Enhancing the Text Mining Process by Implementation of Average-Stochastic Gradient Descent Weight Dropped Long-Short Memory

Sreenivasa Rao Annaluri[11†] and Venkata Ramana Attili [2††],

VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India,
Sreenidhi institute of Science and Technology, Hyderabad, India

Summary

Text mining is an important process used for analyzing the data collected from different sources like videos, audio, social media, and so on. The tools like Natural Language Processing (NLP) are mostly used in real-time applications. In the earlier research, text mining approaches were implemented using long-short memory (LSTM) networks. In this paper, text mining is performed using average-stochastic gradient descent weight-dropped (AWD)-LSTM techniques to obtain better accuracy and performance. The proposed model is effectively demonstrated by considering the internet movie database (IMDB) reviews. To implement the proposed model Python language was used due to easy adaptability and flexibility while dealing with massive data sets/databases. From the results, it is seen that the proposed LSTM plus weight dropped plus embedding model demonstrated an accuracy of 88.36% as compared to the previous models of AWD LSTM as 85.64. This result proved to be far better when compared with the results obtained by just LSTM model (with 85.16%) accuracy. Finally, the loss function proved to decrease from 0.341 to 0.299 using the proposed model

Keywords:
Text Mining, Python, LSTM, AWD-LSTM, IMDB, NLP

1. Introduction

Day-to-day the amount of data keeps increasing at an exponential rate. A huge amount of data is streaming over the internet in the form of textual information, repositories, and many other formats [1]. Traditional data mining methods handle small scale data but it takes more time and effort to extract the information from the big data [2]. Hence, most of the studies/research programs include the concepts of text mining to make their work more powerful/meaningful at the workplace.

Text mining is defined as the process of converting unstructured data into a structured format .useful to identify meaningful patterns [3]. Using NLP tool for text mining performs the text analysis as an automated system. It helps to enhance the understanding between human brains and computer systems, which consists of different types of structures and patterns of languages. Whereas in the case of data mining only structured data is considered unlike text mining, where the data will be in the format of text-only [4]. Generally, text mining can be carried out on any type of data with different sizes/volumes to identify the relationships, facts, and assertions. These techniques are mostly used for classification and clustering to extract the information which mostly needed and important [5]. Once data is extracted then it is converted into structured form and then it can be presented using clustered HTML tables, charts, and can be utilized for other applications.

Text mining techniques are frequently applied towards categorization, sentimental analysis, and extracting hidden content to improve the process of analysing the data. These techniques experienced exponential growth and adoption over the last few years. The tendency of people to store any type of information in the form of text has been increasing with the latest accessories and technologies increasing the challenges for data science engineers. In this context, text mining is a great choice when compared to other technologies to analyse the data from different formats in a meaningful and sensitive manner [6]. Hence, the study is concerned with the implementation of Text mining methods on the Data and analysing potential risks and drawbacks.

1.1 Areas of Text Mining

According to various researchers, there are many perceptions about the different areas involved in text mining. For example, Miner *et al.* suggested seven (7) different areas of text mining as shown in Fig. 1, including search and information retrieval, document clustering, document classification, web mining, information extraction, NLP, and concept extraction [7]. However, few authors/ researchers considered web mining as a separate/isolated it from text mining.

Search Information Retrieval (IR): It refers to the storage and retrieval of text data from various documents, search engines, databases, etc. [8].

Document Clustering: it refers to the Grouping of similar documents into clusters such as categorizing the terms, paragraphs, and documents using the clustering methods [9].

Document Classification: it refers to Categorizing the document using various data mining classification methods that are based on the trained models and labelled examples.

Information Extraction (IE): it refers to the Extraction of structure data from unstructured text to obtain relevant information to create the relationships among the data [10].

Natural Language Processing (NLP): it is used to understand different tasks by using low-level language

Manuscript received July 5, 2022
Manuscript revised July 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.7.43

processing techniques. For example, it can tag some of the parts of a speech in a structured format, which is understood by the computer [11].

Web Mining: it is used to extract meaningful data from ample unstructured data. DM tools can predict the behaviour of businesses to acquire the best results. These tools also predict and resolve many business problems traditionally, which in general is considered a time-consuming process [12].

Concept Extraction: It refers to the grouping of words and phrases into similar groups which somehow helps in understanding the concepts and helps to improve the performance of machines and humans [7, 12].

2. Tables, Figures and Equations

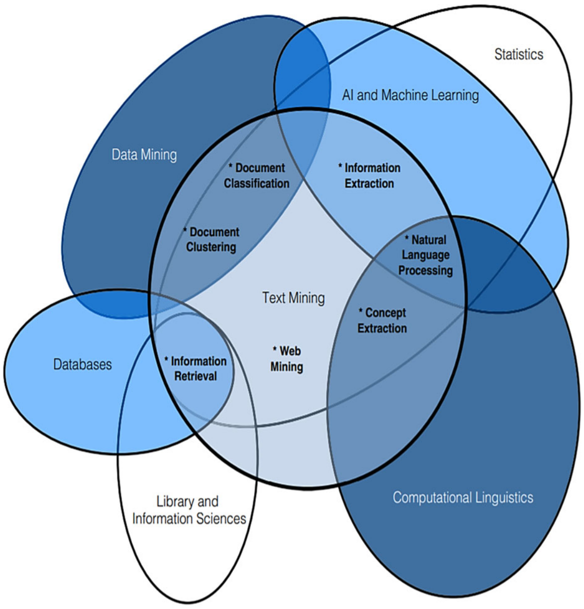


Figure 1: Different Areas of Text Mining.

Source: Miner *et al.* [7, pp. 31]

Hence, there are different areas recognized under the text mining techniques. Most of the applications are sentimental analysis of tweets. However, the notion of these areas when constituted together finds its way to various branches of library sciences like AI, Machine learning, and Deep Learning.

2.1 Related Work

The text mining techniques are implemented on various data types such as books, novels, journals, research articles, etc. In this context, recently Salloum *et al.* used different text mining techniques to extract the information from various research articles. The authors collected the research articles from six renowned databases in scientific research. Except for springer, all the other databases could not detect the interrelations between the topic and their similarities due to obscurities or unstructured interrelations [13].

In a similar line, extraction of data from unstructured and multidimensional big data was carried out by Adnan and Akbar towards their analytical study using text mining techniques [14]. The authors focused on the process of IE for obtaining useful information from various databases consisting of unstructured and semi-structured data, which can be considered as big data in their experiment. Some of the potential solutions were proposed by the authors help

To improve productivity and allow to improve big data analytics. A lot of text mining work is being carried out in big data analysis towards the financial sector according to Pejic et al. to understand different activities that are taking place in the financial institutions. Most of the data in these institutions are in the form of semi-structured or unstructured formats that tends to create massive obstructions for the decision-making process [15]. Therefore, the authors used qualitative analysis for conducting their research in text-mining results. Most of the companies involved with medical records are using various data processing techniques along with text mining technologies according to Sun et al. [16]. The authors applied the text mining techniques on the electronic medical records along with some of the data processing techniques as well. Text mining techniques are used to obtain the simulated and fascinated design data from various sources in this work. Later, alphanumeric textual data has been used to manage the contextual knowledge by Nimmagadda et al. [17].

The authors explained data mining as the process of extracting structured data and text mining as the process of evaluating the characters present in the textual format. Here, diverse views and contexts explained in this research are generally within the document ecosystems. Such types of aspects are analysed for contextual knowledge in a real-time world. Apart from that the authors also discussed, how to design the documents with an optimum number of words, sentences, and alphanumeric characters. In this context, the authors emphasized that they did not try to compromise with the semantics and contextual interpretation of multiple words in new knowledge domains, maintaining the overall quality of the document [17].

In the later stages, big data was critically examined using text mining according to Hassani *et al.* [18]. This has got so much popularity in recent times and become an emerging tool to extract knowledge from most of the unstructured data formats and helped to identify various patterns that are hidden in massive databases. The study of Text mining literature is useful for identifying the developments over a period of time. The experimental results claimed are useful for practitioners and researchers on different types of trends, methodologies, and applications. Mostly implemented in library sciences. Some of the challenges like classification of political speeches were examined in this research and tried to classify the sentiment words such as adjectives or adverbs in the case of movie reviews [18].

In a similar context, a big data analytics platform was developed by Mendhe *et al.* in their university [19]. This experiment allowed the authors to focus on the Twitter database to gain access. In this experiment, machine learning and text mining techniques played a prominent role in the analysis of various applications using Twitter data. Some of the authors experimented the text mining techniques for implementing the NLP. Such experiments provided remarkable outcomes in the area of the biomedical engineering industry according to Boukhari and Omri [20]. The role of big data in biomedical experiments and including the text mining concepts helped most researchers to make strong conclusions in their research area. Some of the machines learning techniques such as SVM models were also implemented as discussed by Luo towards the classification of the text mining approaches [21]. This approach was used to classify the English texts with more than 4,000 features and realized the classification rate up to 90%.

Apart from finding grounds in the business and analytics area. Text mining is also applied in the fields of deep learning and hybrid networks.

2.2. Text Analysis

The first step for a text analytics system is the input as text, which is unstructured data. The unstructured data is processed following the second step that is text processing. It moves all the text data into semantic-syntactic text format and it involves two series of steps [22]. The first step in text processing is cleaning up the text; it removes all the unnecessary information which is unwanted such as it removing unwanted symbols and repeated words. The second step in text processing is tokenization which splits the text into white spaces and improves the quality of the text data [23].

The third step is text transformation; it is a technique that is used to manipulate the style or size of text before proceeding. It also removes the repeated steps and controls the capitalization of the text [24]. After text transformation the fourth step is feature selection, it selects the important features as a subset and uses them in the creation of models. It reduces the input of processing and shows the data into statistics format. The last step is data mining, where the text mining process is merged with the traditional process. Data mining methods are also used in the structural database and the complete data is clustered [25]. The analytics step for text mining is mentioned in Fig. 2.

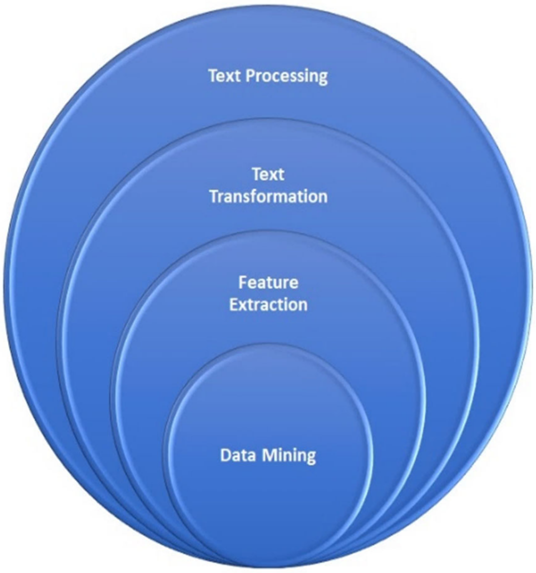


Figure 2: Steps for Text Analytics System.

The programming language python is used to process text data for various textual data analyses. The performance of python when used with text mining is fast when compared to other programming languages. The programming language python is used to perform text mining because of its robustness and flexibility with an ample number of libraries.

Python language includes back-end development, software development, and front-end development such as web development is also done. Python has many applications but a very important application for text processing is NLP. Text mining derives meaningful information from natural languages [26]. When compared to other programming languages it has fewer lines on code and focuses on code readability. Libraries such as Natural Language Toolkit (NLTK) are a group of libraries that help the user to create text processing in an efficient manner. Many other libraries are also available for the user that helps to perform many different tasks.

Python is simple to code and easy to use. Software is available free of cost. Python offers compatibility with various platforms so that no issues are faced which are common in other languages [27]. It supports both the object-oriented programming language and procedural-oriented programming language [28]. The first type utilizes objects based on data and functionality and the second type applies reusable pieces of code. Python is time-savvy as it allows the user to directly go to the research part without spending more time reading the document.

Natural Language Processing (NLP) is a part of computer science and artificial intelligence that deals with human languages [29]. It is used for transforming unstructured data into structured data. Role of NLP is analysing the text and recognizing the text which is in the form of speech, images, and so on. Natural language includes both understanding and generation which is used for creating text. Siri, Alexa, and so on are examples of the NLP which takes the inputs from the users and gives output in human-understandable language.

The text mining process finds frequency, word count, sentence length, and specific words. The main component for processing text mining is NLP which helps in identifying the sentiment, entities in the document, and types of the document.

Table 1: Test Accuracy

Model	Min Loss	Test Accuracy
LSTM	0.341	85.16
LSTM + Weight Dropped	0.338	85.64
LSTM + Weight Dropped + Embedding's	0.299	88.36

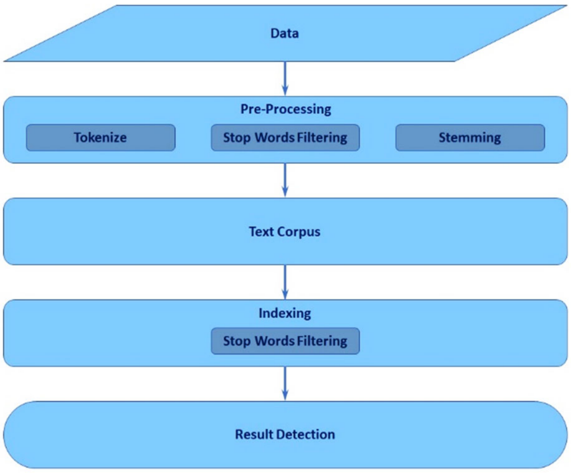


Figure 3: The Block Diagram of Text Mining

The block diagram of text mining gave detail explanation in Fig. 3. First and foremost, data are to be collected from the sources such as social media, Twitter, Google, and so on. The next step is to perform pre-processing where tokenization, stop words filtering, and stemming are done. Tokenization is the first step in pre-processing as it splits the sentences into meaningful words. The tokenization process breaks the text paragraph into a sentence. Similarly; it breaks the text paragraph into words. After this lower casing is done where strings are converted into lower case. All the stop words such as being, he, she, and so on are removed in order to acquire adequate results.

Stemming is defined as transforming the word into its root form. A text corpus is used to get the text in a structured manner. The next step is to check the occurrences of the root word and then display the results in the form of word clouds or graphs. The Data Analysis process consists of different segments to analyse the data. The data requirement specification is the first step to collecting the data and identifying the data. After collecting the data now perform the different processing techniques. Data collection is the process where the data should be in a structured format to perform the processing techniques.

2.3 LSTM and ASD-LSTM

This is one of the best-known models in Character level evolved with methods like Drop Connect and Random Gradient Descent along with other strategies hence, known as Average Weight Dropped LSTM. Text analysis is something related to Natural Language Processing as a part of Artificial Intelligence to normalize the texts in the real world into a structured format. However, NLP takes part as an analysis tool to evaluate the real-world text data. Hence, Long Short-Term Memory is also one of the recurrent neural networks which are most suitable for NLP. The circular connection sometimes leads to over-fitting issues. Drop Connect is introduced to normalize the issue in the network. Changing the weights randomly for the hidden layer to zero is referred to as Dropout as shown in Fig. 4. This is done during model training. However, the weights are retained keeping the nodes temporarily aside. So, dropout and dropout Connect are different unlike these dropouts the random changes are not considered by the drop connects.

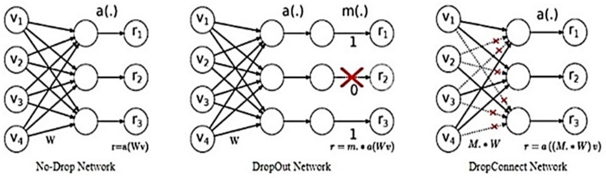


Figure4: DropConnectNetwork

3. Results

3.1. Dataset

The dataset considered in this study is based on collection reviews on movies from critics or viewers. However, the dataset is in the form of tar file which consists of two files namely negative and positive. Then the records in the positive and negative are displayed and found to be 12,500 for positive and 12,500 for negatives. Both of these datasets are combined as reviews dataset and test data is generated with the review (which is the text) and labeled as zero and one. There is a folder with 50000 unlabeled data. However, then an Unlabeled data frame is created

3.2 Packages

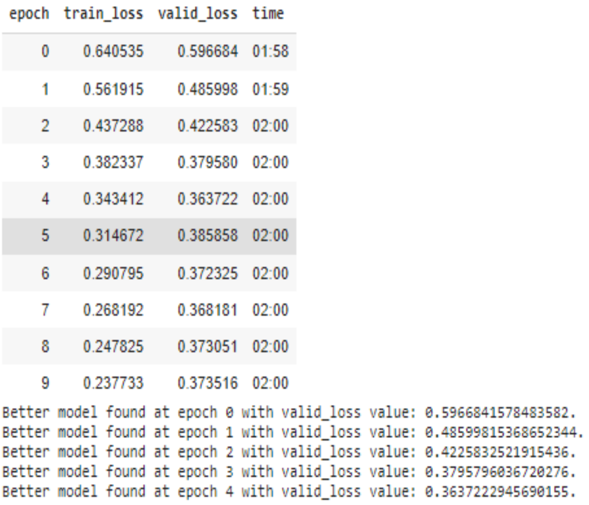
Packages like Pandas, NumPy, Seaborn, Matplotlib, Sklearn, and Fastai are imported to carry out the work.

3.3 Text Pre-Processing

This process is carried out after collecting the data set. This process includes achieving the steps:

- Removing all the unwanted characters using repackage.

Replacing all the white spaces and digits which are with no spacing'. These steps were considered for string pre-processing.



epoch	train_loss	valid_loss	time
0	0.640535	0.596684	01:58
1	0.561915	0.485998	01:59
2	0.437288	0.422583	02:00
3	0.382337	0.379580	02:00
4	0.343412	0.363722	02:00
5	0.314672	0.385858	02:00
6	0.290795	0.372325	02:00
7	0.268192	0.368181	02:00
8	0.247825	0.373051	02:00
9	0.237733	0.373516	02:00

```
Better model found at epoch 0 with valid_loss value: 0.5966841578483582.
Better model found at epoch 1 with valid_loss value: 0.48599815368652344.
Better model found at epoch 2 with valid_loss value: 0.4225832521915436.
Better model found at epoch 3 with valid_loss value: 0.379579603672027б.
Better model found at epoch 4 with valid_loss value: 0.3637222945690155.
```

Figure5: EarlyCall-BacksforGRUModel

- Next step is cleaning the corpus and this includes converting the text from unstructured to lower formatted text and updating the cleaned text to return it as an output.
- Then comes the tokenization phase where the corpus is then sorted to bring out the most basic word creating a dictionary.
- Then finally the text is tokenized. Dataset is divided into training and testing.
- There were 128 as the hidden dimension with 3 input layers and 1 output dimension. The loss function and optimizer are also considered in the study.
- Finally, using Fastai the model stops at the best epochs and gives the results as shown in Fig. 5.

To look back on the frequency of the reviews a graph is shown in Fig. 6 and the colours are given to the type of reviews. The green words represent the positive type/ sentiment-based review whereas the red colour words represent the negative type of review.

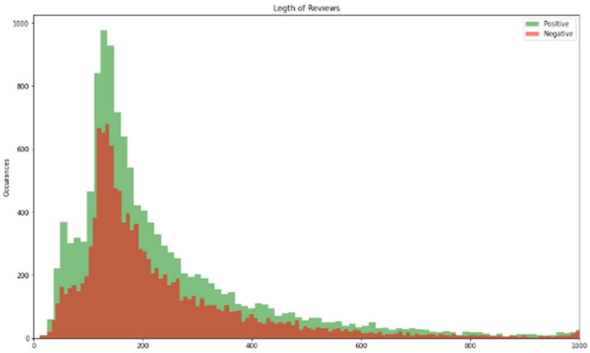


Figure 6: Reviews Frequency.

The best way to represent the words is using a word cloud. This study is based on text mining and it consists of words. Hence, word cloud analysis of the positive and negative words is shown in Fig. 7 and Fig. 8.

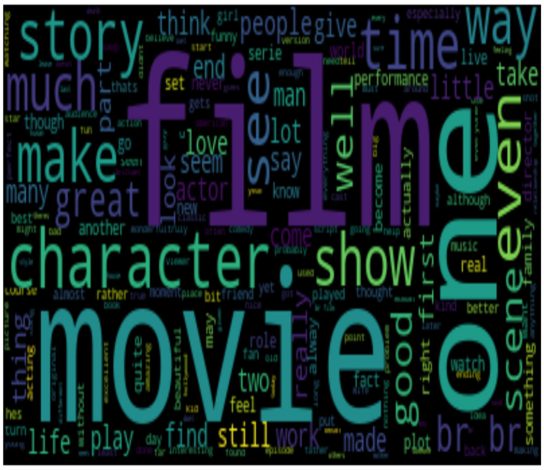


Figure 7: Positive Reviews Frequency.

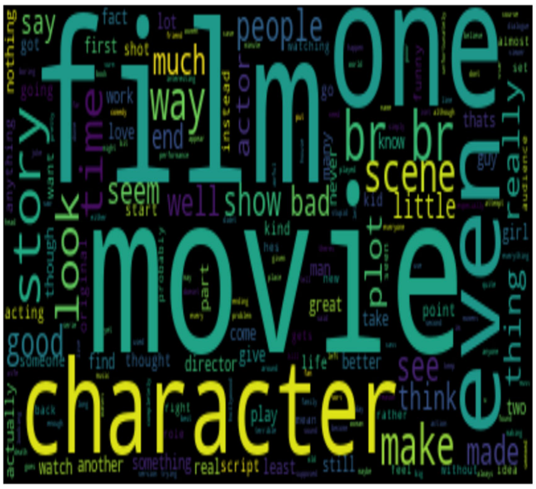


Figure 8: Negative Reviews Frequency

Hence, the LSTM model was implemented and the loss is recognized to be 0.34 initially. The model with LSTM has a recorded accuracy of 0.85 which was the first stage of the study. Then to improvise the model accuracy as well as loss the weight drop method discussed in the methodology section is implemented and the accuracy remains unchanged. However, the loss seems to be improvised from 0.341 to 0.299 and can be seen from Table 1.

Finally, ASD-LSTM is implemented and the accuracy of the model is seen increased 3 times more. The loss function is seen to be improvised as 4 times less than that of the previously implemented method. Hence, improvements are seen based on the accuracy and performance along with the loss of the model.

Acknowledgment:

The authors would like to thank the management of VNR Vignana Jyothi Institute of Engineering and Technology and Sreenidhi Institute of Science and Technology for their support and encouragement to carry out research programs at various stages.

References

[1] Acharjya, D.P. and Ahmed, K. "A survey on big data analytics: challenges, open research issues and tools." International Journal of Advanced Computer Science and Applications 7, no. 2 (2016): 511-518.

[2] Feng, Z. and Zhu, Y. "A survey on trajectory data mining: Techniques and applications." IEEE Access 4, pp. 2056-2067, 2016.

[3] Salloum, S.A. Al-Emran, M. Monem, A. A. and Shaalan, K. "Using text mining techniques for extracting information from research articles." In Intelligent natural language processing: Trends and Applications, pp. 373-397. Springer, Cham, 2018.

[4] Ferreira-Mello, R. André, M. Pinheiro, A. Costa, E. and. Romero, C. "Text mining in education." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9, no. 6 (2019): e1332.

[5] Ignatow, G. and Mihalcea, R. Text mining: A guidebook for the social sciences. Sage Publications, 2016.

[6] Porter, L.A. and Cunningham, S. W. Tech mining: exploiting new technologies for competitive advantage. Vol. 29. John Wiley & Sons, 2004.

[7] Miner, G. Elder IV, J. Fast, A. Hill, T. Nisbet, R. and Delen, D. Practical text mining and statistical analysis for non-structured text data applications. Academic Press, 2012.

[8] Halavais, A. Search engine society. John Wiley & Sons, 2017.

[9] Ayed, A.B. Halima, M.B. and Alimi, A. M. "Survey on clustering methods: Towards fuzzy clustering for big data." In 2014 6th International conference of soft computing and pattern recognition (SoCPaR), pp. 331-336. IEEE, 2014.

[10] Manoharan S. "A smart image processing algorithm for text recognition information extraction and vocalization for the visually challenged." Journal of Innovative Image Processing (JIIP) 1, no. 01 (2019): 31-38.

[11] Eisenstein, J. Introduction to natural language processing. MIT press, 2019.

[12] Jovic, A. Brkic, K. and Bogunovic, N. "An overview of free software tools for general data mining." In 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1112-1117. IEEE, 2014.

[13] Salloum, S.A. Al-Emran, M. Monem, A. A. and Shaalan, K. "Using text mining techniques for extracting information from research articles." In Intelligent natural language processing: Trends and Applications, pp. 373-397. Springer, Cham, 2018.

[14] Adnan, K. and Akbar, R. "An analytical study of information extraction from unstructured and multidimensional big data." Journal of Big Data 6, no. 1 (2019): 1-38.

[15] Pejić Bach, M. Krstić, Ž. Seljan, S. and Turulja, L.. "Text mining for big data analysis in financial sector: A literature review." Sustainability 11, no. 5 (2019): 1277.

[16] Sun, W. Cai, Z. Li, Y. Liu, F. Fang, S. and Wang, G. "Data processing and text mining technologies on electronic medical records: a review." Journal of healthcare engineering 2018 (2018).

[17] Nimmagadda, S.L. Zhu, D. and Reiners, T. "On Managing Contextual Knowledge of Digital Document Ecosystems, characterized by Alphanumeric Textual Data." Procedia Computer Science 159 (2019): 1135-1144.

[18] Hassani, H. Beneki, C. Unger, S. Taj Mazinani, M. and Yeganegi. M. R. "Text mining in big data analytics." Big Data and Cognitive Computing 4, no. 1 (2020): 1

[19] Mendhe, C.H. Henderson, N. Srivastava, G. and Mago, V. "A scalable platform to collect, store, visualize, and analyze big data in real time." IEEE Transactions on Computational Social Systems (2020).

[20] Boukhari, K. and Omri, M.N. "DL-VSM based document indexing approach for information retrieval." Journal of Ambient Intelligence and Humanized Computing (2020): 1-12.

[21] Luo, X. "Efficient english text classification using selected machine learning techniques." Alexandria Engineering Journal 60, no. 3 (2021): 3401-3409.

[22] Vani, K. and Gupta, D. "Unmasking text plagiarism using syntactic-semantic based natural language processing techniques: Comparisons, analysis and challenges." Information Processing & Management 54, no. 3 (2018): 408-432.

[23] Vijayarani, S. and Janani, R. "Text mining: open-source tokenization tools-an analysis." Advanced Computational Intelligence: An International Journal (ACII) 3, no. 1 (2016): 37-47.

[24] Curtis, B. Kellner, M.I. and Over, J. "Process modeling." Communications of the ACM 35, no. 9 (1992): 75-90.

[25] De, S. Musil, F. Ingram, T. Baldauf, C. and Ceriotti, M. "Mapping and classifying molecules from a high-

throughput structural database." Journal of cheminformatics 9, no. 1 (2017): 1-14.

[26] Usai, A. Pironti, M. Mital, M. and Mejri, C.A. "Knowledge discovery out of text data: a systematic review via text mining." Journal of knowledge management (2018).

[27] Zhang, R. Xiao, W. Zhang, H. Liu, Y. Lin, H. and Yang, M. "An empirical study on program failures of deep learning jobs." In 2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE), pp. 1159-1170. IEEE, 2020.

[28] Boer, F.D. Serbanescu, V. Hähnle, R. Henrio, L. Rochas, J. Din, C.C. Johnsen, E.B. et al. "A survey of active object languages." ACM Computing Surveys (CSUR) 50, no. 5 (2017): 1-39.

[29] Bhirud, N.S. "Grammar checkers for natural languages: a review." International Journal on Natural Language Computing (IJNLC)