unsupervised, and reinforcement learning, this research question lost its importance.

This issue is also highlighted and suggested solution for this is presented in the 'Limitation and Future Work' section of this paper.

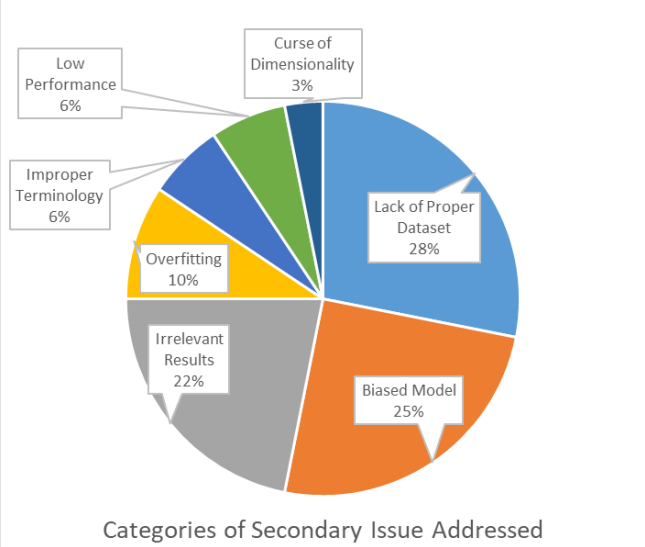

Figure 3. Summery of RQ5: Secondary Issue Addressed

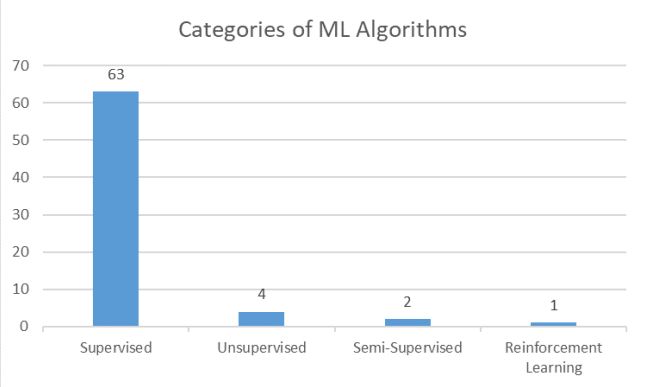

Figure 4. Four Categories of ML Algorithms and its frequency

G. *RQ7: Which ML algorithms were used?*

More than 50 different ML algorithms were in the selected studies. The list of top 10 most frequently used ML algorithms is shown in Table 3.

H. *RQ8: Which ML algorithms were considered the most suitable?*

The purpose of this research question was to find the best ML algorithms which outperformed out algorithms, from the selected studies. Table 5 shows the overall statistics of top 10 best performing algorithms. Table 4 shows the summery of best performing algorithms in different categories i.e. supervised, semi-supervised, unsupervised, and reinforcement learning.

IV. LIMITATION AND FUTURE WORK

As shown in Figure 4, an imbalance can be clearly observed in the selected studies based on the category of ML algorithms used. We cannot conclude with confidence that which ML algorithms have outperformed others for the categories of semi-supervised, unsupervised, and reinforcement learning due to relatively negligible number of studies. Our future work is to revisit the selection criteria in detail, investigate the reason, take necessary steps to collect all possible studies in categories of semi-supervised, unsupervised, and reinforcement learning, and after analysis of the ML algorithms used in each category, we will report the best performing algorithms in each of these categories.

TABLE III. LIST OF TOP 10 MOST FREQUENTLY USED ML ALGORITHMS

ML Algorithm	Count of Studies	Percentage
Support Vector Machine (SVM)	36	51.42 %
Naïve Bayes (NB)	28	40.00 %
Random Forest (RF)	14	20.00 %
Decision Tree (DT)	14	20.00 %
Logistic Regression (LR)	13	18.57 %
K-Nearest Neighbor (KNN)	12	17.14 %
Convolutional Neural Networks (CNN)	12	17.14 %
Recurrent Neural Network (RNN)	11	15.71 %
Long Short Term Memory (LSTM)	9	12.85 %
Bidirectional Long Short Term Memory (BiLSTM)	6	8.57 %

TABLE IV. OVERALL STATISTICS OF TOP 10 BEST PERFORMING ALGORITHMS

ML Algorithm	Count of Studies
Support Vector Machine (SVM)	19
Random Forest (RF)	6
Naïve Bayes (NB)	6
Recurrent Neural Network (RNN)	5
Long Short Term Memory (LSTM)	5
Decision Tree (DT)	4
Bidirectional Long Short Term Memory (BiLSTM)	4
Convolutional Neural Networks (CNN)	4
K-Nearest Neighbor (KNN)	3
Logistic Regression (LR)	2

TABLE V. SUMMERY OF BEST PERFORMING ALGORITHMS IN DIFFERENT CATEGORIES (RL: REINFORCEMENT LEARNING; SS: SEMI-SUPERVISED; SL: SUPERVISED LEARNING; US: UN-SUPERVISED)

ML Algorithm	RL	SS	SL	US
Support Vector Machine (SVM)		1	18	
Random Forest (RF)			5	
Naïve Bayes (NB)			4	
Convolutional Neural Networks (CNN)			4	

ML Algorithm	RL	SS	SL	US
Recurrent Neural Network (RNN)			4	
Decision Tree (DT)			3	1
Bidirectional Long Short Term Memory (BiLSTM)			3	
K-Nearest Neighbor (KNN)	1		2	
Long Short Term Memory (LSTM)		1	2	
Logistic Regression (LR)			2	
K-Mean Clustering				2
Stochastic Gradient Descent (SGD)			1	
Rank SVM-PSO			1	
Symbiotic Gated Recurrent Unit (SGRU)			1	
Conditional Random Fields (CRF)			1	
Latent Dirichlet Allocation (LDA)			1	
Deep Neural Network (DNN)			1	
DCN-BiLSTM			1	
Bidirectional Encoder Representations from Transformers (BERT)			1	
Multinomial Naïve Bayes (MNB)			1	
Random Multimodel Deep Learning (RMDL)			1	
LSTM-RNN			1	
AdaBoost			1	

V. CONCLUSION

In this paper a quick short but concise review have been presented to jump-start the research in the area of text mining and ML. The analysis presented in this paper will help the fresh researchers to directly start with the problem areas instead of implementing the initial algorithms and their combinations. This usually exhaust most of the researchers and only a few can go beyond these experiments and further their investigations. The problem areas, most widely and successfully used algorithms and actual problems have been identified in this field. The %age of the success of the algorithm is also discussed and presented in this paper.

REFERENCES

[1] S. G. Cho and S. B. Kim, "Identification of Research Patterns and Trends through," *International Journal of Information and Education Technology*, 2012.

[2] J.-L. Hung and K. Zhang, "Examining mobile learning trends 2003–2008: A categorical meta-trend analysis using text mining techniques.," *Journal of Computing in Higher education*, pp. 1-17, 2012.

[3] S. Inzalkar and J. Sharma, "A survey on text mining-techniques and application," *International Journal of Research in Science & Engineering*, pp. 1-14, 2015.

[4] A. Gudivada and N. Tabrizi, "A Literature Review on Machine Learning Based Medical Information Retrieval Systems.," *IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 250-257, 2018.

[5] D. Sharma, B. Kumar and S. Chand, "Trend Analysis in Machine Learning Research Using Text Mining.," in *International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, 2018.

[6] N. I. Widiastuti, "Deep Learning-Now and Next in Text Mining and Natural Language Processing," in *IOP Conference Series: Materials Science and Engineering*, 2018.

[7] W. M. Watanabe, K. R. Felizardo, A. C. Jr, É. F. Souza, J. E. d. C. Neto and N. L. Vijaykumar, "Reducing Efforts of Software Engineering Systematic Literature Reviews Updates using Text Classification," *Information and Software Technology*, 2020.

[8] A. Kaur and S. G. Jindal, "Severity Prediction Of Bug Reports using Text Mining: A Systematic Review.," in *International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, 2018.

[9] G. Alfattni, N. Peek and G. Nenadic, "Extraction of Temporal Relations from Clinical Free Text: A Systematic Review of Current Approaches.," *Journal of Biomedical Informatics*, 2020.

[10] M. J. Giummarra, G. Lau, G. Grant and B. J. Gabbe, "A systematic review of the association between fault or blame-related attributions and procedures after transport injury and health and work-related outcomes.," *Accident Analysis & Prevention*, 2020.

[11] J. Lever, M. R. Jones, A. M. Danos, K. Krysiak, M. Bonakdar, J. K. Grewal, L. Culibrk, O. L. Griffith, M. Griffith and S. J. Jones, "Text-mining clinically relevant cancer biomarkers for curation into the CIViC database.," *Genome medicine*, 2019.

[12] M.-Y. Cheng, D. Kusoemo and R. A. Gosno, "Text mining-based construction site accident classification using hybrid supervised machine learning.," *Automation in Construction*, 2020.

[13] S. Sharmin and Z. Zaman, "Spam detection in social media employing machine learning tool for text mining.," in *International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, 2017.

[14] C. J. Court and J. M. Cole, "Magnetic and superconducting phase diagrams and transition temperatures predicted using text mining and machine learning.," *npj Computational Materials*, pp. 1-9, 2020.

[15] H. Hayati, M. K. Idrissi and S. Bennani, "Automatic Classification for Cognitive Engagement in Online Discussion Forums: Text Mining and Machine Learning Approach.," in *International Conference on Artificial Intelligence in Education*, 2020.

[16] K. Christensen, S. Nørskov, L. Frederiksen and J. Scholderer, "In search of new product ideas: Identifying ideas in online communities by machine learning and text mining.," *Creativity and Innovation Management*, pp. 17-30, 2017.

[17] C. Rangu, S. Chatterjee and S. R. Valluru, "Text Mining Approach for Product Quality Enhancement:(Improving Product Quality through Machine Learning).," in *International Advance Computing Conference (IACC)*, 2017.

[18] A. N. Lam, A. T. Nguyen, H. A. Nguyen and T. N. Nguyen, "Bug localization with combination of deep learning and information retrieval.," in *25th International Conference on Program Comprehension (ICPC)*, 2017 .

[19] H. M. S. T. Sandaruwan, S. A. S. Lorensuhewa and M. A. L. Kalyani, "Sinhala Hate Speech Detection in Social Media using Text Mining and Machine learning.," in *19th International Conference on Advances in ICT for Emerging Regions (ICTer)*, 2019.

[20] W. Zhang and S. Ram, "A comprehensive methodology for extracting signal from social media text using natural language processing and machine learning.," in *Proceedings of the Workshop of Information Technology and Systems*, Dallas, 2015.

[21] T. S. Raghavendra and K. G. Mohan, "Web mining and minimization

framework design on sentimental analysis for social tweets using machine learning.," *Procedia Computer Science*, pp. 230-235, 2019.

[22] S. R. Ahmad, D. Harris and I. Sahibzada, "Understanding Legal Documents: Classification of Rhetorical Role of Sentences Using Deep Learning and Natural Language Processing.," in *IEEE 14th International Conference on Semantic Computing (ICSC)*, 2020.

[23] N. Bansal, A. Sharma and R. K. Singh, "An Evolving Hybrid Deep Learning Framework for Legal Document Classification An Evolving Hybrid Deep Learning Framework for Legal Document Classification.".

[24] R. Balyan, K. S. McCarthy and D. S. McNamara, "Combining Machine Learning and Natural Language Processing to Assess Literary Text Comprehension.," *Grantee Submission*, 2017.

[25] L. M. de Campos, J. M. Fernández-Luna, J. F. Huete and L. Redondo-Expósito, "Comparing machine learning and information retrieval-based approaches for filtering documents in a parliamentary setting.," in *International Conference on Scalable Uncertainty Management*, 2017.

[26] S. Shah and X. Luo, "Comparison of deep learning based concept representations for biomedical document clustering.," in *IEEE EMBS international conference on biomedical & health informatics (BHI)*, 2018.

[27] D. A. da Silva, C. S. t. Caten, R. P. d. Santos, F. S. Fogliatto and J. Hsuan, "Predicting the occurrence of surgical site infections using text mining and machine learning.," *PloS*, 2019.

[28] B. E. Perron, B. G. Victor, G. Bushman, A. Moore, J. P. Ryan, A. J. Lu and E. K. Piellusch, "Detecting substance-related problems in narrative investigation summaries of child abuse and neglect using text mining and machine learning," *Child abuse & neglect*, 2019.

[29] G. Thailambal and A. Sheshasaayee, "A different text mining process for classifying journal databases using machine learning algorithms," *International Journal of Recent Technology and Engineering (IJRTE)*, 2019.

[30] J. H. Suh, "SocialTERM-Extractor: Identifying and predicting social-problem-specific key noun terms from a large number of online news articles using text mining and machine learning techniques.," *Sustainability 11*, 2019.

[31] R. I. Rasel, N. Sultana, S. Akhter and P. Meesad, "Detection of Cyber-Aggressive Comments on Social Media Networks: A Machine Learning and Text mining approach.," in *2nd International Conference on Natural Language Processing and Information Retrieval*, 2018.

[32] C. Chukwuocha, T. Mathu and K. Raimond, "Design of an Interactive Biomedical Text Mining Framework to Recognize Real-Time Drug Entities Using Machine Learning Algorithms.," *Procedia computer science*, pp. 181-188, 2018.