# Chapter 9
# A Comprehensive Survey on Text Mining From Theory to Practice

**Danial Zare**
https://orcid.org/0009-0009-8190-328X
*University of Alcala, Alcala de Henares, Spain*

**Luis Fernandez-Sanz**
https://orcid.org/0000-0003-0778-0073
*University of Alcala, Alcala de Henares, Spain*

**Vera Pospelova**
https://orcid.org/0000-0001-5801-1923
*University of Alcala, Alcala de Henares, Spain*

**Ines López-Baldominos**
https://orcid.org/0000-0001-8345-5243
*University of Alcala, Alcala de Henares, Spain*

## ABSTRACT

*Text mining refers to the process of extracting useful information from large volumes of unstructured text data. This paper presents a comprehensive survey of text mining, covering foundational theories and practical applications across various domains within the field of Natural Language Processing (NLP). The study begins by examining the core challenges and historical development of text mining, providing context through an exploration of major areas where text mining techniques have significantly evolved. We offer an in-depth, step-by-step analysis of key algorithms in the text mining pipeline, beginning with data collection and preprocessing, moving through feature generation and selection, and highlighting their roles in transforming*

DOI: 10.4018/979-8-3693-9606-3.ch009

*raw text into valuable insights. This survey serves as a guide for researchers and practitioners by detailing methodological approaches and considerations at each stage of the text mining process. Additionally, Pseudocode and Python implementations of algorithms are provided, facilitating the application of these methods in real-world scenarios.*

## INTRODUCTION

In the digital age, vast amounts of unstructured data are generated daily, with text being one of the most prevalent forms. From social media posts and news articles to scientific journals and customer feedback, the potential insights contained within textual data are immense. However, extracting meaningful information from such a massive volume of raw text presents significant challenges. This is where text mining becomes crucial. Text mining, widely used in knowledge- based organizations, is the process of examining large sets of documents to discover new information or help answer specific research questions. identifies remaining facts, relationships, and assertions. Once extracted, this information is converted into a structured form that can be further analyzed or presented directly using clustered HTML tables, mind maps, charts, etc. (Abdusalomovna et al., 2023).

Text is the most widely used means of communication today (Ly et al., 2020). By leveraging advanced computational techniques, text mining allows organizations and researchers to turn unstructured data into valuable knowledge, providing a foundation for informed decision-making, predictive analysis, and data-driven innovation. Most of the time, text is analyzed by a human who can read the text and transform it into structured information. It can take a lot of time when it comes to analyzing thousands of sentences. In fact, NLP is a tract of Artificial Intelligence and Linguistics, devoted to make computers understand the statements or words written in human languages. It came into existence to ease the user's work and to satisfy the wish to communicate with the computer in natural language, and can be classified into two parts i.e. Natural Language Understanding or Linguistics and Natural Language Generation which evolves the task to understand and generate the text (Khurana et al., 2023).

Text mining and NLP have rapidly gained traction in both academia and industry, making them some of the most dynamic and intriguing fields in artificial intelligence. The allure of these areas lies in their ability to make sense of the vast amounts of unstructured text data generated daily, from social media posts and blogs to research papers and customer feedback. As these fields evolve, the demand for more sophisticated tools to analyze and understand human language grows, resulting in significant advancements in machine learning models, deep learning, and