UDC 004.89:004.93 DOI: https://doi.org/10.30837/IYF.CVSAMM.2024.006

RESEARCH OF SEMANTIC TEXT ANALYSIS METHODS AND MODELS

Kravchenko O.O.
Supervisor – Candidate of Technical Sciences, Associate Prof. Vechirska I.D.
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
e-mail: oleksii.kravchenko@nure.ua

This work provides a comprehensive review and evaluation of existing techniques for semantic text analysis. It explores a wide range of methodologies, including Bag-of-Words, TF-IDF, various word embedding models, semantic analysis techniques, sentiment analysis, and others. The analysis delves into the strengths, limitations, and applications of each method, highlighting their effectiveness in capturing semantic relationships and extracting meaningful insights from textual data. This work aims to provide valuable insights to inform the selection and implementation of semantic text analysis techniques, thus advancing the field of natural language processing.

In today's digital era, the exponential growth of textual data across various domains has propelled the development of sophisticated techniques for extracting meaningful insights from unstructured text. Natural Language Processing (NLP) is a branch of artificial intelligence (AI) that focuses on enabling computers to understand, interpret, and generate human language in a meaningful and useful way. NLP encompasses a wide range of tasks, including language understanding, language generation, information extraction, and facilitating human-computer interaction. These tasks include understanding the meaning of text or speech, recognizing entities and relationships within text, generating coherent responses, and extracting structured information from unstructured text data.

Semantic text analysis, a vital subfield of NLP, aims to decipher the semantic meaning embedded within textual content, enabling a wide range of applications such as sentiment analysis, information retrieval, document summarization, and machine translation [1]. Amidst the plethora of semantic text analysis methods available, researchers and practitioners are often confronted with the challenge of selecting the most appropriate approach for their specific task or application. The landscape of semantic text analysis is characterized by a diverse array of methodologies, ranging from traditional statistical techniques to state-of-the-art deep learning models.

The significance of conducting a comprehensive comparison analysis of these methods cannot be overstated. Such an endeavor not only facilitates a deeper understanding of the strengths and limitations of individual approaches but also serves as a roadmap for guiding future research directions in the field of NLP. By systematically evaluating and benchmarking different semantic text

analysis methods, researchers can gain valuable insights into their relative performance across various tasks, datasets, and evaluation metrics.

Having established the relevance and significance of semantic text analysis, we now turn our attention to a comprehensive review of existing methods in the field. This review encompasses a diverse range of techniques, each offering unique approaches to semantic analysis. Some of the prominent methods of semantic text analysis include:

Bag-of-Words (BoW): BoW represents text as a collection of words, disregarding grammar and word order but focusing on word frequency. This method is simple and efficient but lacks context and semantic understanding [2].

Term Frequency-Inverse Document Frequency (TF-IDF): TF-IDF assigns weights to words based on their frequency in a document relative to their frequency across all documents in a corpus [2].

Latent Semantic Analysis (LSA): LSA uses singular value decomposition to transform a term-document matrix into a lower-dimensional space, capturing latent semantic relationships between terms and documents.

Word Embeddings (Word2Vec, GloVe, FastText): Word embedding methods represent words as dense vectors in a continuous vector space, capturing semantic relationships between words based on their context [2].

Contextual Word Embeddings (ELMo, BERT): Contextual word embedding models generate word representations that are sensitive to the context in which they appear. These models use deep neural networks to capture contextual information, leading to more accurate representations for downstream tasks such as named entity recognition and sentiment analysis [3].

Semantic Role Labeling (SRL): SRL identifies the predicate-argument structure of a sentence, labeling words with their semantic roles such as agent, patient, or location. This method helps in understanding the meaning of sentences and is crucial for tasks like information extraction and question answering [3].

Named Entity Recognition (NER): NER identifies and classifies named entities mentioned in the text. It is essential for information extraction tasks and is often a precursor to more advanced semantic analysis.

Sentiment Analysis: Sentiment analysis methods classify text into positive, negative, or neutral sentiment categories. These techniques are widely used in social media monitoring, customer feedback analysis, and opinion mining.

Our analysis of existing semantic text analysis methods highlights the diverse array of techniques available for extracting meaningful insights from textual data. As a result of the research, we were able to extract the core advantages and disadvantages of each semantic text analysis method which are presented in Table 1. Our research underscores the importance of understanding the strengths and weaknesses of each method in the context of specific tasks and applications.

Table 1 – Comparison of semantic text analysis methods

Method	Advantages	Disadvantages
BoW	Simplicity and efficiency, interpretability	Lacks context and semantic understanding. Vulnerability to sparsity
TF-IDF	Highlights the importance of words, and is robust to document length	Lacks semantic understanding, and has limited handling of synonyms and polysemy
LSA	Reduces dimensionality; uncovers hidden patterns and semantic similarities	It may be difficult to interpret the extracted dimensions. Struggles with synonyms and polysemy
Word Embeddings	Has good semantic understanding, and considers contextual information	Requires large training data sets, and may struggle with out-of-vocabulary words
Contextual Word Embeddings	Sensitivity to context, task agnosticism	Requires a large amount of annotated pre-training data, and is computationally complex
SRL	Deep understanding of sentence structure, and rich semantic representation	Dependency on syntax and parsing, complexity, and ambiguity
NER	Information extraction, entity disambiguation	Ambiguity and variability, domain dependence
Sentiment Analysis	Deep understanding of textual data meaning	Subjectivity and context dependency, accuracy challenges

By leveraging the insights gained from our review, one can make informed decisions in selecting and implementing semantic text analysis techniques, ultimately advancing the state-of-the-art in natural language processing and enabling new opportunities for knowledge discovery, and innovation.

References:

1. Goddard C., Shalley A. Handbook of natural language processing / ed. by I. Nitin, J. Damerau. Boca Raton : Taylor & Francis, 2010. 93 p.

2. Demystifying Text Representation: BoW, TF-IDF, and Word Embeddings Explained // Medium. URL: https://medium.com /@datailm/2171d9feecfd (date of access: 21.02.2024).

3. Semantic Role Labeling: Unveiling the Meaning Behind Language // Medium. URL: https://medium.com/@evertongomede/a4d48d4986af (date of access: 22.02.2024).