



The Impact of mHealth and NLP on Big Data Analytics in Healthcare

Gordana Jelić¹
Danica Mamula Tartalja²

Received: January 30, 2025

Accepted: June 5, 2025

Published: November 3, 2025

Keywords:

mHealth;
Natural language processing;
Big data in healthcare;
Text mining

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits non-commercial use, reproduction and distribution of the work without further permission.

Abstract: *The widespread adoption of mobile health (mHealth) has recently led to a significant expansion in the volume of medical data generated daily from diverse sources, including laboratory information systems, electronic health records (EHR), wearable devices, and social media. This trend is one reason why mHealth is increasingly associated with the concept of big data. Big data in healthcare encompasses a wide range of data types, differing in both source and level of structuring.*

Text mining and semantic analysis techniques within big data analytics have shown promising potential in addressing challenges in this field. Unstructured textual data, often referred to as big text data, contains invaluable insights that can help healthcare practitioners in clinical decision-making, enhance healthcare outcomes, and benefit scientific discoveries. Consequently, text mining has been widely employed to convert unstructured textual data gathered from different sources into a structured format.

This paper provides a comprehensive overview of the tools and methods offered by Natural Language Processing (NLP) for extracting and analyzing meaningful information and patterns from large volumes of natural language text data. It focuses on various text preprocessing techniques used to prepare raw text files to obtain relevant linguistic units interpretable by computers.

Given that human language remains central for documenting diseases and treatments, the integration of rich terminology systems and NLP in healthcare and biomedical research is of tremendous importance for statistical analytics and knowledge discovery. Moreover, healthcare professionals require access to scalable and expandable big data infrastructure to facilitate decision-making and diagnosis, improve healthcare outcomes, and, importantly, reduce costs in this sector.

1. INTRODUCTION

Due to the rapid development of technology and the enormous amounts of available data, analytics has become an immanent practice in many different fields of study, referring to the analysis which involves the detailed and careful examination of data to identify causes, crucial components, and potential outcomes. As a logical process of analyzing data, analytics is often supported by disciplines like statistics and computer science. Healthcare organizations are investing in the development of advanced tools and techniques to analyze, process, accumulate, and manage large amounts of healthcare data, both structured and unstructured (Nambiar et al., 2013; Rehman et al., 2021). Efficient management, analysis, and interpretation of big data can unlock new opportunities for modern healthcare, enabling healthcare providers to revolutionize medical therapies and personalized medicine (Dash et al., 2019). In many fields and industries, data analytics can be categorized into four types based on its outcomes: descriptive, diagnostic, predictive, and prescriptive analytics.

¹ Academy of Technical and Art Applied Studies Belgrade, Department School of Information and Communication Technologies, Zdravka Čelara 16, Belgrade, Serbia

² Academy of Technical and Art Applied Studies Belgrade, Department School of Information and Communication Technologies, Zdravka Čelara 16, Belgrade, Serbia

According to [Alghamdi et al. \(2021\)](#), among healthcare analytics methods, descriptive analytics plays a foundational role by analyzing raw data to discover patterns and provide insights for better patient care and more sophisticated decision-making. Healthcare companies can more effectively detect problems and their effects by using diagnostic analytics, which provides a deeper understanding of the reasons and motives behind their goal to generalize findings appropriate for wider populations. By predicting future trends and disease courses, predictive analytics expands these capabilities through the use of sophisticated data mining and machine learning techniques. This method has become increasingly integral in healthcare since it enables the development of frameworks and support systems that improve disease prevention and diagnostic precision. Descriptive, diagnostic, and predictive analytics work together to provide healthcare practitioners with a powerful set of resources that supports data-driven decision-making and operational effectiveness.

The digitization of healthcare data has led to the accumulation of massive heterogeneous datasets generated from various sources such as electronic health records, clinical notes, and social media ([Asri et al., 2015](#)). To best describe big data, [Laney \(2001\)](#) singled out three key dimensions of big data, known as the 3Vs: volume, velocity, and variety. Several researchers introduced other Vs to this model, such as veracity ([Schroeck et al., 2012](#)) and value ([Dijcks, 2013](#)).

In the paper about the potential and challenges of big data analytics in healthcare, the authors ([Raghupathi & Raghupathi, 2014](#)) describe four Vs of analytics. The volume of healthcare data is immense and continues to grow rapidly due to the continuous creation of data. Velocity is the term used to describe the real-time accumulation of data at unprecedented speeds, which makes timely processing and analysis challenging. Healthcare data has also evolved, becoming more complex due to the shift from structured formats like electronic health records to semi-structured and unstructured formats like multimedia files. This variety makes healthcare data both valuable and difficult to evaluate, requiring the application of advanced and scalable techniques to extract meaningful insights. Veracity, the fourth component, emphasizes how important data accuracy and dependability are in the medical industry.

[Hammad et al. \(2020\)](#) discuss the challenges and potential solutions for managing the increasing volume, velocity, and variety of healthcare data. Traditional data processing methods struggle to handle this influx of information, so the authors argue that semantic web technologies can complement big data approaches to address these challenges. Moreover, existing semantic ontologies used in healthcare can improve data acquisition and integration. However, there is a constant necessity for improving methods and techniques for the semantic storage of healthcare data.

Exploring how big data and semantic web technologies can revolutionize personalized medicine, [Panahiazar et al. \(2014\)](#) argue that the massive amounts of health-related data offer remarkable opportunities to improve clinical decision-making and patient outcomes. Semantic web technologies have a significant role in transforming “big data” into “smart data,” which further provides useful information for personalized treatment. The paper proposes a framework for empowering personalized medicine through a robust, scalable, and flexible semantic-driven big data infrastructure.

2. THE POWER OF NLP - TEXT PREPROCESSING TECHNIQUES

Text mining and semantic analysis techniques within big data analytics have emerged as promising solutions to extract valuable insights from this wealth of information ([Raghupathi & Raghupathi, 2014](#)). The potential of these techniques lies in their ability to convert unstructured textual data into a structured format, which can then be utilized by healthcare practitioners to drive scientific

discoveries. Leveraging the power of big data analytics in healthcare can lead to many benefits, including early detection, prediction, and prevention of health-related issues (Rehman et al., 2021).

Natural language processing (NLP) has emerged as a powerful tool for extracting and analyzing meaningful information from vast volumes of textual data (Torfi et al., 2020; Akerkar, 2018; Bahja, 2020). This paper provides a comprehensive overview of the various text preprocessing techniques employed in NLP to prepare raw text for further analysis and interpretation.

One of the fundamental steps in NLP is text preprocessing, which involves cleaning and transforming the raw text data into a format that computational algorithms can effectively process (Rajput, 2019). This typically includes steps such as tokenization, stop-word removal, stemming, and lemmatization, among others. (Kumar & Babu, 2018; Singh, 2018; Nhlabano & Lutu, 2018).

Tokenization is the process of breaking down the text into its fundamental units, known as tokens, which are typically words, numbers, or punctuation marks. This step is crucial as it allows the NLP system to identify and analyze individual linguistic elements within the text.

Following tokenization, stop-word removal is often performed to eliminate common words that hold little semantic value, such as articles, conjunctions, possessive adjectives, etc. (Akerkar, 2018). This helps to focus the analysis on more meaningful and informative words (Khurana et al., 2022). Stemming and lemmatization are two techniques used to further simplify the text by reducing words to their base or root form. Stemming involves the removal of suffixes from words to obtain their stem, while lemmatization utilizes more sophisticated linguistic knowledge to identify the base or dictionary form of a word, known as the lemma (Singh, 2018; Alaparathi & Mishra, 2020).

The effectiveness of these text-preprocessing techniques has been well documented in the literature. Studies have shown that proper text preprocessing can significantly improve the performance of various NLP tasks, such as sentiment analysis, information extraction, and text classification. For example, a study on sentiment analysis of Twitter data found that text preprocessing methods, such as handling hashtags, emojis, and URL links, can increase the predictive accuracy of sentiment classification models (Bao et al., 2014; Kumar & Babu, 2018; Nhlabano & Lutu, 2018). Another study on the impact of text preprocessing on sentiment analysis models for social media data demonstrated that applying various preprocessing techniques, including negation handling, stemming, and lemmatization, can significantly enhance the performance of the resulting models. Furthermore, the importance of text preprocessing is not limited to specific NLP tasks. Review on Sentiment Lexicons (Jagdale et al., 2018) highlights the critical role of lexicons in text preprocessing for a wide range of text mining applications.

3. CASE STUDY

As a multidisciplinary field, text mining integrates tools from information retrieval, data mining, machine learning, statistics, and computational linguistics to process natural language text. It converts large volumes of unstructured or semi-structured text data into a structured format, enabling further analysis or visualization through tables, charts, or mind maps (Tyagi, May 3, 2021). Using diverse methodologies, text mining explores facts, associations, and affirmations and applies text preprocessing to transform raw text into linguistically meaningful units.

This paper aims to investigate to what extent text mining relies on NLP by applying NLP text preprocessing techniques to five posts downloaded from PatientsLikeMe (2024), an open platform

where patients with chronic illnesses or conditions share their health journeys, experiences, and treatments. All posts are part of the discussion within the Diabetes online community and are used for healthcare research purposes as publicly available anonymized data.

Post 1: *I had type 2 diabetes and underwent a mini gastric bypass surgery. Amazingly, my diabetes was gone just one day after the surgery and it has stayed that way ever since!*

Post 2: *I am looking for shoes that are good for amputated toes. I have no toes left. I need a pair of shoes that are easy to put on and take off, but have a back because I have neuropathy and can't feel it when a shoe falls off. I also need a sandal. I have my Medicare shoes, but they are boots and sneakers. I'm looking for slippers and sandals.*

Post 3: *I'm new here and just curious. Are there people here not taking insulin shots? I have been diagnosed with type 2 diabetes and was prescribed with Insulin. I have been holding off and haven't gone back to the doctor but I plan to do it when I am able to. I have tried doing the low carb no sugar diet and my blood sugar is still always at 290 to 350.*

Post 4: *I have been diagnosed with diabetes for 2 years now. I have undergone several treatments which include daily insulin injections and medication. But i did like to know if there are alternative remedies that i can use to treat my neuropathy.*

Post 5: *They diagnosed me with type 2 diabetes a few years ago. Due to MS Osteoarthritis and Lupus it's hard to move around. I eat my green smoothies, veggies and minimum (ROM) range of motion when my body permits me to do so. The problem is nothing is still really helping drop my A1C. Tried metformin, januvia, majourna etc.*

All posts have undergone five NLP text preprocessing techniques, such as tokenization, stop-word removal, lemmatization, POS tagging, and parsing, but due to the scope of this paper, only the analysis of Post 1 will be presented.

POST 1 text preprocessing steps:

- Tokenization refers to breaking the text into individual words: [*'type', '2', 'diabetes', 'undergo', 'mini', 'gastric', 'bypass', 'surgery', 'amaze', 'diabetes', 'go', 'day', 'surgery', 'stay'*]
- Stop-word removal involves removing common but uninformative words like *'and', 'a', 'my', 'was', etc.*
- Lemmatization reduces a word to its base or dictionary form (lemma) and, unlike stemming, produces a linguistically correct base form: [*'type', '2', 'diabetes', 'undergo', 'mini', 'gastric', 'bypass', 'surgery', 'amazing', 'diabetes', 'go', 'day', 'surgery', 'stay'*]
- Part-of-speech tagging assigns grammatical tags to each token, indicating its part of speech (e.g. noun, verb, adjective, etc.): *type (noun), 2 (number), diabetes (noun), undergo (verb), mini (adjective), gastric (adjective), bypass (noun), surgery (noun), amaze (adjective), diabetes (noun), go (verb), day (noun), surgery (noun), stay (verb)*
- Parsing involves analyzing the grammatical structure of a sentence to understand how different words and phrases relate to each other: Subject: *"I"* → Verb: *"undergo"* → Object: *"mini gastric bypass surgery"* → Outcome: *"diabetes gone"*

This processed text demonstrates how raw healthcare text is transformed into a form suitable for machine learning algorithms or other analytical purposes. After having applied NLP preprocessing techniques to all selected posts, a comparison has been conducted to enable text classification and information extraction. The input data has been entered into the ChatGPT AI (OpenAI, 2023) and the detailed outputs have been structured and presented in Table 1.

The preprocessed data is structured in the Table according to different criteria. Category classifies the purpose of the post: sharing success - Post 1 highlights a resolved condition through surgery;

product recommendations - Post 2 focuses on specialized footwear for neuropathy and amputated toes; medical advice seeking - Posts 3, 4, and 5 explore challenges with diabetes management, alternative treatments, and multi-condition impacts. Themes and medical conditions identify key issues discussed, including health conditions, treatments, or lifestyle aspects: diabetes is present in all posts, either as a resolved issue (Post 1) or a current challenge (Posts 2-5). Treatment options are explored in Posts 3-5, including medications, dietary changes, and alternative remedies. Product recommendation is sought in Post 1. Challenges/Queries highlight the specific challenges faced or questions asked in Posts 2-5. Posts 3, 4, and 5 explore challenges with diabetes management, alternative treatments, and multi-condition impacts. Posts 2, 3, and 5 emphasize mobility, diet, and physical challenges due to underlying conditions. Outcomes state the results or unresolved issues, where applicable.

Table 1. Structured data table summarizing the five Posts

Post	Category	Themes	Medical Conditions	Treatments/Products	Challenges/Queries	Outcomes
1	Treatment Success Stories	Diabetes management, surgical intervention	Type 2 diabetes	Mini gastric bypass surgery	None	Diabetes resolved post-surgery
2	Product Recommendations	Neuropathy, footwear for mobility issues	Amputated toes, neuropathy	Shoes, slippers, sandals, Medicare footwear	Difficulty finding suitable shoes and sandals	None
3	Medical Advice Seeking	Diabetes management, dietary challenges	Type 2 diabetes	Insulin (prescribed but not used), low-carb diet	High blood sugar (290–350), looking for non-insulin treatment options	Blood sugar remains uncontrolled
4	Medical Advice Seeking	Neuropathy treatment, long-term diabetes	Diabetes, neuropathy	Insulin injections, medication	Seeking alternative remedies for neuropathy	None
5	Medical Advice Seeking	Chronic conditions, ineffective treatments	Diabetes, MS, osteoarthritis, lupus	Metformin, Januvia, Majourna, diet (green smoothies, veggies)	Difficulty lowering A1C despite medications and lifestyle adjustments	A1C levels remain high

Source: Own research

As can be seen from this small-scale analysis, NLP text preprocessing is crucial for transforming raw text into structured data that can be analyzed and summarized effectively. Tokenization allows the identification of individual entities (e.g., *diabetes*, *neuropathy*) and actions (e.g., *underwent surgery*) and facilitates the mapping of key themes and concepts to specific table fields like Medical Conditions and Treatments/Products. Stop-word removal reduces noise in the text, ensuring only meaningful words (e.g., *insulin*, *neuropathy*) are considered for analysis. Lemmatization normalizes variations of the same word, ensuring consistency (e.g., *treated*, *treatments* → *treatment*) and helps group related concepts under unified terms, improving the accuracy of fields like Treatments/Products and Challenges/Queries. Part-of-speech (POS) tagging enables the identification of key entities (nouns) like *diabetes* and *surgery*, and helps extract actions (verbs) like *underwent*, *prescribed*, and descriptors (adjectives) like *chronic*. Parsing identifies relationships between entities and actions (e.g., *diabetes* is treated with *insulin*).

In order to extract structured data in a variety of ways, NLP preprocessing is essential. More precise identification of key themes can be achieved by segmenting texts into smaller parts and concentrating on meaningful words. The correct categorization of data into table fields is enabled by methods such as entity extraction. Furthermore, by identifying cause-and-effect links, parsing maps connections between data points. Eliminating irrelevant words guarantees that only significant content is examined. Consequently, it reduces the possibility of incorrect classification or overcrowding tables with unnecessary details. In conclusion, NLP preprocessing aids in the extraction of insightful information from text, producing an output that is organized, structured, and simple to understand.

4. CONCLUSION

The choice of appropriate preprocessing techniques can be highly context-dependent and may vary depending on the specific application and the characteristics of the text data. For instance, in the domain of sentiment analysis, studies have found that the inclusion of emoticons, emojis, and abbreviations can provide valuable insights into the writer's sentiment. They thus may require specialized handling during the preprocessing stage (Resyanto et al., 2019). An in-depth analysis of the language of electronic discourse, in SMS communication media in particular (Jelić & Polovina, 2024), highlights the intensive usage of different graphostylistic innovations, such as emoticons, multiplication of graphemes, code-switching, verbalization of laughter, etc. that the writers of the messages use to add special nuance to the meaning of their utterances. Additionally, when gathering information from social media, special attention should be paid to the shortening and clipping of words and phrases - linguistic processes that need to be understood to analyse the text.

The field of NLP is continuously evolving, and new techniques and approaches are constantly being developed to enhance the effectiveness of text preprocessing. As the volume and complexity of textual data continue to grow, the importance of efficient and robust text preprocessing methods will only become more pronounced.

High data quality is critical, as inaccurate or unreliable information can directly impact life-saving decisions. Despite advancements in analytics, maintaining veracity, the fourth key feature of data analytics, remains a challenge, particularly with unstructured data, where errors, such as misinterpreted handwritten prescriptions, are common. Thus, the ultimate goal of healthcare analytics should be to ensure error-free and credible data.

The combination of NLP-driven big data analytics and mHealth solutions has the potential to drastically change healthcare services. More individualized and proactive medical care is made possible by mobile health applications, which enable continuous collection and analysis of patient-generated data. But there are still a lot of important issues that must be resolved in the future. Significant issues include interoperability, data imbalance, missing data, data noise, sophisticated integration of varied data kinds, and combining data from multiple sources, in addition to the previously mentioned heterogeneity and large data volumes. To overcome these and other challenges, a multidisciplinary team effort combining technological know-how with a solid ethical basis is needed. Finding this balance will allow the healthcare industry to fully utilize mHealth and NLP to improve patient care, reduce costs, and advance medical research.

References

- Akerkar, R. (2018). Natural Language Processing. In SpringerBriefs in business (p. 53). Springer Nature. https://doi.org/10.1007/978-3-319-97436-1_5
- Alaparthy, S., & Mishra, M. (2020). Bidirectional Encoder Representations from Transformers (BERT): A sentiment analysis odyssey. In arXiv (Cornell University). Cornell University. <https://doi.org/10.48550/arxiv.2007.01127>
- Alghamdi, A. M., Alsubait, T., Baz, A., & Alhakami, H. (2021). Healthcare Analytics: A Comprehensive Review [Review of Healthcare Analytics: A Comprehensive Review]. Engineering Technology & Applied Science Research, 11(1), 6650. Engineering, Technology & Applied Science Research. <https://doi.org/10.48084/etasr.3965>
- Asri, H., Mousannif, H., Moatassime, H. A., & Noël, T. (2015). Big data in healthcare: Challenges and opportunities (p. 1). <https://doi.org/10.1109/cloudtech.2015.7337020>
- Bahja, M. (2020). Natural Language Processing Applications in Business. In IntechOpen eBooks. IntechOpen. <https://doi.org/10.5772/intechopen.92203>
- Bao, Y., Quan, C., Wang, L., & Ren, F. (2014). The Role of Pre-processing in Twitter Sentiment Analysis. In Lecture notes in computer science (p. 615). Springer Science+Business Media. https://doi.org/10.1007/978-3-319-09339-0_62
- Dash, S., Shakyawar, S. K., Sharma, L., & Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. In Journal Of Big Data (Vol. 6, Issue 1). Springer Science+Business Media. <https://doi.org/10.1186/s40537-019-0217-0>
- Dijcks, J. (2013). Oracle: Big data for the enterprise. Oracle White Paper. Redwood Shores, CA: Oracle Corporation.
- Hammad, R., Barhoush, M., & Abed-alguni, B. H. (2020). A Semantic-Based Approach for Managing Healthcare Big Data: A Survey [Review of A Semantic-Based Approach for Managing Healthcare Big Data: A Survey]. Journal of Healthcare Engineering, 2020, 1. Hindawi Publishing Corporation. <https://doi.org/10.1155/2020/8865808>
- Jagdale, R. S., Shirsat, V. S., & Deshmukh, S. N. (2018). Review on Sentiment Lexicons. <https://doi.org/10.1109/cesys.2018.8723913>
- Jelić, G., & Polovina, V. (2024). *Diskurs SMS poruka – jezički procesi i inovacije*. Beograd: Filološki fakultet.
- Khurana, D., Koli, A., Khatter, K., & Singh, S. (2022). Natural language processing: state of the art, current trends and challenges. In Multimedia Tools and Applications (Vol. 82, Issue 3, p. 3713). Springer Science+Business Media. <https://doi.org/10.1007/s11042-022-13428-4>
- Kumar, C. S. P., & Babu, L. D. D. (2018). Novel Text Preprocessing Framework for Sentiment Analysis. In Smart innovation, systems and technologies (p. 309). Springer Nature. https://doi.org/10.1007/978-981-13-1927-3_33
- Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note, 6.
- Nambiar, R., Bhardwaj, R., Sethi, A., & Vargheese, R. (2013). A look at challenges and opportunities of Big Data analytics in healthcare. <https://doi.org/10.1109/bigdata.2013.6691753>
- Nhlabano, V. V., & Lutu, P. E. N. (2018). Impact of Text Pre-Processing on the Performance of Sentiment Analysis Models for Social Media Data (p. 1). <https://doi.org/10.1109/icabcd.2018.8465135>
- OpenAI. (2023). *ChatGpT* (Mar 14 version) [Large language model] <https://chat.openai.com/chat>
- Panahiazar, M., Taslimitehrani, V., Jadhav, A., & Pathak, J. (2014). Empowering Personalized Medicine with Big Data and Semantic Web Technology: Promises, Challenges, and Use Cases. PatientsLikeMe. (2024, December 15). Diabetes Community. Discussion. Retrieved January 11, from patientslikeme.com

- Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential [Review of Big data analytics in healthcare: promise and potential]. *Health Information Science and Systems*, 2(1). Springer Nature. <https://doi.org/10.1186/2047-2501-2-3>
- Rajput, A. E. (2019). Natural Language Processing, Sentiment Analysis and Clinical Analytics. In arXiv (Cornell University). Cornell University. <https://doi.org/10.48550/arxiv.1902.00679>
- Rehman, A., Naz, S., & Razzak, I. (2021). Leveraging big data analytics in healthcare enhancement: trends, challenges and opportunities. In *Multimedia Systems* (Vol. 28, Issue 4, p. 1339). Springer Science+Business Media. <https://doi.org/10.1007/s00530-020-00736-8>
- Resyanto, F., Sibaroni, Y., & Romadhony, A. (2019). Choosing The Most Optimum Text Preprocessing Method for Sentiment Analysis: Case:iPhone Tweets. In 2019 Fourth International Conference on Informatics and Computing (ICIC) (p. 1). <https://doi.org/10.1109/icic47613.2019.8985943>
- Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., & Tufano, P. (2012). Analytics: The real world use of big data. New York, NY: IBM Institute for Business Value, Said Business School.
- Singh, S. (2018). Natural Language Processing for Information Extraction. In arXiv (Cornell University). Cornell University. <https://doi.org/10.48550/arxiv.1807.02383>
- Torfi, A., Shirvani, R. A., Keneshloo, Y., Tavaf, N., & Fox, E. A. (2020). Natural Language Processing Advancements By Deep Learning: A Survey. In arXiv (Cornell University). Cornell University. <https://doi.org/10.48550/arxiv.2003.01200>
- Tyagi, N. (2021, May 3). What is Text Mining? Process, Methods and Applications. Retrieved January 12, from www.analyticssteps.com