BOOK REVIEW

Singh, J. (2023). Natural Language Processing in the Real World: Text Processing, Analytics, and Classification. CRC Press, Taylor & Francis Group

https://doi.org/10.17356/ieejsp.v1li1.1421

Natural Language Processing (NLP) in the Real World presents a structured guide to the fundamentals of NLP, combining theoretical concepts with practical applications. Covering essential techniques like text preprocessing, sentiment analysis, text classification, and named entity recognition (NER), the author emphasizes hands-on Python implementations, making it accessible to beginners and professionals alike. The book's strength lies in its focus on real-world problems, such as social media analysis and customer feedback. NLP has emerged as an indispensable tool for analysing large amounts of text data, especially with the exponential growth of unstructured information across industries like e-commerce, social media, healthcare, and finance. The book comprehensively explores NLP techniques with a focus on real-world applications. It aims to bridge the gap between theoretical understanding and practical implementation of NLP tasks (Simske & Vans, 2021). This review critically assesses the book's chapter-wise content, evaluating the strengths and weaknesses of its methodology, as well as its contributions to social science research, data science and academia. The analysis also highlights the book's limitations, especially in addressing the complexities of modern NLP research.

Natural Language Processing (NLP) is a branch of artificial intelligence that focuses on the interaction between computers and human languages by transferring it to the algorithm (Dyshel & Lane, 2023). It involves text processing, sentiment analysis, machine translation, speech recognition, and text classification. By using techniques like tokenization, stemming, and machine learning, NLP enables machines to understand, interpret, and generate human language. Modern applications include chatbots, virtual assistants, language models (e.g. GPT), and more. NLP is critical in various domains, including social media analysis, healthcare, and customer service automation. Through different sections and chapters, the author explains important NLP concepts, ideas, data curation, data processing, implementation of NLP application, and implementation of NLP application in the real world.

The book deals with questions such as how can we preprocess and prepare text data for machine learning models, what are the most effective methods for performing sentiment analysis on textual data, which classification algorithms can be used to categorize

text, how can Named Entity Recognition (NER) be applied to identify specific entities within text (Fortino, 2021), and finally, what are the practical challenges in applying NLP to real-world scenarios like social media analysis or customer feedback? These questions guided the author to articulate the book systematically.

Natural Language Processing in the Real World explains how to clean, tokenize, and normalize text using methods like stemming, lemmatization, and stop-word removal, making it ready for machine learning. It demonstrates how machine learning algorithms, such as Naive Bayes or Support Vector Machines (SVM), classify text sentiment, social media and product reviews (Atkinson-Abutridy, 2022). It explains how to identify and classify entities like the names of people, organizations, or places using NER algorithms, employing libraries like spaCy. Most importantly, the book highlights challenges in applying NLP to real-world contexts, such as noisy text data, social media, and business-related use cases, and provides step-by-step Python code to overcome these issues.

The book has six sections covering twelve chapters to explain NLP as its central idea. The introductory discussion in Section I (Chapter 1) introduces fundamental NLP concepts in an accessible manner. While it covers standard material like tokenization, part-of-speech tagging, stemming, and lemmatization, it offers little to experts in the field (Iezzi et al., 2020). It could benefit from broader coverage of recent innovations like transformer-based tokenization or advanced embeddings.

Section II (Chapter 2) discusses real-world data challenges, especially in environments where pre-labelled datasets are scarce. It allows us to learn how to identify, access, and extract data from diverse sources. The author's experience with industry datasets adds depth here, but it could benefit from more exploration of cutting-edge scraping techniques and API integration methods. Greater attention to data ethics, especially in large-scale extraction, could enhance the strength of this section.

Section III (Chapter 3) gives substantial coverage of key preprocessing steps, including dealing with unstructured text and feature extraction. The author provides practical Python code examples, which enhance the book's usability. The chapter thoroughly covers text cleaning, tokenization, and vectorization, offering clear Python implementations, however, it could offer more advanced techniques like sentence embeddings or multilingual text preprocessing. This section permits learners to capture master techniques for cleaning and transforming raw text data into structured formats for analysis (Iezzi et al., 2020), and to understand feature extraction and vectorization methods.

Chapter 4 of the section is one of the book's highlights. The author seamlessly transitions into various NLP models, from traditional ones like TF-IDF (Term Frequency-Inverse Document Frequency) to more advanced machine learning algorithms (Dyshel & Lane, 2023). The balance between theoretical explanation and application is well-maintained, though a more detailed discussion on transformer-based models would have added value, given their dominance in the field (Dyshel & Lane, 2023). The model diversity here is commendable, ranging from traditional machine learning to newer deep learning models, ensuring a broad overview. More focus on state-of-the-art models such as transformers could be beneficial, as these are now industry-standard in NLP tasks. Nevertheless, the omission of detailed hyperparameter tuning or model optimization strategies leaves gaps for more advanced readers.

Section IV (Chapters 5 & 6) focuses on real-world use cases of NLP, emphasizing industry applications. The author provides rich examples across different domains, making the content particularly valuable for professionals in e-commerce, finance, and customer service. However, the sections are somewhat disjointed, lacking a unifying narrative between the case studies. In Chapter 5, the real-world applications in various industries are well-covered, giving readers clear insights into NLP's practical impacts in sectors like finance and healthcare. It indicates that better integrating practical examples with theoretical insights could strengthen the learning experience. Chapter 6 highlights emerging areas of NLP, such as legal text mining and healthcare applications. Though the range of applications is limited, the Chapter could explore how emerging NLP applications evolve in response to large language models.

Section V (Chapters 7 & 8) presents information extraction techniques, covering both rule-based and machine-learning approaches. However, the book could benefit from deeper coverage of neural network-based extraction models, such as BERT and GPT-based frameworks, which are now becoming industry standards. In Chapter 7, the author offers solid coverage of entity extraction and summarization techniques, providing practical examples. This Chapter is an excellent source to gain knowledge of information extraction techniques and text summarization (Ignatow & Mihalcea, 2018). Also, it explores models for entity recognition and transformation. However, while practical, it lacks depth in explaining sophisticated extraction models, particularly those utilizing pre-trained models. Advanced techniques such as abstractive summarization models like PEGASUS are underexplored.

In Chapter 8, the author explains various text categorization algorithms, from Naïve Bayes to Support Vector Machines and the use of affinity models. The inclusion of performance metrics and error analysis for each model is a commendable touch. However, the absence of more recent advances like zero-shot classification is noticeable, which is meticulously relevant in today's fast-evolving NLP landscape (Wang et al., 2023; Yu et al., 2023). However, excellent explanations of classification techniques, from logistic regression to support vector machines, with relevant code examples are promising. However, the challenge lies in balancing foundational theory with cutting-edge developments to meet a broad audience's needs. Nonetheless, the section is equipped enough to develop skills in text categorization techniques among learners who can learn to use models like Naïve Bayes and Support Vector Machines for classification tasks.

Section VI comprises four chapters (9-12) and delves into advanced NLP applications, which marks an essential shift from foundational concepts to real-world solutions. Each chapter focuses on practical implementation, moving beyond theoretical discussions and offering insights into how NLP can be applied to develop functional, scalable systems in real-world environments. Chapter 9 focuses on one of the most visible and widespread applications of NLP –chatbots. The author explains the basic principles of chatbot development, including rule-based approaches and machine learning-driven solutions. The chapter provides step-by-step guidance, with Python code and tools such as NLTK and spaCy, for building a basic chatbot. The real-world focus on chatbots ties in well with the book's overarching theme of practical NLP. The explanation of both rule-based and machine learning methods allows readers to explore multiple development paths, based on their

needs or technical expertise. However, the chapter's limitation lies in its exclusion of more advanced conversational AI models, such as transformers (e.g. GPT-3), which dominate modern chatbot development. The absence of neural-based approaches could leave readers wanting more depth in this rapidly evolving area.

Chapter 10 provides an in-depth look at how NLP can be used to analyse customer feedback from platforms like e-commerce sites and social media. It focuses on extracting actionable business insights through sentiment analysis, topic modelling, and aspect-based sentiment analysis, helping businesses improve products and enhance user experiences (Sailunaz et al., 2018; Seyeditabari et al., 2018). It includes practical Python examples, making the concepts accessible to business analysts, data scientists, and product managers. The Chapter primarily relies on traditional machine learning models, such as Naïve Bayes and SVM, for sentiment classification and feature extraction, offering a comprehensive guide to these techniques. While this makes it suitable for those looking to apply NLP to real-world problems quickly, it lacks coverage of advanced deep learning methods, such as transformers and BERT, which are now widely used in NLP tasks. That limits the chapter's relevance for readers seeking more cutting-edge techniques. The author also addresses challenges in handling noisy, unstructured text by explaining effective preprocessing methods like tokenization and stop-word removal (Atkinson-Abutridy, 2022). The chapter provides valuable strategies for identifying sentiment polarity and extracting specific product features mentioned in reviews. However, it offers only a superficial treatment of more complex issues, such as sarcasm detection and opinion spam, which are critical in customer review analysis but not fully explored.

Chapter 11 focuses on integrating NLP into recommendation systems, a vital tool in e-commerce and media platforms. Singh demonstrates how sentiment analysis and topic modelling can predict user preferences by analysing text-based interactions, product descriptions, and reviews. The chapter's strength lies in its practical approach, offering step-by-step guides with Python libraries like NLTK and spaCy, making it accessible for professionals. However, it has limitations, particularly its lack of discussion on advanced deep learning techniques such as neural networks and transformers, which limits its relevance for cutting-edge research. It also misses a critical discussion on ethical issues like bias and over-personalization in recommendation systems. Despite these gaps, the author addresses challenges like noisy text data and the user cold start problem, providing practical solutions.

The last chapter broadly examines how NLP can enhance business intelligence (BI). It shows how textual data from various sources – such as emails, customer reviews, or news articles – can be processed and analysed to deliver insights that support business decision-making. This chapter highlights one of the most crucial applications of NLP–turning unstructured text into actionable data for businesses. The author effectively showcases how BI tools can integrate NLP and provides relevant case studies to show its real-world impact. The coverage of this chapter is somewhat generic, focusing more on traditional NLP techniques. It overlooks a discussion on how emerging trends like deep learning, transformers, and knowledge graphs can enhance business intelligence. This omission reduces the relevance of the chapter for businesses seeking more cutting-edge solutions.

Section VI of the book highlights practical NLP applications like chatbots and recommendation systems, providing accessible explanations and Python code for implementation. However, it mainly focuses on traditional methods and neglects advanced models like transformers, limiting its demand for researchers. Though there is a lack of discussion on ethical implications, the section offers useful insights for practitioners. Its limitations may hinder its relevance for advanced research.

The methodology of the book is practical, emphasizing traditional machine learning techniques like Naive Bayes and SVM, with essential NLP preprocessing steps such as tokenization and stemming. The accessible, Python-driven approach is ideal for real-world applications like sentiment analysis. While the methodology is more suited for practical implementation and introductory learning, the book provides a solid foundation for applying NLP techniques in real-world settings, contributing to academic learning and practical applications across fields. However, it lacks depth in advanced methods like deep learning, limiting its use for cutting-edge research. Additionally, the book does not adequately address ethical concerns in NLP applications, such as bias or more complex issues like sarcasm detection and opinion spam.

The volume is particularly recommended for professionals in fast-paced industries as it offers practical, code-based solutions for immediate implementation. Its interdisciplinary approach appeals to a wide audience, including data scientists and social researchers, with accessible explanations and coding examples suited for beginners. Covering sentiment analysis, information extraction, and text categorization, it provides actionable skills applicable to business, technology, and social science research, especially for analysing large-scale textual data such as social media posts or surveys (Grimmer et al., 2022). It addresses common challenges in text processing, such as noisy data, context understanding, bias in models, and language diversity. The author offers practical solutions and applies NLP in real-world contexts such as sentiment analysis and customer reviews. Although the book briefly touches on bias mitigation, it primarily focuses on providing scalable, real-time Python-based solutions for large-scale data, making it especially valuable for social researchers tackling complex societal issues using NLP techniques. Though it could explore more advanced techniques like transformer models, its focus on real-world applications ensures it remains a useful resource for those looking to apply NLP in industry or research settings.

Mohammad Ashraful Alam[1,2,*]

[1] ELTE Eotvos Lorand University, Doctoral School of Sociology, Faculty of Social Sciences, Budapest-1117, Hungary.

[2] Department of Criminology and Police Science, Mawlana Bhashani Science and Technology University (MBSTU), Tangail-1902, Bangladesh.

* Correspondence: maalam@student.elte.hu; maalam.cps@gmail.com; Orcid: 0009-0005-7064

References

Atkinson-Abutridy, J. (2022). *Text analytics: An introduction to the science and applications of unstructured information analysis.* Chapman and Hall/CRC. https://doi.org/10.1201/9781003280996

Dyshel, M., & Lane, H. (2023). *Natural Language Processing in Action* (2nd, Version 8 ed.). Manning Publications. https://www.manning.com/books/natural-language-processing-in-action-second-edition

Fortino, A. (2021). *Text Analytics for Business Decisions: A Case Study Approach.* Mercury Learning and Information.

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as data: A new framework for machine learning and the social sciences.* Princeton University Press.

Iezzi, D. F., Mayaffre, D., & Misuraca, M. (Eds.). (2020). *Text Analytics: Advances and Challenges (Studies in Classification, Data Analysis, and Knowledge Organization)* (1st ed.). Springer.

Ignatow, G., & Mihalcea, R. (2018). *An Introduction to Text Mining: Research Design, Data Collection, and Analysis.* Sage Publications.

Sailunaz, K., Dhaliwal, M., Rokne, J., & Alhajj, R. (2018). Emotion detection from text and speech: a survey. *Social Network Analysis and Mining*, *8*, 28. https://doi.org/10.1007/s13278-018-0505-2

Seyeditabari, A., Tabari, N., & Zadrozny, W. (2018). Emotion detection in text: a review. *ArXiv Preprint ArXiv:1806.00674*. https://doi.org/10.48550/arXiv.1806.00674

Simske, S., & Vans, M. (2021). *Functional Applications of Text Analytics Systems.* River Publishers.

Wang, Y., Feng, L., Song, X., Xu, D., & Zhai, Y. (2023). Zero-Shot Image Classification Method Based on Attention Mechanism and Semantic Information Fusion. *Sensors*, *23*(4), 2311. https://doi.org/10.3390/s23042311

Yu, Y., Zhuang, Y., Zhang, R., Meng, Y., Shen, J., & Zhang, C. (2023). ReGen: Zero-Shot Text Classification via Training Data Generation with Progressive Dense Retrieval. In A. Rogers, J. Boyd-Graber, & N. Oazaki (Eds.) *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 11782–11805). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.748