International Journal of Technology Research and Management
ISSN (Online): 2348-9006
Vol 10 Issue 05 May 2023

Text Analysis with different approaches in social media environment: A Review

Himanshu Yadav[1] and Raju Baraskar[2]
Department of Computer Science & Engineering[1,2]
University Institute of Technology, RGPV, Bhopal[1,2]
himanshuyadav86@gmail.com[1], rajubaraskar@rgtu.net[2]

Abstract: *The integration of technology into everyday life has driven an exponential increase in data creation and management. This surge in data, particularly human-generated in forms such as text, audio, and video, has led to a growing interest in automated methods to extract useful information from large volumes of unstructured data. Text analysis, encompassing techniques like data mining, machine learning, and computational linguistics, has emerged as a crucial tool for extracting information and patterns from textual data. This paper explores various methodologies in text analysis, including sentiment analysis, information extraction, natural language processing (NLP), text classification, and deep learning. Each methodology is examined for its applications, particularly in domains like social media, biomedical fields, e-commerce, healthcare, and agriculture. The paper discusses the challenges faced by researchers in text analysis and the potential future scope for these methods. By leveraging advanced algorithms and computational techniques, text analysis can significantly enhance the efficiency and effectiveness of data processing across multiple industries.*

Keywords:*Text analysis, twitter, machine learning, deep learning*.

1. INTRODUCTION

Technology is now deeply integrated into everyone's lives, with nearly every activity, from phone calls to satellite launches, evolving exponentially due to technological advancements. The increasing capacity to create and manage information has significantly driven this technological growth. The National Security Agency of the United States reports handling an average of 1826 petabytes daily over the Internet. With the rapid rise in data and information transmitted online, it has become necessary to regulate and streamline this flow. Various commercial and social applications have been introduced to address these needs. Data and information aspects, such as security, research, and sentiment analysis, are invaluable to organizations, governments, and the public. Optimized techniques aid in tasks like classification, summarization, and data management, with machine learning and deep learning algorithms being instrumental in processing the vast amounts of available information.

Text analysis has become indispensable in the era of big data, where extracting relevant information from vast textual sources is crucial for decision-making and knowledge discovery. From sentiment analysis in social media to extracting medical insights from clinical notes, text analysis has a wide array of applications. This paper aims to provide an in-depth review of text analysis methodologies, focusing on their evolution, current state, and future prospects.

Human-generated data in the form of text, audio, video, and more is rapidly increasing. This surge has led to a growing interest in methods and tools that can automatically extract useful information from enormous amounts of unstructured data. One crucial method is text analysis, which combines techniques such as data mining, machine learning, and computational linguistics to extract information and patterns from textual data. Traditionally, text analysis was done manually, with humans reading and searching for useful information. However, automated text analysis is more logical, providing efficiency in terms of speed and cost.

Text analysis is applied in various industries. On social media, text data is generated in posts, blogs, and web forum activities. Despite the vast quantity of data available, the relatively low proportion of high-quality content remains a problem that text analysis can address. In the biomedical field, effective text-mining and classification methods are needed. On e-commerce websites, text analysis prevents repetitive information to the same audience and improves product listings through reviews. In healthcare, applications include identifying healthcare topics from personal messages, classifying online data, and analyzing patient feedback. The agriculture industry uses text analysis for classifying regulations, ontology-based clustering, and analyzing public opinions. Text analysis is also utilized in detecting malicious web URLs that evolve over time and have complex features. This paper discusses the use of text analysis across various domains, considering different areas of application, widely used methodologies and techniques, challenges faced by researchers, and future scope for text analysis methods.

2. TEXT ANALYSIS

Text analysis [1] is the process of deriving high-quality information from text. The demand for text analysis has significantly increased in recent years, evolving continuously alongside big data analytics. Various sectors can greatly benefit from these techniques, as analyzing large volumes of data is both a necessity and an advantage. This section discusses some important and widely used techniques for analyzing textual data.

Text analysis [2] refers to the process of using computational methods to analyze and extract meaningful information from textual data. This field encompasses a variety of techniques and methodologies aimed at understanding and interpreting text

2.1. Sentiment Analysis (SA)

Sentiment analysis (SA) [3] is a crucial technique in text analysis. It extracts underlying opinions from textual data and is also known as opinion mining. SA is widely used across domains such as e-commerce, blogs, online social media, and microblogs. The motives behind SA can be broadly divided into emotion recognition and polarity detection. Emotion detection focuses on extracting emotion labels, while polarity detection is a classifier-oriented approach with discrete outputs (e.g., positive and negative).

There are two main approaches to SA: lexicon-based (dictionary-based) and machine learning (ML). ML approaches are further classified into supervised and unsupervised learning. Lexicon-based approaches use word maps, while ML considers SA as a classification problem and employs established techniques. In lexicon-based approaches, the overall sentiment score is calculated by dividing the sentiment frequency by the sum of positive and negative sentiments. In ML approaches, major techniques include Naïve Bayes (NB) classifier and support vector machines (SVMs), which use labeled data for classification. ML-based SA has an edge over lexicon-based approaches as it doesn't require costly word dictionaries, though it does need domain-specific datasets, which can be a limitation. After data preprocessing, feature selection is performed, followed by obtaining final results through the adopted approach.

2.2. Information Extraction

Information extraction (IE) [4] aims to extract predefined data types from text documents. IE systems identify objects by extracting relevant information from text fragments and then organizing the extracted pieces into a structured framework. Methods such as DiscoTEX (Discovery from TextEXtraction) convert structured data into meaningful information to discover knowledge from it.

2.3. Natural Language Processing (NLP)

NLP, [2,3]a part of artificial intelligence, transforms imprecise and ambiguous messages into precise and unambiguous ones. NLP is often employed to mine documents for insights to develop conclusions. NLP helps perform analyses such as named-entity recognition (NER), identifying relationships and other information to pinpoint key concepts. However, NLP lacks a comprehensive dictionary for all named entities used for identification.

NLP has been applied to bridge gaps between NLP and various fields by considering topics of mutual interest. NLP generates significant textual data in acquiring information, understanding attributes, and making activities more efficient.

2.4. Text Classification

Text classification [5] is a four-step process comprising feature extraction, dimension reduction, classifier selection, and evaluation. Feature extraction can be done using techniques such as term frequency and Word2Vec; dimensionality reduction is performed using methods such as principal component analysis and linear discriminant analysis. Choosing a classifier is critical, with deep learning approaches often surpassing other machine learning algorithms. The evaluation step assesses model performance

using parameters like the Matthews correlation coefficient (MCC), area under the ROC curve (AUC), and accuracy.[6,7]

2.5. Deep Learning

Deep learning [8], a subset of machine learning, trains a data model to make predictions about new data. It has a layered architecture where input data is transformed through middle levels by applying algorithms to extract and transform features. Studies have shown that deep learning models, such as long short-term memory models, outperform traditional machine learning algorithms when applied to analyze data.

In other wordsDeep learning, a subset of machine learning, employs neural networks to enable computers to learn from examples, similar to how humans do. In deep learning, models are trained to perform tasks like classification or regression using data such as images, text, or sound. These models often achieve state-of-the-art accuracy, sometimes surpassing human performance.

How Does Deep Learning Work?

Deep learning [9] models are built on neural network architectures, which are inspired by the human brain. A neural network comprises interconnected nodes, or neurons, organized in layers that connect inputs to the desired outputs. The neurons between the input and output layers are called hidden layers. The term "deep" refers to the number of hidden layers within the neural network. Deep learning models can contain hundreds or even thousands of these hidden layers, enabling them to process complex patterns and large amounts of data.

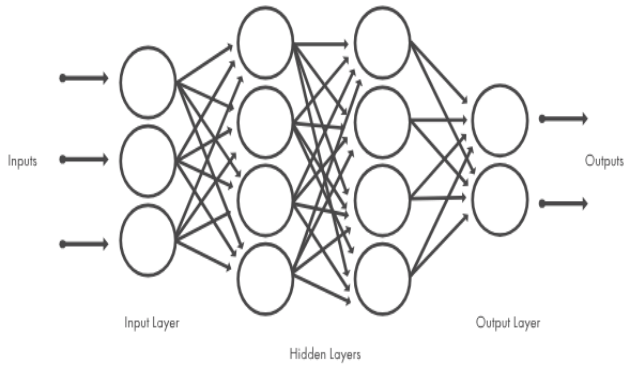

Figure 1 : Deep Learning Model

Text analysis involves the process of converting unstructured text into structured data to derive insights and support decision-making. This includes several key tasks such as:

1. **Tokenization:** Splitting text into individual words or phrases.
2. **Stop Words Removal:** Filtering out common words that do not carry significant meaning.
3. **Stemming and Lemmatization:** Reducing words to their base or root form.
4. **Vectorization:** Converting text into numerical representations.

3. DIFFERENT APPROACHES OF TEXT ANALYSIS

The basic four methods are available in processing of text analysis. [10] Some of them has been discussed here

3.1. Statistical Methods:

- **Bag of Words (BoW):** Represents text as a collection of words, disregarding grammar and word order.
- **Term Frequency-Inverse Document Frequency (TF-IDF):** Weighs the importance of words based on their frequency in a document relative to their frequency in the entire corpus.

3.2. Machine Learning Approaches:

- **Supervised Learning:** Techniques such as Naive Bayes, Support Vector Machines (SVM), and neural networks, which require labeled training data.
- **Unsupervised Learning:** Clustering algorithms like K-means and Latent Dirichlet Allocation (LDA) for topic modeling.

3.3. Deep Learning Approaches:

- **Recurrent Neural Networks (RNN):** Suitable for sequential data but can struggle with long-term dependencies.
- **Long Short-Term Memory Networks (LSTM):** Addressing the limitations of RNNs by maintaining long-term dependencies.
- **Transformers**: Revolutionized NLP with models like BERT and GPT, providing state-of-the-art performance in various text analysis tasks.

3.4. Hybrid Approaches:

- Combining statistical methods, machine learning, and deep learning to leverage the strengths of each technique.

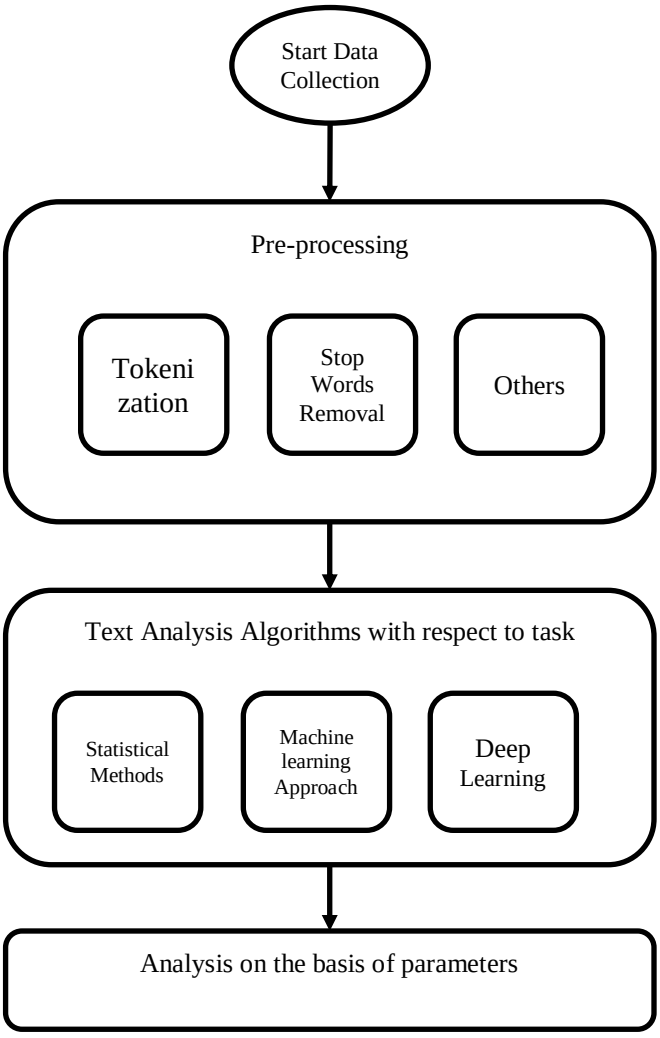

Figure 2: Approach used in text analysis

Above figure have 3 major section in which all are dependent to the previous one.

4. APPLICATIONS OF TEXT ANALYTICS

1. **Fraud Detection**

Insurance and finance sectors use text analytics and mining to process claims quickly and detect fraud by integrating text analysis results with structured data.

2. **Social Media Analysis**

Text mining helps assess social media performance by tracking and interpreting content from various sources, such as news, blogs, and emails. It evaluates posts, likes, and connections to understand public reactions and market trends.

3. **Customer Care Service**

Companies use text analytics, especially natural language processing (NLP), to enhance customer experience by analyzing surveys, feedback, and conversations. This helps reduce response times and efficiently resolve customer issues.

4. **Knowledge Management**

Managing vast text data, particularly in healthcare, is crucial. Text analytics organizes and makes accessible massive amounts of data, aiding in collaborative efforts during events like epidemics to trace sources and manage outbreaks.

5. **Risk Management**

Text mining tools assist businesses in staying current with industry trends and improving risk mitigation by aggregating and linking insights from various text data sources, ensuring timely access to relevant information.

6. **Market Research**

Text analytics enables companies to understand market trends, customer needs, and competitive landscapes by analyzing data from surveys, reviews, social media, and forums. This helps in developing strategies and products aligned with market demands.

6. **Content Recommendation**

Streaming services, news portals, and e-commerce platforms use text analytics to recommend personalized content to users based on their preferences and past behavior. This enhances user engagement and satisfaction.

7. **Healthcare Diagnostics**

Analyzing medical records, clinical notes, and research papers using text analytics helps in diagnosing diseases, identifying treatment options, and predicting patient outcomes. This aids in providing personalized healthcare solutions.

8. **Human Resources**

Text analytics streamlines recruitment by analyzing resumes and cover letters to identify the best candidates. It also helps in employee sentiment analysis by evaluating feedback and surveys, leading to better HR decisions.

9. **Academic Research**

Researchers use text analytics to conduct literature reviews, identify research gaps, and analyze academic papers. This facilitates the discovery of new insights and the advancement of knowledge in various fields.

10. **Legal Document Review**

Law firms and legal departments use text analytics to review and analyze large volumes of legal documents, contracts, and case files. This improves efficiency in legal research, e-discovery, and compliance checks.

11. **Supply Chain Management**

Text analytics helps in monitoring and analyzing supply chain data from various sources, such as emails, reports, and shipment logs. This improves logistics, demand forecasting, and inventory management.

12. **Opinion Mining**

Companies analyze customer opinions and sentiments from reviews, social media, and forums to gauge public perception of their products and services. This helps in making informed business decisions and improving customer satisfaction.

13. **Cybersecurity**

Text analytics is used to detect and analyze potential cyber threats by monitoring and analyzing communication channels, forums, and dark web discussions. This helps in identifying and mitigating security risks.

14. **Education**

Educational institutions use text analytics to evaluate student feedback, analyze academic performance, and improve curriculum design. This enhances the quality of education and student satisfaction.

5. CONCLUSIONS

Text analysis has become an indispensable tool in managing and deriving insights from the vast amounts of unstructured data generated daily. Techniques such as sentiment analysis, information extraction, natural language processing, text classification, and deep learning offer powerful means to handle complex textual data. These methodologies have broad applications across various domains, including social media, biomedical research, e-commerce, healthcare, and agriculture, showcasing the versatility and impact of text analysis.

Despite the advancements, several challenges remain, such as the need for domain-specific datasets, the complexity of natural language, and the high costs associated with certain approaches. Addressing these challenges requires continued research and innovation in algorithm development and computational resources. The future of text analysis lies in improving the accuracy, efficiency, and scalability of these techniques, enabling more sophisticated and real-time data processing capabilities.

In conclusion, text analysis represents a critical intersection of technology and data science, offering significant benefits in understanding and utilizing the wealth of information available in textual form. As technology continues to evolve, so will the methodologies and applications of text analysis, paving the way for more intelligent and automated data processing solutions.

REFERENCES

[1]. H. Q. Abonizio, E. C. Paraiso and S. Barbon, "Toward Text Data Augmentation for Sentiment Analysis," in IEEE Transactions on Artificial Intelligence, vol. 3, no. 5, pp. 657-668, Oct. 2022, doi: 10.1109/TAI.2021.3114390.

[2]. J. Tian, Y. Huang, Z. Guo, X. Qi, Z. Chen and T. Huang, "A Multi-Modal Topic Model for Image Annotation Using Text Analysis," in IEEE Signal Processing Letters, vol. 22, no. 7, pp. 886-890, July 2015, doi: 10.1109/LSP.2014.2375341.

[3]. P. Wang et al., "Classification of Proactive Personality: Text Mining Based on Weibo Text and Short-Answer Questions Text," in IEEE Access, vol. 8, pp. 97370-97382, 2020, doi: 10.1109/ACCESS.2020.2995905.

[4]. S. Cunningham-Nelson, M. Baktashmotlagh and W. Boles, "Visualizing Student Opinion Through Text Analysis," in IEEE Transactions on Education, vol. 62, no. 4, pp. 305-311, Nov. 2019, doi: 10.1109/TE.2019.2924385.

[5]. C. -W. Tseng, J. -J. Chou and Y. -C. Tsai, "Text Mining Analysis of Teaching Evaluation Questionnaires for the Selection of Outstanding Teaching Faculty Members," in IEEE Access, vol. 6, pp. 72870-72879, 2018, doi: 10.1109/ACCESS.2018.2878478.

[6]. H. Q. Abonizio, E. C. Paraiso and S. Barbon, "Toward Text Data Augmentation for Sentiment Analysis," in IEEE Transactions on Artificial Intelligence, vol. 3, no. 5, pp. 657-668, Oct. 2022, doi: 10.1109/TAI.2021.3114390.

[7]. X. Rong, C. Yi and Y. Tian, "Unambiguous Text Localization, Retrieval, and Recognition for Cluttered Scenes," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 3, pp. 1638-1652, 1 March 2022, doi: 10.1109/TPAMI.2020.3018491.

[8]. H. Wang, J. Cao and D. Lin, "Deep Analysis of Power Equipment Defects Based on Semantic Framework Text Mining Technology," in CSEE Journal of Power and Energy Systems, vol. 8, no. 4, pp. 1157-1164, July 2022, doi: 10.17775/CSEEJPES.2019.00210.

[9]. O. Wu, T. Yang, M. Li and M. Li, "Two-Level LSTM for Sentiment Analysis With Lexicon Embedding and Polar Flipping," in IEEE Transactions on Cybernetics, vol. 52, no. 5, pp. 3867-3879, May 2022, doi: 10.1109/TCYB.2020.3017378.

[10]. Y. Xu, Z. Yu, W. Cao and C. L. P. Chen, "Adaptive Dense Ensemble Model for Text Classification," in IEEE Transactions on Cybernetics, vol. 52, no. 8, pp. 7513-7526, Aug. 2022, doi: 10.1109/TCYB.2021.3133106.