RADemics

Natural Language Processing and Text Mining Using NLTK and SpaCy for Sentiment and Topic Analysis

Pranali L. Katore, Archana Vasudeo Bhoyar, B. Persis Urbana Ivy
PRIYADARSHINI BHAGWATI COLLEGE OF ENGINEERING, MOTHER TERESA INSTITUTE OF ENGG AND TECHNOLOGY

Natural Language Processing and Text Mining Using NLTK and SpaCy for Sentiment and Topic Analysis

[1]Pranali L. Katore, Assistant Professor, Computer Science and Engineering, Priyadarshini Bhagwati College of Engineering, Nagpur, Mobile number: 824 800 2831, Mail id: pranalikrose05@gmail.com.

[2]Archana Vasudeo Bhoyar, Assistant Professor, Computer Science and Engineering, Priyadarshini Bhagwati College of Engineering, Nagpur, Mobile number: 824 800 2831, Mail id: archanapauyar475@gmail.com.

[3]B. Persis Urbana Ivy, Dean (CSE & Allied branches), Mother Teresa Institute of Engg and Technology, Melumoi (Post), Palamaner - 517408, Mail id: urbana23@gmail.com.

Abstract

The exponential growth of unstructured textual data across digital platforms has necessitated the development of scalable and interpretable frameworks for automated text analysis. This chapter presents a comprehensive exploration of Natural Language Processing (NLP) and text mining methodologies with a specific focus on sentiment classification and topic modeling using NLTK and SpaCy. These two widely adopted Python libraries are analyzed in terms of their linguistic processing capabilities, computational efficiency, and compatibility with external machine learning frameworks. The chapter introduces an integrative framework that combines the linguistic depth of NLTK with the performance-driven architecture of SpaCy to construct end-to-end pipelines for extracting sentiment polarity and latent thematic structures from diverse corpora. Through rigorous comparative benchmarking, the chapter evaluates preprocessing techniques, annotation quality, and model outcomes across various domains, including news articles, customer reviews, and social media streams. Emphasis is placed on the role of linguistic preprocessing in enhancing semantic coherence, as well as on visualization strategies that aid in interpreting sentiment trends and topic distributions. The study also addresses domain adaptability, multilingual processing, and integration with modern deep learning libraries to improve both analytical precision and system scalability. Experimental results and case studies demonstrate the effectiveness of hybrid pipelines in real-world applications, offering valuable insights for researchers and practitioners in computational linguistics, data science, and artificial intelligence.

Keywords: Natural Language Processing, Sentiment Analysis, Topic Modeling, Text Mining, NLTK, SpaCy

Introduction

The vast proliferation of digital content across social media, online forums, customer reviews, and news portals has transformed text data into a critical resource for analytical and decision-making processes [1]. Extracting meaningful information from such unstructured data requires

advanced computational methodologies capable of interpreting linguistic patterns, semantic relationships, and contextual nuances [2]. Natural Language Processing (NLP) and text mining have become indispensable in addressing this challenge by enabling automated systems to process, understand, and derive insights from large volumes of textual information [3], [4]. These technologies facilitate a range of applications, including sentiment analysis, topic modeling, information retrieval, summarization, and entity recognition, thereby contributing significantly to fields such as business intelligence, healthcare analytics, political science, and digital marketing [5].

In this evolving landscape, Python has emerged as a dominant programming language due to its extensive ecosystem of NLP libraries [6]. Among these, the Natural Language Toolkit (NLTK) and SpaCy have gained prominence for their linguistic depth and computational efficiency, respectively [7]. NLTK provides an extensive collection of corpora, grammars, and processing modules tailored for academic research and teaching. It supports intricate linguistic analysis, including morphological parsing, syntax trees, and semantic role labelling [8]. Conversely, SpaCy emphasizes speed, scalability, and production readiness, offering high-performance tools for tokenization, dependency parsing, part-of-speech tagging, and named entity recognition [9]. Together, these libraries represent a versatile toolkit for constructing sophisticated NLP pipelines suitable for diverse analytical tasks. Their integration allows for the development of modular, efficient, and domain-adaptive systems capable of handling real-world text processing challenges [10].

Sentiment analysis is a focal point within NLP, aimed at identifying and categorizing opinions expressed in text [11]. It enables organizations to gauge public sentiment, monitor brand perception, and predict consumer behavior [12]. Accurate sentiment classification depends heavily on preprocessing quality, feature selection, and context-aware algorithms. Techniques such as polarity detection, emotion tagging, and aspect-based sentiment analysis are implemented using rule-based systems, traditional machine learning, or deep learning architectures [13]. The synergy between linguistic tools like NLTK and SpaCy enhances sentiment analysis by providing accurate tokenization, part-of-speech tagging, and lemmatization, which form the foundation for sentiment lexicon mapping and syntactic feature extraction [14]. These enriched features enable improved sentiment detection across noisy and informal text formats, such as tweets or customer feedback [15].