Ms. Pooja, 2025, 13:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

A Survey of Techniques, Methods and Approaches in the Field of Natural Language Processing and Data Mining

Assistant Professor Ms. Pooja, Ms. Neeharika Sengar
Department of Computer Science, Raffles University

Abstract- This paper first describes the history of text mining technology, highlights its drawbacks, and then develops a text mining system based on natural language processing technology. The incessant generation of data has presented novel research obstacles because of its intricacy, variety, and magnitude. As a result, big data is gradually being acknowledged as a legitimate scientific discipline. An overview of the current state of big data science research is given in this article, with a focus on the theoretical underpinnings and applications of the field. Natural Language Processing (NLP) is one of the domains where data has a significant impact. The majority of NLP applications, including automatic speech recognition and machine translation, have not performed as well as they could in the past due to the proliferation of data. As such, a lot of NLP applications are regularly shifting from data-driven strategies to knowledge- and rule-based systems. On the other hand, gathered data that are based on vague design specifications or on forms that are not technically appropriate will be meaningless.

Keywords: Text mining; NLP; Bayesian networks, C4.5 Algorithm, K-mean Algorithm, Support Vector Machines, Apriori Algorithm.

I. INTRODUCTION

Natural Language Processing (NLP) and Data Mining are two interconnected domains that play a crucial role in extracting valuable insights and knowledge from vast and complex datasets. NLP focuses on the interaction between computers and human language, enabling machines to understand, interpret, and generate human-like text. On the other hand, Data Mining involves the discovery of patterns, trends, and knowledge from large volumes of data. The amalgamation of NLP and Data Mining has led to significant advancements in various applications, ranging from sentiment analysis and text summarization to predictive modeling and information retrieval. This introduction provides an overview of the methods and approaches employed in these fields, highlighting their synergies and the transformative impact on diverse industries.

II. BACKGROUND OF INFORMATION EXTRACTION AND TEXT MINING

Natural Language Processing (NLP):

- **Text Processing:**
- **Tokenization and Parsing:** Breaking down textual data into smaller units (tokens) and analyzing their syntactic structure.
- **Part-of-Speech Tagging:** Assigning grammatical categories to words, aiding in understanding the role of each word in a sentence.

- **Information Extraction:**
- **Named Entity Recognition (NER):** Identifying entities such as names, locations, and organizations within text.
- **Relation Extraction:** Determining relationships between entities to extract structured information.

- **Sentiment Analysis:**
- **Opinion Mining:** Analyzing text to determine the sentiment expressed, valuable for gauging public opinion and customer feedback. d. Machine
- **Translation:**
- **Statistical and Neural Approaches:** Translation of text between languages using statistical models or neural networks.

Data Mining:

- Association Rule Mining:
- **Apriori Algorithm:** Discovering associations and relationships within large datasets, commonly used in market basket analysis. b. Clustering:
- **K-means Algorithm:** Grouping similar data points together to uncover patterns and structures in data.

- **Classification:**
- **Decision Trees and Random Forests:** Predicting categorical outcomes based on input features.
- **Support Vector Machines (SVM):** Classifying data points by finding optimal hyperplanes in high-dimensional space.

Sequential Pattern Mining:

GSP (Generalized Sequential Pattern): Identifying sequential patterns in data, crucial in fields like bioinformatics and web usage mining.

Integration of NLP and Data Mining:

- **Text Mining:**
- **Information Retrieval:** Utilizing NLP techniques to enhance search engines by understanding user queries and retrieving relevant information.
- **Text Summarization:** Extracting key information from large text corpora for concise representation.
- **Entity Linking:** Connecting mentions in text to real-world entities for richer contextual understanding.
- **Event Extraction:** Identifying and categorizing events from unstructured text data.

III. LITERATURE SURVEY

- **Data Mining:** The practise of extracting valuable information from massive amounts of data kept in databases, data warehouses, or other information repositories, such as associations, trends, changes, noteworthy structures, and anomalies, is known as data mining [1]. Because there is a need to transform vast amounts of data that are available in electronic form into knowledge and information that is helpful for massive applications, it has become increasingly popular in recent years. Artificial Intelligence, Machine Learning, Market Analysis, Statistics and Database Systems, Business Management, and Decision Support are some of the domains where these applications are used [2].

- **Natural Language Processing:** As a branch of artificial intelligence, natural language processing, or NLP, is primarily used for Chinese word segmentation, question answering, machine translation, automatic summarization, textual entailment, part-of-speech tagging, parsing, natural language generation, text categorization, information retrieval, information extraction, text-proofing, and question answering. NLP has the advantage of being quick and effective. It supports constructive motivation free from harmful influences, which can successfully boost potential and promote ongoing learning, development, and growth [3]. NLP is frequently utilised in medical text processing for entity name and formation extraction, which includes word and phrase segmentation, syntactic, grammatical, and pragmatic analysis T. T. Kuo, P. Rao, C. Maehara et al. [4]. developed a multichannel processing technique, enhanced

data extraction capabilities, and extracted key disease-related themes from clinical notes using NLP tools. Jonnagaddala et al. [5] suggested a hybrid natural language processing (NLP) model to recognise clinical notes and electronic health records (EHR) signs and symptoms of Framingham heart failure. Trivedi et al. [6] created an interactive natural language processing (NLP) application that may be used by physicians to extract information from clinical texts following evaluation. Datta et al. [7] Analysed the NLP technology for cancer information extraction from EHRs, summarised each framework's implementation functions, and discovered a lot of duplicate content across several NLP frameworks, which led to some resource wastage. Decision support and medical data analysis modes will be transformed by the potential for diverse medical text data. Roberts and Demner-Fushman [8] In order to create a corpus that supported medical data mining, 468 electronic medical records were manually annotated. The complexity of manual data processing in data mining has significantly decreased with the emergence of NLP technology. Shikhar Vashishth et al. [9] a new semantic type prediction module for the biomedical NLP pipeline by using semantic type filtering to increase the performance connectedness of medical entities across all toolkits and datasets. Topaz et al. [10] reduced the manual workload in medical text data mining by using a recurrent neural network (RNN), support vector machine (SVM), NLP-based classification system, and other machine learning techniques to identify diabetic patients from clinical records.

IV. PAST & FUTURE WORK OF ALGORITHMS

S.No.	Algorithm Name	Past Work	Future Work
1.	Bayesian Networks	Medical Diagnosis: Bayesian networks have been extensively used in medical diagnosis, where they model relationships between symptoms and diseases. Risk Assessment: Applied in risk assessment and decision-making processes, such as in finance or project management.	Dynamic Bayesian Networks: Explore applications of dynamic Bayesian networks for 3odelling time-dependent relationships in various fields. Causal Inference: Enhance methods for inferring causal relationships in complex systems using Bayesian networks. Large-Scale Bayesian Networks: Develop efficient algorithms for learning and inference in large-scale Bayesian networks.

2.	C4.5 Algorithm	Data Classification: C4.5 has been widely used for classification tasks, particularly in machine learning and data mining. Decision Trees: The algorithm is known for generating decision trees, making it interpretable for non-experts.	Ensemble Methods: Explore ensemble methods involving C4.5 to improve classification accuracy. Handling Imbalanced Data: Investigate techniques to improve C4.5's performance on datasets with imbalanced class distributions.
3.	K-means Algorithm	Image Compression: K-means has been used for image compression, reducing the number of colors while maintaining visual quality. Clustering in Customer Segmentation: Applied in market research for customer segmentation and targeted marketing.	Dynamic K-means: Develop variants of Kmeans that adapt to changes in data distribution over time. Handling Noisy Data: Research techniques to make K-means more robust to noisy or outliers in the data.
4.	Support Vector Machines (SVM)	Image Recognition: SVMs have been successfully applied in image recognition and classification. Text Classification: Used for text classification tasks, such as spam detection and sentiment analysis.	Kernel Methods: Investigate novel kernel functions for SVMs to improve their performance on complex data. Large-Scale SVMs: Develop scalable versions of SVMs for handling large datasets.
5.	Apriori Algorithm	Market Basket Analysis:	Parallelization:
		Apriori has been widely used in market basket analysis to discover associations between products. Web Usage Mining: Applied in web usage mining to discover patterns in user behavior.	Explore parallel algorithms to improve the efficiency of the Apriori algorithm on large datasets. Incremental Mining: Develop techniques for incremental mining to update association rules in real-time.

Table 1: Past & Future work of Algorithms

V. CONCLUSION

In the dynamic landscape of NLP and Data Mining, the integration of sophisticated methods and approaches continues to propel advancements in artificial intelligence and data analytics. As technology evolves, researchers and practitioners strive to enhance the accuracy, efficiency, and applicability of these methods, leading to a deeper understanding of natural language and more effective knowledge extraction from diverse datasets. This paper will delve into specific methodologies within each domain, exploring recent developments and future directions in this ever-evolving intersection of Natural Language Processing and Data Mining.

REFERENCES

1. Han, J. and Kamber, M., (2006) "Data Mining: Concepts and Techniques", Elsevier.
2. Dunham, M.H., (2003) "Data Mining: Introductory and Advanced Topics", Pearson Education Inc.
3. K. Kreimeyer, M. Foster, A. Pandey et al., (2017) "Natural language processing systems for capturing and standardizing un structured clinical information: a systematic review," Journal of Biomedical Informatics, vol. 73, pp. 14–29.
4. T. T. Kuo, P. Rao, C. Maehara et al., (2017) "Ensembles of NLP tools for data element extraction from clinical notes," AMIA Annual Symposium proceedings AMIA Symposium, vol. 2016, pp. 1880– 1889.
5. J. Jonnagaddala, S.-T. Liaw, P. Ray, M. Kumar, N.-W. Chang, and H.-J. Dai, (2015) "Coronary artery disease risk assessment from unstructured electronic health records using text mining," Journal of Biomedical Informatics, vol. 58, no. Suppl, pp. S203–S210.
6. G. Trivedi, E. R. Dadashzadeh, R. M. Handzel, W. C. Wendy, V. Shyam, and H. Harry, (2019) "Interactive NLP in clinical care: identifying incidental findings in radiology reports," Applied Clinical Informatics, vol. 10, no. 4, pp. 655–669.
7. S. Datta, E. V. Bernstam, and K. Roberts, (2019) "A frame semantic overview of NLP-based information extraction for cancer related EHR notes [J]," Journal of Biomedical Informatics, vol. 100, no. 1, pp. 03–301.
8. K. Roberts and D. Demner-Fushman, (2016) "Annotating logical forms for EHR questions [J]. LREC," International Conference on Language Resources & Evaluation: [proceedings] Interna tional Conference on Language Resources and Evaluation, vol. 2016, no. 3, pp. 772–778.
9. S. Vashishth, D. Newman-Griffis, R. Joshi, D. Ritam, and P. R. Carolyn, (2021) "Improving broadcoverage medical entity linking with semantic type prediction and large-scale data sets," Journal of Biomedical Informatics, vol. 121, no. 10, pp. 38–80.
10. M. Topaz, L. Murga, O. Bar-Bachar, M. McDonald, and K. Bowles, (2019) "NimbleMiner," CIN: Computers, Informatics, Nursing, vol. 37, no. 11, pp. 583–590.
11. R. J. Oskouei, N. M. Kor, and S. A. Maleki, (2017) "Data mining and medical world: breast cancers' diagnosis, treatment, prognosis and challenges [J]," American journal of cancer research, vol. 7, no. 3, pp. 610–627.
12. Y. Zhang, S.-L. Guo, L.-N. Han, and T.-L. Li, (2016) "Application and exploration of big data mining in clinical medicine," Chinese Medical Journal, vol. 129, no. 6, pp. 731–738.
13. B. Polnaszek, A. Gilmore-Bykovskyi, M. Hovanes et al., (2019) "Overcoming the challenges of unstructured data in multisite, electronic medical record-based abstraction," Medical Care, vol. 54, no. 10, pp. e65–e72.
14. E. Ford, M. Oswald, L. Hassan, K. Bozentko, G. Nenadic, and J. Cassell, (2020) "Should free-text data in electronic medical records be shared for research? A citizens' jury study in the UK," Journal of Medical Ethics, vol. 46, no. 6, pp. 367–377.
15. S. M. Ayyoubzadeh, S. M. Ayyoubzadeh, H. Zahedi, M. Ahmadi, and S. R Niakan Kalhori, (2020) "Predicting COVID-19 incidence through analysis of google trends data in Iran: data mining and deep learning pilot study," JMIR public health and surveillance, vol. 6, no. 2, Article ID e18828.
16. X. Ren, X. X. Shao, X. X. Li et al., (2020) "Identifying potential treatments of COVID-19

from Traditional Chinese Medicine (TCM) by using a data-driven approach," Journal of Ethnopharmacology, vol. 258, no. 1, Article ID 12932.
17. E. Massaad and P. Cherfan, (2020) "Social media data analytics on telehealth during the COVID19 pandemic," Cureus, vol. 12, no. 4, Article ID e7838.
18. J. Dong, H. Wu, D. Zhou et al., (2021) "Application of big data and artificial intelligence in COVID-19 prevention, diagnosis, treatment and management decisions in China," Journal of Medical Systems, vol. 45, no. 9, p. 84.
19. L. B. Moreira and A. A. Namen, (2018) "A hybrid data mining model for diagnosis of patients with clinical suspicion of dementia [J]," Computer Methods and Programs in Biomedicine, vol. 165, no. 1, pp. 39–49.
20. S. Vilar, C. Friedman, and G. Hripcsak, (2018) "Detection of drugdrug interactions through data mining studies using clinical sources, scientific literature and social media," Briefings in Bioinformatics, vol. 19, no. 5, pp. 863–877.
21. H. S. Cha, T. S. Yoon, K. C. Ryu et al., (2015) "Implementation of hospital examination reservation system using data mining technique," Healthcare informatics research, vol. 21, no. 2, pp. 95–101.
22. B. L. Gudenas, J. Wang, S.-z. Kuang, A.-q. Wei, S. B. Cogill, and L.-j. Wang, (2019) "Genomic data mining for functional annotation of human long noncoding RNAs," Journal of Zhejiang University - Science B, vol. 20, no. 6, pp. 476–487.
23. R. S. Evans, (2016) "Electronic health records: then, now, and in the future," Yearbook of medical informatics, vol. Suppl 1, no. Suppl 1, pp. S48–S61.
24. P. C. Austin, I. R. White, D. S. Lee, and S. van Buuren, (2021) "Missing data in clinical research: a tutorial on multiple imputation," Canadian Journal of Cardiology, vol. 37, no. 9, pp. 1322–1331.
25. L. Yu, L. Liu, and K. E. Peace, (2020) "Regression multiple imputation for missing data analysis," Statistical Methods in Medical Research, vol. 29, no. 9, pp. 2647–2664.
26. P. C. Chang, C. L. Wang, F. C. Hsiao et al., (2020) "Sacubitril/valsartan vs. angiotensin receptor inhibition in heart failure: a real-world study in Taiwan," ESC heart failure, vol. 7, no. 5, pp. 3003–3012.
27. E. Tavazzi, S. Daberdaku, R. Vasta, C. Andrea, C. Adriano, and D. C. Barbara, (2020) "Exploiting mutual information for the imputation of static and dynamic mixed-type clinical data withan adaptive k-nearest neighbours approach," BMC Medical Informatics and Decision Making, vol. 20, no. Suppl 5, p. 174.
28. A. Idri, I. Kadi, I. Abnane, and J. L. Fernandez-Aleman, (2020) "Missing data techniques in classification for cardiovascular dysautonomias diagnosis," Medical, & Biological Engineering & Computing, vol. 58, no. 11, pp. 2863–2878.
29. C. Wang, C. Yao, P. Chen, S. Jiamin, G. Zhe, and Z. Zheying, (2021) "Artificial intelligence algorithm with ICD coding technology guided by the embedded electronic medical record system in medical record information management," Journal of healthcare engineering, vol. 2021, Article ID 3293457, 9 pages.