Knowledgeable Research (KR) 2026, Special Vol. (05): 03
ISSN: 2583-6633, CODEN: KRABAU
International open-access peer-reviewed refereed monthly journal
https://knowledgeableresearch.com/index.php/1

Digital Literary Analysis Using Text Mining and Sentiment Analysis

[1]Nilesh Ingale* and Amol Mandle[2]

[1]Indira Gandhi Senior College, CIDCO, Nanded, Mahaarashtra-431603

[2]Sambhajirao Kendre Mahavidyalaya Jalkot, Latur, Maharshtra-413532

Abstract:
This paper explores the use of text mining and sentiment analysis techniques to analyze Amazon product reviews. Python libraries such as Pandas, Matplotlib, NLTK, and Scikit-learn are used to preprocess and analyze customer feedback. Text preprocessing techniques including tokenization, stop-word removal, and stemming or lemmatization are applied. The analysis focuses on sentiment and emotion distribution, as well as word frequency patterns, to gain insights into customer opinions and overall product perception.

Keywords*: Artificial intelligence, Text mining, Sentiment analysis.*

Received: 11 December 2025
Accepted: 24 January 2026
Published: 30 January 2026
***Corresponding Author:**
Nilesh Ingale
Email: ingaleyash7@gmail.com

Introduction

Artificial Intelligence (AI) is a field of computer science that enables machines to perform tasks requiring human intelligence, such as language understanding, pattern recognition, and decision-making. AI integrates concepts from mathematics, statistics, and engineering to develop systems that learn from data and adapt to new situations. AI has wide-ranging applications across various domains, including healthcare, education, finance, robotics, and e-commerce. In the digital era, AI-driven techniques such as text mining and sentiment analysis have significantly enhanced literary and textual studies by allowing the systematic analysis of large text datasets and uncovering patterns beyond manual analysis.

1.1 AI in Literature

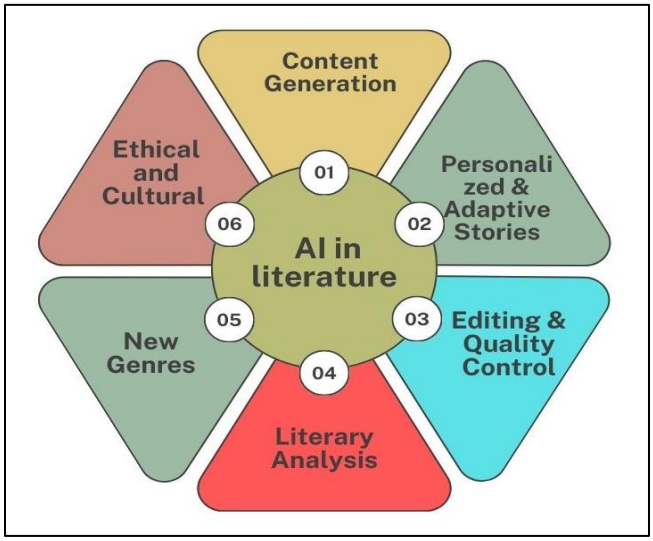


Fig.1. AI in Literature

AI can play crucial role in literature such as creating; analysing and experiencing the stories, articles, editorials etc. following are few application of AI in literature analysis.

a) Content creation

AI tools help writers overcome creative blocks by generating plot ideas, characters, and dialogue, accelerating the drafting process and enabling broader creative exploration.

E.g. ChatGPT, Gemini, Jasper, Copy.ai, Writesonic, Gramerlly [2].

b) Personalized adaptive stories

AI can adapt narratives in real time to individual reader choices, resulting in immersive and interactive storytelling experiences.

c) Editing and quality control

AI-powered tools enhance written work by improving clarity, ensuring consistency, and refining language to produce more polished and professional texts.

E.g. Gramerlly, Prowritingaid, Hemingway Editor and Narrato[3].

d) Literary analysis

AI analyses complex texts to generate summaries, explain key themes, and provide translations, thereby making classic works accessible to broader audiences.

E.g. EliCit, ResearchRabbit and Scite.ai

e) New genres

The emergence of AI may give rise to entirely new literary styles and forms, challenging conventional understandings of authorship.

f) **Ethical and cultural**

AI will help for plagiarism and copyright issues. In context of cultural aspect, use of AI might affect human feelings, emotions. However, use of AI will make literature field transparent.

1.2. Text mining and sentiment analysis

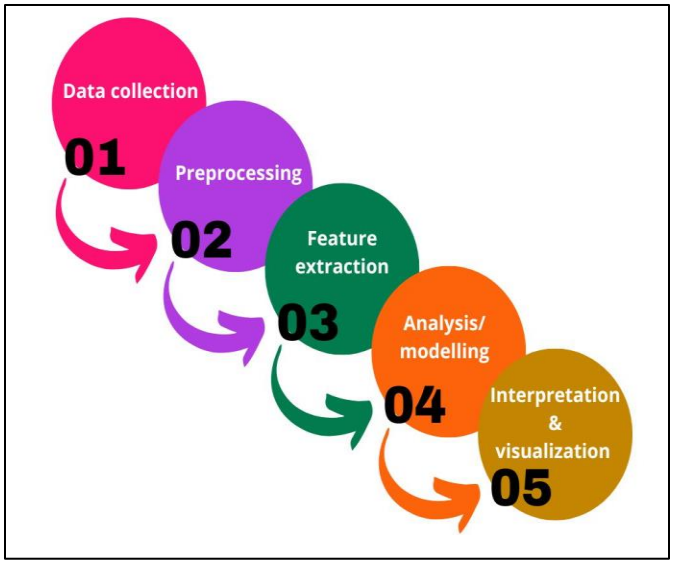
The digital age has transformed literary studies by introducing computational methods that complement traditional close reading. Among the most influential of these methods are text mining and sentiment analysis, which enable scholars to analyse large corpora of texts systematically, revealing patterns that would be difficult or impossible to detect manually [4].

1.2.1. Text mining

Text mining is computational technique that is used to extract useful information, patterns, trends and insights from large volume of unstructured text data. Unstructured data referrers to emails, text messages, web pages, PDFs and word documents social media posts, reviews, presentations etc. Text mining turns this data in to structured data that can be analysed by using AI techniques. Text mining can be carried out by using Python and R programing languages. There are numbers of libraries associated with python as well as R for text mining [5-6].

Python: NLTK, spaCY, scikit-learn, Gensim, Hugging Face Transformers.

The typical steps involved in text mining shown in following workflow diagram.



(i) Data collection

The very first step in text mining is collection of data. The data in documents (PDF and word) format, web pages, social media posts, reports etc are collected.

(ii) Pre-processing

It prepares unstructured data for further analysis such as

(a) Tokenization

It is the process of splitting continuous text into smaller meaningful parts, called tokens, such as words or sentences, for further analysis.

(b) Lower casing

It is text processing step that convert all text to lowercase to create consistency, reduce vocabulary volume and treat words like “Tiger”, and “tiger” as the same.

(c) Stop-word removal

In this technique commonly, occurring words that are less meaningful removed from data.

E.g. The, is, am, are etc.

(d) Stemming/lemmatization

Stemming and lemmatization are text pre-processing techniques used to reduce words to their base or root form, helping normalize text for analysis.

(iii) Feature extraction

Feature extraction converts raw text in to numerical representations that AI models can use for analysis. It improves performance and efficiency of AI models. Few feature extraction techniques are listed below.

(a) Bag of words(BoW)

The Bag-of-Words (BoW) model represents text as numerical vectors based on word frequency, treating documents as unordered sets of words while ignoring grammar, syntax, and word order, and is commonly used in text classification and information retrieval.

(b) Term Frequency–Inverse Document Frequency(TF-IDF)

TF-IDF is an NLP method that measures word importance by balancing how frequently a term appears in a document against how uncommon it is across the full document collection

(c) Word embedding

Word embedding represent words as low-dimensional numerical vectors, placing similar

words closer together to capture semantic and syntactic relationships for efficient text processing.

E.g. Word2Vec, GloVe, FastText

(iv) Analysis/ modelling

Analysis or modelling in Natural Language Processing (NLP) refers to applying statistical, machine learning, or deep learning techniques to extracted text features in order to identify patterns, make predictions, or generate insights from textual data.

(a) Text classification

Text classification also referred to as text tagging or text categorization is the task of organizing text into structured categories. Using Natural Language Processing (NLP), classifiers automatically analyse textual content and assign predefined labels based on its meaning.

Examples: spam detection, sentiment analysis, topic categorization

(b) Clustering

Text clustering is the task of organizing documents into groups based on content similarity, enabling the discovery of underlying patterns and trends that may not be immediately apparent.

Example: document organization, topic discovery

(c) Topic modelling

Topic modelling in NLP is an unsupervised machine learning approach that identifies latent topics, groups of related words within large text collections, enabling the organization, understanding and summarization of unstructured data without the need for labelled examples.

Common model: Latent Dirichlet Allocation (LDA)

(d) Sentiment analysis

Sentiment analysis is the task of examining textual data to identify the underlying emotional tone, classifying it as positive, negative, or neutral, and in some cases detecting more fine-grained emotions such as happiness, sadness, anger, or frustration.

(e) Named entry recognization(NER)

Named Entity Recognition (NER) in NLP is the task of detecting and classifying key pieces of information, known as entities, within text, such as names of people, locations, organizations, and dates. It converts unstructured text into structured data, supporting applications like text summarization, knowledge graph construction, and question answering.

(v) Interpretation and visualization

Interpretation in NLP focuses on enabling computers to understand human language in terms of meaning and context, while visualization translates complex text analysis results—such as sentiment, topics, and structural patterns—into clear visual forms like charts, graphs, and diagrams. Together, they bridge the gap between raw textual data and actionable human insights by presenting extracted patterns through intuitive tools such as word clouds, syntax trees, and analytical dashboards, which support informed decision-making.

2. Methodology

In the present study, dataset is analyzed and explored by using Jupyter notebook with Python code, an application of Anaconda open source Artificial intelligence, data science distribution platform developed by Anconda Inc. [7]. The dataset used in this study contains 21,255 Amazon product reviews designed for sentiment analysis, emotion detection, and topic modelling, with each review enriched by sentiment metrics, emotion probability scores, and clustering assignments . This dataset is collected from Kaggle platform[8].

3. Result and discussion

The main objective of this study is to extract meaningful insights from the amazon predict review dataset by examining the complex interplay between product reviews, emotions, and sentiments in global market. Through the application of advanced text mining and sentiment analysis techniques, this study aims to identify patterns, trends, and sentiment distributions across diverse product categories. The findings have the potential to support improved customer-focused strategies, product innovation, and marketing initiatives within e-commerce ecosystem.

3.1 Data pre-processing

The loaded data cleaned in this step by searching missing values, removing duplicates and irreverent columns. It is observed that there were 21255 rows with 19 columns. The number of missing values, duplicate values and irreverent columns are zero in the present dataset.

3.2 Text normalization

After the data cleaning, the text normalization carried out by

(a) Lowercasing

df['cleaned_review']=df['cleaned_review'].str.lower( )

(b) Removing special character

df['cleaned_review']=df['cleaned_review'].str.replac e('[^\w\s]', '')

(c) Removing stop words

stop_words=set(stopwords.words('english'))

df['cleaned_review']=df['cleaned_review'].apply(la mbdax:''.join(word for word in word_tokenize(x) if word not in stop_words))

3.3 Exploratory data analysis

(a) Distribution of ratings and sentiments

The rating and reviews given by customers on product distributed in positive, negative and neutral sentiments seen in bar plot shown in Fig.3.

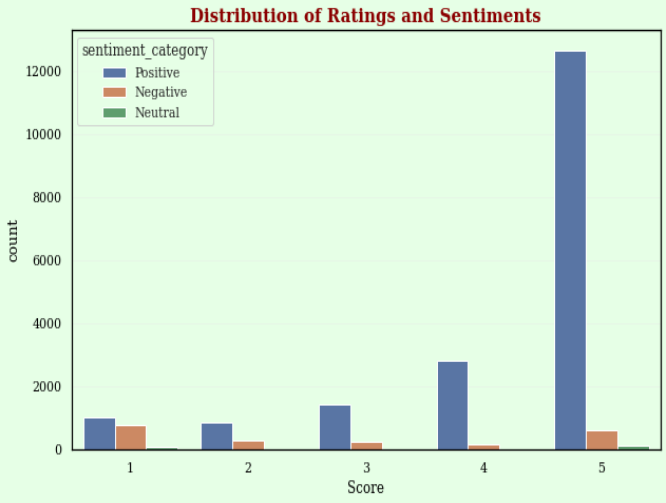


Fig. 3. Distribution of ratings and sentiments

(b) Word frequency analysis

It provide a quick way to visualize the most common words in a text, helping reveal key themes, trends, and sentiment in large datasets such as customer feedback or surveys, and making complex text data easier to understand and communicate. The word frequency analysis have carried out to look for the words, which are most frequently received in customers review in the present dataset, and it is plotted by using word cloud plot as shown in Fig. 4.

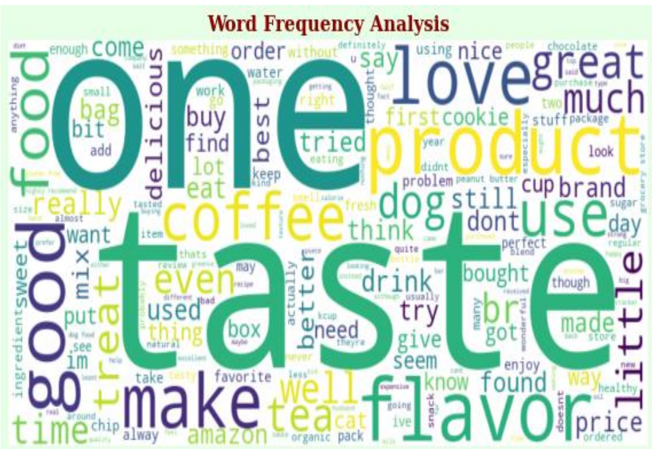


Fig. 4. Word cloud plot for amazon product reviews

The word cloud plot shows that, most frequently listed words in the customers product reviews include, “taste”, “one”, “product”, “favour”, “good”, “food”, “make”, “love”, “great”, “coffee”, “well”, “tea”, “use”, “brand”, “well”, “delicious”, “treat”, “really” “price”, “little”, “great” etc. It also shows most of the words belongs to positive sentiment.

3.4 Sentiment analysis.

In the present study, sentiment analysis is performed using sentiment labelling as well sentiment distribution.

(a) Sentiment labelling

It carried out by two ways, using customer ratings i.e. by numbers as well as by using customer review means by text.

(b) Sentiment distribution

The sentiment distribution plot shows that, the most of the reviews are with positive sentiment for the products purchased with number 17500 and around 1600 reviews are under negative sentiment category, few neutral sentiments also listed in the reviews. The sentiment distribution bar plot is plotted in the Fig.5.

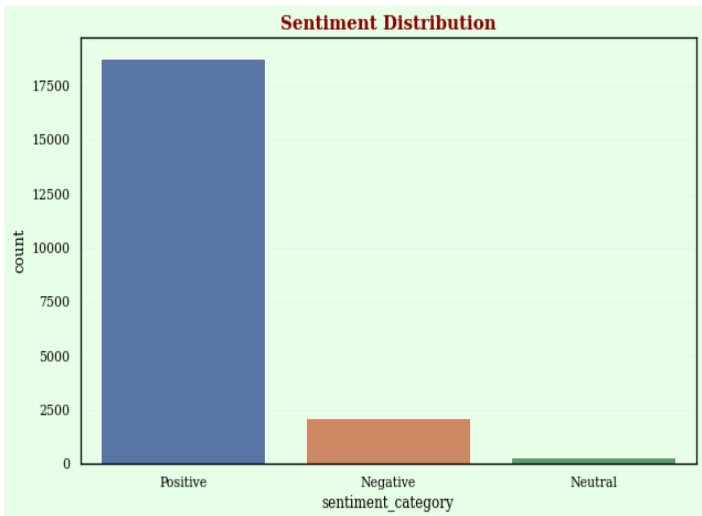


Fig. 5. Sentiment distribution bar blot

3.5 Emotion analysis

From the present dataset, the customer’s emotions are identified and extracted.

(a) Emotion distribution

The present dataset consists of human emotion trust, fear, anger, anticipation, surprise, sadness, joy and disgust. It shows that review received from customers are positive, negative and neutral. This distribution of emotion will help the company to reach out the customers and solve their issues regarding service of company, quality of product, delivery speed etc. In addition, these emotions visualized using bar plot as shown in Fig. 6.

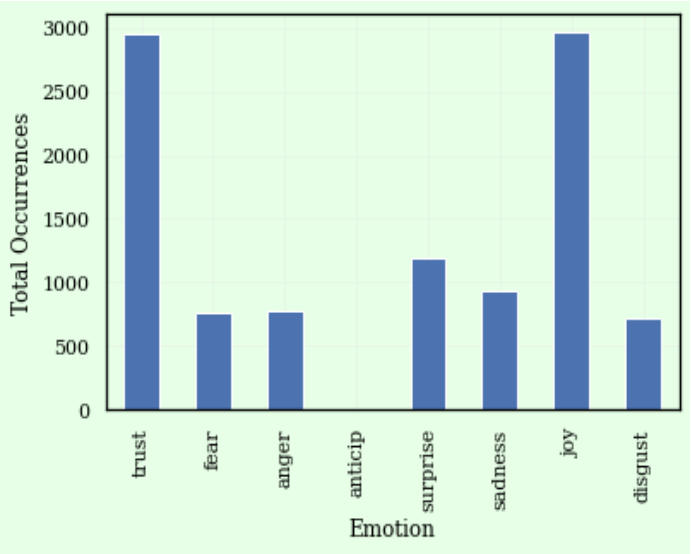


Fig.6. Emotion distribution bar plot.

It shows the most of the reviews have emotion of trust and joy. While disgust and anger are with less numbers.

4. Conclusion

In conclusion, this study demonstrates the effective use of text mining and sentiment analysis to analyse Amazon product reviews using Python libraries. The analysis of 21,255 reviews reveals that most customer feedback is positive, with trust and joy being the dominant emotions. Overall, the findings indicate high customer satisfaction, as reflected by the predominance of positive sentiments and five-star ratings.

References

1. Artificial intelligence: A modern approach 3rd edition, 2015, Pearson, India.
2. https://www.getblend.com/blog/10-best-ai-tools-to-use-for-content-creation/
3. Boondoggle Studio, LLC. Hemingway Editor. Hemingway Editor, 2023-2026, Accessed 21 Jan. 2026.
4. Weiss, S. M., Indurkhya, N., Zhang, T., & Damerau, F. J. (2010). Text Mining: Predictive Methods for Analyzing Unstructured Information.
5. Sebastiani, F. (2002), Machine learning in automated text categorization, ACM Computing Surveys, 34(1), 1–47.
6. Benoit, K., et al. (2018), quanteda: An R package for the quantitative analysis of textual data, Journal of Open Source Software, 3(30), 774.
7. Anaconda Software Distribution. (2020). Anaconda Documentation. Anaconda Inc. Retrieved from https://docs.anaconda.com/
8. https://www.kaggle.com/datasets/srinandanv/amazon-product-reviews/code/data, Accessed 21 Jan. 2026

.