

Sixth International Conference on Futuristic Trends in Networks and Computing Technologies (FTNCT06) held in Uttarakhand, India

Exploring the Landscape of Natural Language Processing for Text Analytics: A comprehensive Review

Om Prakash Suthar^{a*}, Ankita Mishra^b, Shilpa Singhal^c

^aMarwadi University, Rajkot, Gujarat, 36003, India

^bBanasthali Vidyapeeth, Jaipur, Rajasthan, 304022, India

^cMarwadi University, Rajkot, Gujarat, 36003, India, *opsuthar18@gmail.com

Abstract

The amount of textual information that can be analyzed in order to look for meaningful information has become a constraint as the amount of digital content that is being produced everyday increases. When it comes to mining large text datasets for useful information, NLP methods and models are necessary and extremely effective tools. The paper's primary objective is to present a comprehensive review of the NLP methods and models that are utilized for text analytics, sentiment analysis, topic modelling, text summarization, and text generation. In this paper, we will discuss the trending methodologies for social media analysis, consumer opinion analysis, and content creation. In addition, we will discuss the methodologies, methods, and evaluation metrics that are utilized in these types of contexts. This analysis aims to provide context for the development of natural language processing (NLP) in the context of text analytics, both historically and prospectively. This will be accomplished by providing context for the development of natural language processing in the context of text analytics.

© 2025 The Authors. Published by Elsevier B.V.

This is an open access article under the CC BY-NC-ND license (<https://creativecommons.org/licenses/by-nc-nd/4.0>)

Peer-review under responsibility of the scientific committee of the Sixth International Conference on Futuristic Trends in Networks and Computing Technologies (FTNCT06)

Keywords: NLP, LDA, NMF, text analytics, text extraction.

1. Introduction

There is now an unprecedented amount of textual data that permeates every aspect of modern life, thanks to the information explosion. Text data, such as social media posts, feedback from consumers, scientific papers, and legal documents, is a gold mine of information and knowledge [1]. However, it can be challenging for human beings to sort through such vast amounts of text with no structure in search of relevant information. This is where Natural Language Processing (NLP), a multidisciplinary field that draws from artificial intelligence (AI), linguistics, and computing,

comes in. By equipping computers with the ability to comprehend, interpret, and produce human language, NLP allows us to leverage the wealth of written information alongside the efficacy of computational analysis and analytical texts. One of the most important applications of natural language processing is text analytics, which entails mining text for useful information [1], [2]

Natural language processing (NLP) shows great promise in text analytics because it can aid businesses in uncovering hidden insights [2],[3]. When applied to customer sentiment analysis, systems for retrieving data, content digitization, document summarization, and the creation of artificially intelligent chatbots and AI-powered assistants, NLP techniques have great potential to improve business operations. Due to NLP's growing significance in text analytics, it's important to be well-versed in the most recent developments, strategies, and challenges in this field. This paper is helpful for scholars, practitioners, and industry professionals because it summarizes knowledge and expertise from a variety of fields to demonstrate how to use NLP to maximize the value of text data [4].

The challenges and limitations of using natural language processing for the analysis of texts are also brought to light in this article. These challenges and limitations include uncertainty, sensitivity to context, variations in language, and the requirement for large datasets that have been annotated [5]. Within the context of natural language processing (NLP), we highlight the importance of accountable and transparent AI development while discussing bias, privacy, and fairness. The purpose of this paper is to conduct a survey of the state of the art of natural language processing (NLP) for text analytics by synthesizing previous studies and examples from real-world applications. It outlines potential avenues for future research as well as places where novel methods can be tested [5, 6]. The purpose of this paper is to contribute to the development of natural language processing methods and to pave the way for innovative text analytics tools. These tools have the potential to revolutionize the way that we extract meaning from large textual datasets and use that knowledge to inform better decision making.

2. Sentiment Analysis Methodologies

From lexicon-based methods and machine learning algorithms to more recent innovations like deep learning models, we cover the gamut of techniques used in sentiment analysis here. It delves into methods of feature extraction, sentiment classification, and lexicon building for expressing emotions [7].

Understanding the emotional tone of text data is the goal of sentiment analysis, also called opinion mining. Here we will look at the many ways sentiment analysis can be done, from the tried-and-true to the cutting edge. The detailed literature review has been presented in (Table 1).

2.1. Lexicon-based Methods:

Pre-defined sentiment lexicons or dictionaries that include words or phrases annotated with sentiment scores are the backbone of lexicon-based methods. Polarity (positive, negative, or neutral) is represented by these numbers. The overall sentiment of a text is determined by adding up the sentiment scores of its individual words using a lexical approach [7],[8]. Common lexicon-based methods include counting words, matching words, and assigning weights to words.

2.2. Machine Learning Algorithms:

With their adaptability and scalability, machine learning techniques have found widespread use in sentiment analysis. Supervised learning algorithms are widely used, including SVM, Naive Bayes, and Random Forest. Training data for these algorithms consists of texts that have been annotated with labels indicating their general tone. A classifier is trained using features extracted from text data, such as word frequencies, n-grams, or syntactic patterns, to predict sentiment labels for unseen texts [9]. You don't need labelled data to perform sentiment analysis; you can use unsupervised learning algorithms instead, like clustering and topic modelling.

2.3. Deep Learning Models:

Sentiment analysis is just one area where deep learning has made a huge impact in the field of natural language processing. Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Transformer-based architectures like the **BERT (Bidirectional Encoder Representations from Transformers)** have all demonstrated impressive performance in capturing the context and nuance of sentiment in text. Word and sentence representations can be learned automatically by these models, with complex dependencies and relationships captured [10].

2.4. Hybrid Approaches:

Hybrid methods strengthen sentiment analysis by combining different approaches. Combining lexicon-based sentiment capture with machine learning or deep learning to account for context and interdependencies between sentiments is one example of a hybrid approach [11],[12]. Hybrid models, which combine elements of multiple methods, can improve sentiment analysis results, especially in complex settings.

2.5. Sentiment Lexicon Construction:

The foundation for lexicon-based approaches, such as sentiment analysis, are sentiment lexicons. **Sentiment lexicons are built by assigning ratings of happiness, sadness, anger, etc. to words and phrases.** Constructing something manually by domain experts takes a lot of time and effort. Another option is to use automated methods, such as bootstrapping, which uses seed words and their context to iteratively expand the sentiment lexicon. **WordNet and similar lexical resources can also be used to infer emotion from the relationships between words.**

3. Topic Modelling:

3.1. Algorithms:

Discovering hidden topics or themes in a set of documents is a common goal of topic modelling. The fundamentals, model assumptions, and inference methods of well-known algorithms for topic modelling are explored here.

Table 1. Summary of different Technologies for Sentiment Analysis and Text Analytic.

Cited No.	Year	Technology	Findings	Limitations	Results
[13]	2011	Machine Learning	Twitter sentiment analysis is done using various ML algorithms such as Decision tree, Naïve Bayes and SVM	Not work with objective sentences	Decision Tree performs the best
[14]	2014	Hybrid Approach	Proposed pipeline architecture based on three classification approaches: rule based, lexicon based and ML	work with limited dataset	Improved accuracy as compared to previous 2013 Test model
[15]	2015	Lexicon Based approach	Lexicon-based sentiment analysis algorithm has been designed with sentiment normalization and evidence-based combination function	limited to single language	More accuracy as compared to standard lexicon-based algorithm
[16]	2016	Machine Learning	Various methods of textual analysis using ML is compared	does not identify multiple objects within single document	Neural Network gives the best result
[17]	2016	Hybrid Approach	Hybrid approach combining NLP, Fuzzy Logic and ML	Technique works for only sentence level	Combination gives better results as compared to Naïve Bayes and Maximum Entropy techniques

[18]	2018	Deep Learning	Deep Learning model using LDA based text analytics	lack of elimination of hidden characteristics in text description	Deep neural Network with LDA outperforms SVM, Random Forest
[19]	2018	Lexicon Based approach	Lexical-based sentiment analysis is done with ML algorithms Decision Tree, KNN, SVM	lack of neutral sentiment polarity	KNN with lexicon method gives highest accuracy
[20]	2021	Hybrid Approach	Hybrid approach combining long short-term memory (LSTM) networks, convolutional neural networks (CNN), and support vector machines (SVM)	work on limited dataset, large computational time	L-CNN and C-LSTM in combination with SVM give the highest accuracy
[21]	2022	Machine Learning	Live Twitter sentiment analysis is done using various ML and text processing algorithms SVM, Naïve Bayes, Textblob and Lexicon approach	worked on limited dataset	Based on Majority vote combination of all four algorithms results best
[22]	2022	Deep Learning	Proposed a word embedding and splicing based text emotion analysis framework (called BCDF) along with Convolutional Long Short-Term Memory (Conv-LSTM) network	more time and space complexity	BCDF has highest accuracy as compared to LSTM based model
[23]	2022	Lexicon Based approach	Lexical-based sentiment analysis approach with polarity determination	limited dataset, lack of polarity parameters	proposed method gives highest accuracy compared to SVM, KNN, Naïve Bayes (NB) classifiers
[24]	2023	Hybrid Approach	Hybrid approach using NLP and Deep learning classifier LSTM is proposed	limited dataset, language dependent	HFV+LSTM Model outperforms the existing methods
[25]	2023	Machine Learning	Use of ML and Text Mining (TM) technology for risk analysis along with Term Frequency Inverse Document Frequency (TF-IDF) for feature extraction.	limited dataset, limited parameters for risk were considered	Proposed approach quickly identifies the risk factors by K means algorithm with text mining approach.

3.1.1 Latent Dirichlet Allocation (LDA):

One of the most popular algorithms for topic modelling is called Latent Dirichlet Allocation (LDA). Each document is viewed as a composite of various topics, with each topic represented by a word-by-word probability distribution. Bayesian inference is used by LDA to estimate the distributions of both topics and words within documents [26],[27]. These distributions are assumed to have a Dirichlet prior in the model. The latent variables are estimated using inference methods like Gibbs sampling or variational inference. In addition to its widespread use for a variety of text analysis and information retrieval tasks, LDA has proven effective at unearthing meaningful topics across a number of different domains.

3.1.2 Probabilistic Latent Semantic Analysis (PLSA):

Each document is thought to be produced by a pool of latent topics in PLSA, a generative probabilistic model. In contrast to LDA, PLSA does not rely on a hierarchy or state any prior assumptions. It uses a multinomial distribution to model the combined probabilities of words and documents [27]. The latent topic proportions and word distributions in PLSA are estimated using the Expectation-Maximization (EM) algorithm. PLSA's ease of use and success in a wide range of text analysis tasks belie its sophistication.

3.1.3 Non-Negative Matrix Factorization (NMF):

Topic modelling makes use of NMF, a matrix factorization method. The premise of NMF is that the product of two non-negative matrices representing topics and their corresponding weights in documents can serve as an approximation of a term-document matrix. The identification of meaningful and interpretable topics is facilitated by the non-negativity constraint. Matrix factorization and topic representations are achieved via iterative optimization algorithms like multiplicative update rules, which are used by NMF [28],[29].

In text mining applications, NMF has been used for tasks like topic identification and document clustering.

3.2 Applications:

Many fields have found use for topic modelling because it facilitates efficient knowledge discovery and data retrieval from massive text collections. The importance of topic modelling in various fields, including text classification, document clustering, recommender systems, and trend analysis, is discussed.

Topic modelling allows for the automatic assignment of categories or tags to documents based on their underlying topics, which is useful in text categorization. Topic modelling helps organize and structure a corpus by identifying its most prominent topics, which in turn improves search results and allows for more precise content filtering. Users can browse the document archive according to specific interests. Similar documents can be clustered together according to their topic distributions, making document clustering another application of topic modelling. Topic modelling helps with tasks like document organization, summarization, and recommendation by clustering documents so that similarity between them can be explored [30].

4. Proposed Methodology:

The process of utilizing NLP for text analytics is depicted in Figure 1. It all starts with the Text Data, which will ultimately be used as the analysis's input. In the step known as "Text Pre-processing," you will be responsible for preparing the text for analysis [32]. This includes tasks such as tokenization and sentence segmentation, stop word removal, text cleaning, and word normalization (such as stemming or lemmatization). Depending on the particular requirements of the analysis, these steps can be carried out singly or in any order that works best for the situation.

The next step is called **Feature Extraction**, and it is the process of extracting meaningful features from the text that has been pre-processed. Depending on the desired representation of the text data, this may involve employing methods such as Bag-of- Words, TF-IDF, Word Embedding, or other Representations.

The step known as Text Analysis Models incorporates a wide variety of NLP approaches and models [33],[34]. A few examples of these are sentiment analysis, text classification, named entity recognition, topic modelling, and text summarization, but there are many more. These models perform an analysis on the text data in order to gain insights and obtain information that is relevant.

During the step labelled "Output and Visualization," the analyzed text data are presented in a format that makes sense to the user. This may take the form of visualizations, reports, or summaries that provide insights that can be put into action and help decision- making proceed more smoothly. Figure.1 describes the general framework of NLP for text Analytics.

5. Text Extractive Summarization:

This section limelight's several approaches to extractive summarization that are applied most frequently today. When performing extractive summarization, it is typical to make use of methods that are centered on graphs. Nodes are the individual phrases or clauses that make up this representation, and edges are the connections that exist between the nodes. The representation being discussed here can be thought of as a tree. Before incorporating the sentences into a summary, graph algorithms like PageRank and Text Rank are applied to the text in order to ascertain which sentences contain the most pertinent and significant information [35]. Clustering techniques are used in extractive summarization as well. These methods cluster sentences or phrases that are semantically or contextually similar. Redundancy is cut down by grouping similar sentences together, and then representative sentences from each group are selected to form the summary.

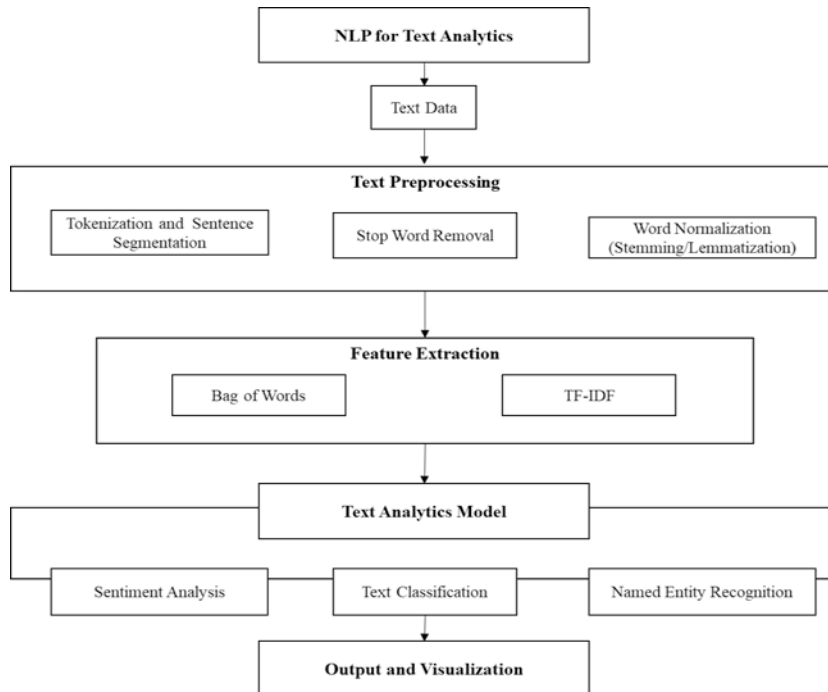


Fig. 1. General Framework of NLP for Text Analytics.

In extractive summarization, optimization algorithms like Integer Linear Programming (ILP) and Integer Quadratic Programming (IQP) are used to select the best possible sentences [35],[36]. Predefined scoring functions are used by these algorithms, and those functions consider things like sentence importance, coherence, and diversity. Summarization algorithms choose a diverse and informative set of sentences by treating the task as an optimization problem.

A more sophisticated method, abstractive summarization seeks to distil the essential meaning of a piece of text into a brief synopsis that reads like a human wrote it. In the field of abstract summarization, sequence-to-sequence models have recently received a lot of attention [37]. These models, which are frequently based on recurrent neural networks (RNNs) or transformer architectures, study the input text and figure out how to map it sequentially to a summary. To enhance the quality of generated summaries, reinforcement learning strategies have been investigated for use in abstractive summarization [38]. These methods involve training a summarization model with feedback on the quality of its generated summaries within a reward-based framework. The models can produce better quality summaries in terms of accuracy, concision, and coherence thanks to the optimizations made possible by reinforcement learning during the summarization process. To conclude, abstractive summarization techniques aim to understand and rephrase the text to generate concise summaries, while extractive summarization techniques focus on selecting important sentences or phrases from the original text [39]. Both methods have pros and cons, and which one is used depends on the nature of the information being summarized and the constraints that must be met. Text summarization systems that aid in information retrieval, document analysis, and content comprehension across domains are made possible by the evolution of these methods.

6. Text Generation:

This section describes several popular language models for text generation. Discussed are parameters, teaching strategies, and obstacles to writing texts for specific contexts. Text generators are useless without language models.

Next, you'll learn how n-gram, RNN, and transformer models can help writers. It examines how they're made, taught, and the challenges of writing context-appropriate text [40]. The n-gram language model is popular. The sentence's n words after the target word determine probability. N-gram models predict the next most likely word to use when generating text based on training data n-gram frequency. Thus, models reflect local conditions. Despite this, they rarely write meaningfully and coherently outside of the immediate context and fail to identify long-term dependencies.

The ability of Recurrent Neural Networks (RNNs) to capture sequential information has led to their increased popularity in language modelling. Word by word, recurrent neural networks (RNNs) absorb text, secretly encoding its context in a hidden state. Because this hidden state is updated recursively as new words are input, the model is able to capture dependencies across longer sequences. RNNs can be trained to predict the next word based on the current one, but only if the model is optimized for this task. However, RNNs' inability to reliably capture long-range dependencies is hampered by issues like vanishing or exploding gradients.

7. Evaluation Metrics:

Accuracy, precision, recall, F1-score, perplexity, and ROUGE scores are some of the important metrics used in sentiment analysis, topic modelling, summarization, and even text generation.

Accuracy: is a standard measure of a model's NLP performance's quality. The accuracy rate is the proportion of instances or labels that were correctly predicted. Accuracy in sentiment analysis, for instance, refers to the percentage of positive, negative, and neutral sentiment labels that were correctly assigned to data points. It offers a straight forward metric for evaluating the accuracy with which a model can predict the emotional tone of a given text sample [41].

Precision and recall have widespread application in binary classification tasks, where precision measures the fraction of correct predictions among all correct predictions and recall measures the fraction of correct predictions among all actual positive instances. Precision measures how well a model can predict positive sentiments, while recall measures how well it can capture all positive sentiments in a dataset. The F1-score is a balanced measure of the performance of the model because it incorporates both precision and recall into a single metric.

Perplexity is a metric for judging the success of language modelling projects. The accuracy with which a language model can predict a given text is evaluated. The model's ability to confidently and accurately predict the next word in a sequence, as measured by the perplexity score, improves as the score decreases. The measure of perplexity is widely used to evaluate the success of various language models in capturing the essential features of a language [40], [41].

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores are commonly used in tasks involving the summarization and creation of text. Similarity between the generated summary or text and the reference summary or text is quantified using ROUGE metrics. The quality and similarity of the generated text in comparison to the reference text can be gleaned from the various ROUGE scores, such as ROUGE-N (measuring n-gram overlap), ROUGE-L (measuring longest common subsequence), and ROUGE-S (measuring skip-bigram overlap).

Your evaluation metrics will depend on the text analytics task and your success criteria. These metrics help NLP researchers and practitioners quantify model comparisons, optimize parameters, and identify systemic weaknesses.

8. Current Advancements and Challenges:

This review paper's focus is on the state of natural language processing (NLP) for text analytics, including recent developments, obstacles, and potential future directions. It discusses the current state of the field, the obstacles that researchers and practitioners must overcome, and possible future avenues for development [41].

Further, this section delves into recent developments in pre-training methods, such as transformer-based models like **BERT (Bidirectional Encoder Representations from Transformers)**. Models can learn contextual representations and achieve state-of-the-art results in various text analytics applications when pre-trained on large-scale corpora and fine-tuned on specific tasks. Another major development in NLP for text analytics is the incorporation of domain-specific

-knowledge and resources. The accuracy and effectiveness of NLP models in domain-specific tasks, like sentiment analysis in the healthcare or financial domains, have been enhanced through the use of domain adaptation techniques and the incorporation of domain-specific lexicons, ontologies, or knowledge graphs. The scarcity of labelled training and evaluation data is a major obstacle. It may be difficult, expensive, or even impossible to collect labelled data for some tasks or domains. Because of this restriction, NLP models struggle to perform well and can only be used in specific contexts [42]. The difficulty in understanding and explaining NLP models is another obstacle [43]. It is difficult to comprehend the decision-making process of deep learning models like transformer models because they are typically treated as black boxes. In highly sensitive areas, such as healthcare or the law, interpretable models are crucial for establishing credibility, exposing biases, and explaining results.

Another major difficulty involves NLP models' susceptibility to and inability to withstand adversarial attacks due to their inherent fragility [44]. It is possible to trick natural language processing models with adversarial examples, which can then lead to inaccurate predictions or manipulated output text. Protecting natural language processing (NLP) models from malicious attacks is a top priority.

8.1. Future Directions:

Finally, text analytics NLP innovations are discussed. It tackles research-intensive issues. Text analytics may integrate multimodal data. Text, images, audio, and video can be combined for context and deeper analysis. Multimodal techniques can improve sentiment analysis, text summarization, and content generation. Interpretable NLP models also promise. Model interpretability and explainability research helps users trust complex models' predictions by revealing how they make decisions. This study affects secrecy-sensitive areas. Bias mitigation and fairness in NLP models need further study. NLP model deployment requires methods to identify and mitigate biases, ensure fairness in predictions and outputs, and promote ethics.

Data scarcity and domain adaptation must be addressed in the future. Transfer learning, few-shot learning, and the use of unlabeled data can reduce NLP models' reliance on large, labelled datasets and make them applicable to many domains and applications. In conclusion, the incorporation of deep learning models, pre-training techniques, and domain-specific knowledge into natural language processing (NLP) for text analytics has led to significant advances. Data interpretation, bias, availability, and robustness remain issues. Future research could focus on multimodal analysis, explainable models, addressing bias and fairness issues, and overcoming data scarcity and domain adaptation issues. These research areas will improve and ethically implement NLP in text analytic applications.

8.2 Conclusion:

This extensive review paper offers insights into the recent developments, methodologies, and applications of NLP techniques and models for text analytics. It highlights the importance of sentiment analysis, topic modelling, text summarization, and text generation in the process of extracting meaningful insights from large text datasets. The conclusion of the paper identifies the current challenges and future directions for further advancements in NLP for text analytics. Particular emphasis is placed on the potential impact that these advancements could have on a variety of domains, including social media analysis, customer feedback analysis, and content generation.

References

- [1] Roberts, M.E.; Stewart, B.M.; Airolidi, E.M. *A model of text for experimentation in the social sciences*. J. Am. Stat. Assoc., 111, 988–1003 (2016).
- [2] Roberts, M.E.; Stewart, B.M.; Tingley, D.; Lucas, C.; Leder-Luis, J.; Gadarian, S.K.; Albertson, B.; Rand, D.G. *Structural topic models for open-ended survey responses*. Am. J. Political Sci., 58, 1064–1082 (2014).
- [3] Jung, H.; Hong, S.; Park, J.; Park, M.; Sun, J.; Lee, S.; Ahn, J.; Ahn, M.; Park, K. MA19. 06 *Successful Development of Realtime Automatically Updated Data Warehouse in Health Care (ROOT-S)*. J. Thorac. Oncol., 14, S328. (2019)
- [4] Kreimeyer, K.; Foster, M.; Pandey, A.; Arya, N.; Halford, G.; Jones, S.F.; For-shee, R.; Walderhaug, M.; Botsis, T. *Natural language*

processing systems for capturing and standardizing unstructured clinical information: A systematic re-view. J. Biomed. Inform., 73, 14–29, (2017).

- [5] Zhang, J.; Gong, J.; Barnes, L. *HCCNN: Heterogeneous convolutional neural net-works for comorbid risk prediction with electronic health records*. In Proceed- ings of the 2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), 2017, Phila- delphia, PA, USA (2017).
- [6] Bian, J.; Barnes, L.E.; Chen, G.; Xiong, H. *Early detection of diseases using electronic health records data and covariance-regularized linear discriminant analysis*. In Proceedings of the 2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), 2017, Orlando, FL, USA, 16–19 Feb-ruary (2017).
- [7] Löbe, M.; Stäubert, S.; Goldberg, C.; Haffner, I.; Winter, A. *Towards Pheno- typing of Clinical Trial Eligibility Criteria*. Stud. Health Technol. Inform., 248, 293–299, (2018).
- [8] Sheikhalishahi, S.; Miotto, R.; Dudley, J.T.; Lavelli, A.; Rinaldi, F.; Osmani, V. *Natural language processing of clinical notes on chronic diseases: Systematic review*. JMIR Med. Inform., 7, e12239 (2019)
- [9] Dreisbach, C.; Kolecik, T.A.; Bourne, P.E.; Bakken, S. *A systematic review of natural language processing and text mining of symptoms from electronic pa- tient-authored text data*. Int. J. Med. Inform. (2019)
- [10] O. Alonso, K. Shiells, H. J. Lee and C. Carson, *System and method for generat-ing social summaries*, (2019).
- [11] K. C. Yang and Y. Kang, "Microblogs jasmine revolution and civil unrest: Re- assessing the emergence of public sphere and civil society in people's republic of china", *Censorship Surveillance and Privacy: Concepts Methodologies Toolsand Applications*, pp. 1153-1178, (2019).
- [12] Z. Lv, H. Song, J. Lloret, D. Kim and J.-N. De Souza, "Ieee access special sec-tion editorial: Big data analytics in the internet-of-things and cyber-physical sys-tems", *IEEE Access*, vol. 7, pp. 18070-18075, (2019).
- [13] I. Hemalatha, G. Varma, and A. Govardhan, "Machine Learning Methods in Classification of Text by Sentiment Analysis of Social Networks," *Int. J. Adv. Res. Comput. Sci.*, vol. 2, no. 5, (2011).
- [14] P. P. Balage Filho, L. Avanço, T. A. S. Pardo, and M. G. V. Nunes, "NILC USP: An Improved Hybrid System for Sentiment Analysis in Twitter Messages," *8th Int. Work. Semant. Eval. SemEval 2014 - co-located with 25th Int. Conf. Com-put. Linguist. COLING 2014, Proc.*, no. SemEval, pp. 428–432, (2014).
- [15] A. Jurek, M. D. Mulvenna, and Y. Bi, "Improved lexicon-based sentiment anal-ysis for social media analytics," *Secur. Inform.*, vol. 4, no. 1, doi: 10.1186/s13388-015-0024-x. (2015)
- [16] L. Guo, F. Shi, and J. Tu, "Textual analysis and machine leaning: Crack unstruc-tured data in finance and accounting," *J. Financ. Data Sci.*, vol. 2, no. 3, pp. 153–170, doi: 10.1016/j.jfds.2017.02.001 (2016).
- [17] O. Appel, F. Chiclana, J. Carter, and H. Fujita, "A hybrid approach to the senti-ment analysis problem at the sentence level," *Knowledge-Based Syst.*, vol. 108, pp. 110–124, doi: 10.1016/j.knosys.2016.05.040. (2016)
- [18] Y. Wang and W. Xu, "Leveraging deep learning with LDA-based text analyticsto detect automobile insurance fraud," *Decis. Support Syst.*, vol. 105, pp. 87– 95, doi:10.1016/j.dss.2017.11.001 (2018).
- [19] A. Aryo Prakoso, B. Winantesa Yananta, A. Fitra Setyawan, and Muljono, "A Lexicon-Based Sentiment Analysis for Amazon Web Review," *Proc. - 2018 Int. Semin. Appl. Technol. Inf. Commun. Creat. Technol. Hum. Life, iSemantic 2018*, pp. 503–508, doi: 10.1109/ISEMANTIC.2018.8549812 (2018).
- [20] C. N. Dang, M. N. Moreno-García, and F. De La Prieta, "Hybrid Deep Learning Models for Sentiment Analysis," *Complexity*, vol. 2021, doi: 10.1155/2021/9986920. (2021)
- [21] A. Motz, E. Ranta, A. S. Calderon, Q. Adam, F. Alzhouri, and D. Ebrahimi, "Live Sentiment Analysis Using Multiple Machine Learning and Text Pro- cessing Algorithms," *Procedia Comput. Sci.*, vol. 203, pp. 165–172, doi: 10.1016/j.procs.2022.07.023. (2022)
- [22] B. She, "Deep Learning-Based Text Emotion Analysis for Legal Anomie," *Front. Psychol.*, vol. 13, no. June, pp. 1–7, doi: 10.3389/fpsyg.2022.909157. (2022)
- [23] D. S. A. M. Mrtdaa Mohammed Almosawi, "Lexicon-Based Approach For Sen-timent Analysis To Student Feedback," *Webology*, vol. Volume 19, no. No. 1, pp. 6971–6989, [Online]. Available: <http://www.webology.org/abstract.php?id=1537>. (2022)
- [24] G. Kaur and A. Sharma, "A deep learning-based model using hybrid feature extractionapproach for consumer sentiment analysis," *J. Big Data*, vol. 10, no. 1, doi: 10.1186/s40537-022-00680-6. (2023)
- [25] Li, Zheng, Min Yao, Zhenmin Luo, Xinping Wang, Qianrui Huang, and Chang Su. "Analysis of risk factors of coal chemical enterprises based on text mining." *Journal of environmental and public health* 2023 (2023).
- [26] H.R. Marucci-wellman, H.L. Corns, M.R. Lehto, "Classifying injury narrativesof large administrative databases for surveillance - a practical approach combin-ing machine learning ensembles and human review, *Accid. Anal. Prev.*, 98 pp. 359-371, 10.1016/j.aap.2016.10.014 (2017).
- [27] S.J. Bertke, A.R. Meyers, S.J. Wurzelbacher, A. Measure, M.P. Lampl, D. Rob-ins Comparison of methods for auto-coding causation of injury narratives *Acid. Anal. Prev.*, 88, pp. 117-123, 10.1016/j.aap.2015.12.006, (2016).
- [28] Z. Liao et al., "Medical Data Inquiry Using a Question Answering Model", 2020IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1490- 1493, (2020).
- [29] P. Jeyaraj and E. R. S. Nadar, "Deep Boltzmann Machine Algorithm for Accu- rate Medical Image Analysis for Classification of Cancerous Region", *CognitiveComputation and Systems*, (2019).
- [30] A. M. Shah, X. Yan, S. A. A. Shah and G. Mamirkulova, "Mining patient opin-ion to evaluate the service quality in healthcare: a deep-learning approach", *Journal of Ambient Intelligence and Humanized Computing*, (2019).
- [31] S. Amin et al., "Recurrent Neural Networks With TF-IDF Embedding Tech- nique for Detection and Classification in Tweets of Dengue Disease", *IEEE Ac-cess*, vol. 8, pp. 131522-131533, (2020).
- [32] A. Gupta and R. Katarya, "Social Media based Surveillance Systems for Healthcare using Machine Learning: A Systematic Review", *Journal of Biomed-ical Informatics*, pp. 103500, (2020).

- [33] M. Bates, "Tracking Disease: Digital Epidemiology Offers New Promise in Pre-dicting Outbreaks", IEEE Pulse, vol. 8, pp. 18-22, (2017).
- [34] A. Mike and C. Daniel, "*Social media big data and mental health: current ad- vances and ethical implications*", Curr. Opin. Psychol., (2016).
- [35] K. Nargund and S. Natarajan, "*Public health allergy surveillance using micro- blogs*", 2016 Int. Conf. Adv. Comput. Commun. Informatics ICACCI 2016, pp.1429-1433, (2016).
- [36] V. Sahayak, V. Shete and A. Pathan, "*Sentiment analysis on twitter data*", Inter-national Journal of Innovative Research in Advanced Engineering (IJIRAE), vol. 2, no. 1, pp. 178-183, (2015).
- [37] F. Ceci, A. L. Goncalves and R. Weber, "*A model for sentiment analysis basedon ontology and cases*", IEEE Latin America Transactions, vol. 14, no. 11, pp. 4560-4566, (2016).
- [38] B. Liu, *Sentiment analysis: Mining opinions sentiments and emotions*, Cam- bridge University Press, (2015).
- [39] V. Agarwal, H. P. Sultana, S. Malhotra and A. Sarkar, "*Analysis of Classifiers for Fake News Detection*", Procedia Comput. Sci., vol. 165, no. 2019, pp. 377- 383, (2019).
- [40] W. Haitao, H. Jie, Z. Xiaohong and L. Shufen, "*A short text classification method based on n-gram and cnn*", Chinese J. Electron., vol. 29, no. 2, pp. 248-254, (2020).
- [41] S. Deepak and B. Chitturi, "*Deep neural approach to Fake-News identification*",Procedia Comput. Sci., vol. 167, no. 2019, pp. 2236-2243, (2020).
- [42] S. A. Cammel et al., "*How to automatically turn patient experience free-text responses into actionable insights: A natural language programming (NLP) ap-proach*", BMC Med. Inform. Decis. Mak., vol. 20, no. 1, pp. 1-10, (2020).
- [43] C. Zhang, A. Gupta, C. Kauten, A. V. Deokar and X. Qin, "*Detectingfake newsfor reducing misinformation risks using analytics approaches*", Eur. J. Oper. Res., vol. 279, no. 3, pp. 1036-1052, (2019).