ISSN (Online): 3033-0106

BUSINESS & MANAGEMENT COMPASS
Volume: 70 (2026), Issue: 1
DOI: https://doi.org/10.56065/9y517n52**, pp. 61-66**

Issues since 2024: https://bi.ue-varna.bg/ojs/index.php/bmc
Previous issues: http://journal.ue-varna.bg/en/home

Literature review of algorithms for determining semantic keyword sets

Yana PANCHEVA [1]

1 Department of Informatics, University of Economics - Varna, Varna, Bulgaria, yana.pancheva@ue-varna.bg

JEL: O14, P24, C41, E13

Key words:
semantic keyword grouping; natural language processing (NLP); word embeddings; clustering; SEO optimization

Abstract

The purpose of this paper is to present a literature review of algorithms and methods for determining semantic keyword sets in the context of search engine optimisation (SEO) and natural language processing (NLP). The practical significance lies in identifying applicable approaches for automated semantic keyword grouping in real business environments. Limitations are associated with the insufficient scientific coverage of the topic for morphologically complex languages such as Bulgarian. The research method is a desk study of existing publications. Expected results include a systematised overview of leading algorithmic approaches and identified research gaps.

Published by University of Economics – Varna

Citation: Pancheva, Y. (2026). Literature review of algorithms for determining semantic keyword sets. *Business & Management Compass.* University of Economics Varna, 70 (1), pp. 61 - 66. DOI: 10.56065/9y517n52

Received: 02.03.2026.
Revised: 26.03.2026.
Accepted: 30.03.2026.
Published: 31.03.2026.

1. Introduction

The purpose of this paper is to conduct a literature review of algorithms for determining semantic keyword sets and to identify existing research gaps. The practical significance lies in systematising the knowledge required for the subsequent development of an algorithm applicable in a real business environment. Limitations arise from the scarce research coverage of morphologically complex languages such as Bulgarian. The method is a desk study of scientific publications. The expected result is a clearly defined research gap justifying the need for further development.

Keyword selection and grouping are fundamental processes in building online content and optimising for search engines. Traditional approaches rely on frequency analysis without accounting for the semantic relationships between words, which often results in keyword lists that fail to reflect actual user searches. Contemporary literature demonstrates that Google, holding over 91% of the global market (Bianchi, 2024), increasingly focuses on understanding the meaning behind queries rather than literal word matching (Dean, 2024; Guelailia & Bouziane, 2024). In this context, Guelailia and Bouziane (2024) emphasise that modern search engines use semantic signals for ranking, shifting keyword selection toward topics, intentions, and context. Jindal (2025) introduces the concept of Generative Engine Optimisation (GEO), whereby language models automatically generate relevant keyword phrases based on semantics.

2. Research Methodology

The research is conducted as a desk study of scientific publications indexed in Google Scholar, Emerald Insight, DOAJ, and EBSCO for the period 2013–2025. The search targets publications on keyword extraction, semantic grouping, word embeddings, and NLP-based SEO optimisation. Peer-reviewed articles addressing semantic analysis algorithms, vectorisation methods, and clustering approaches are prioritized. The reviewed literature is organised into thematic blocks: frequency-based methods, vector models, clustering algorithms, and SEO applications. Table 1 presents the systematic mapping of the reviewed sources, summarising each study's main focus, key contribution, and relevance to the present review.

Table 1

Systematic mapping of reviewed literature

Author(s)	Year	Main Focus	Key Contribution	Relevance to Review
Ziakis et al.	2021	SEO and machine learning	ML improves strategic decision-making in search optimization	Frames transition from traditional to algorithmic SEO
Matošević et al.	2021	Web page classification in SEO	ML supports classification of web content for improved search relevance	Links SEO practice with automated content analysis
Vyawahare	2022	ML for complex problems	Presents ML as a general methodology for complex analytical tasks	Broad methodological justification for semantic grouping
Sancanin & Penjisevic	2022	AI in media content generation	AI plays a growing role in automated content creation	Context for automated semantic modeling
Kar	2023	AI in digital marketing	AI transforms marketing through automation and improved analytics	Semantic analysis as part of AI-driven marketing
Srivastava	2023	AI/ML in SEO	AI and ML improve SEO processes and performance	Directly links algorithms with search optimization

Vinutha & Padma	2023	NLP and ML in SEO	NLP captures user intent, context, and thematic relationships	Most relevant: connects semantic NLP with SEO
Shawn	2023	AI in backlink analysis	AI supports link-building and off-page SEO	Supporting evidence for expanding AI role in SEO
Bianchi	2024	Search engine market share	Statistical context on search engine dominance	Contextual support for SEO research importance
Dean	2024	Google ranking factors	Summarizes factors affecting Google rankings	Explains why semantic relevance matters in practice
Guelailia & Bouziane	2024	AI-enhanced SEO	Modern search engines rely on semantic signals over exact matching	Supports shift from lexical to semantic approaches
Jindal	2025	Generative Engine Optimization (GEO)	LLMs generate semantically relevant keyword phrases	Positions topic within latest AI-driven optimization

Source: Own elaboration, 2026

3. Literature Review

3.1. From Frequency Analysis to Semantic Understanding

Early keyword extraction approaches rely primarily on TF-IDF, which treats words as independent units without accounting for their semantic interrelations. Ziakis et al. (2021) argue that frequency-oriented approaches are becoming increasingly inadequate as search engines interpret queries semantically rather than lexically. Dean (2024) identifies semantic coherence and thematic density as ranking factors of growing importance among the over 200 signals used by Google. The consensus across the reviewed sources is that frequency analysis alone is insufficient as a tool for semantic keyword organisation, and that a shift toward context-aware, meaning-based algorithms is both necessary and ongoing.

3.2. Word Embedding Models

The shift toward semantically oriented methods is primarily associated with the introduction of Word2Vec by Mikolov et al. (2013), in which words are represented as numerical vectors reflecting their semantic relationships. Bojanowski et al. (2017) extend this concept with fastText, which builds vectors at the level of subword units (character n-grams), making it more robust for morphologically rich languages. Devlin et al. (2019) introduce BERT – an architecture generating contextualised word representations in which the meaning of a word depends on the surrounding sentence. Vinutha and Padma (2023) confirm that BERT achieves significantly better results in semantic keyword grouping compared to earlier models. However, Xu et al. (2021) find that the effectiveness of multilingual transformers for morphologically complex languages remains limited due to insufficient training corpora.

As illustrated in Figure 1, the three core NLP architectures differ in their core innovation and language strength. Word2Vec introduces static word vectors enabling simple semantic analogies. fastText extends this by processing subword n-grams, achieving better performance for morphologically rich languages. BERT, relying on transformer-based bidirectional encoding, provides deep contextual understanding – the most powerful approach for capturing nuanced semantic relationships between keywords.

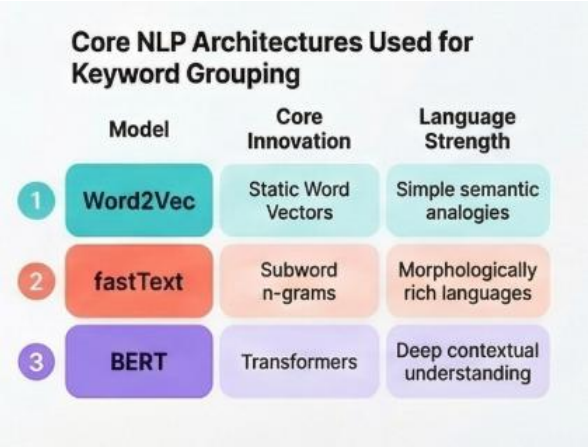


Fig. 1. Core NLP Architectures Used for Keyword Grouping Source
Source: Own elaboration, 2026

3.3. Clustering Algorithms

Word embeddings serve as input data for clustering algorithms through which thematic keyword groups are formed. Vinutha and Padma (2023) apply k-means to Word2Vec-vectorized keywords and obtain thematically coherent clusters outperforming frequency-based methods, while noting the algorithm's limitation for non-linearly distributed data. As a more flexible alternative, DBSCAN groups points based on local density without requiring a predefined number of clusters. Srivastava (2023) indicates that HDBSCAN provides even finer granularity for real SEO data, as it does not assume uniform cluster density. Comparative testing of all three algorithms on real data from diverse market contexts remains underrepresented in the literature.

Figure 2 illustrates the fundamental structural difference between the two approaches. k-means partitions data into a fixed number of spherical groups, which can distort natural thematic boundaries when keyword clusters are irregularly shaped or overlap. HDBSCAN, by contrast, identifies clusters of varying density and shape, isolating noise points as outliers. This distinction has direct practical implications for keyword grouping: real-world semantic spaces rarely conform to the uniform geometry assumed by k-means, making density-based approaches more suitable for capturing natural thematic organisation.

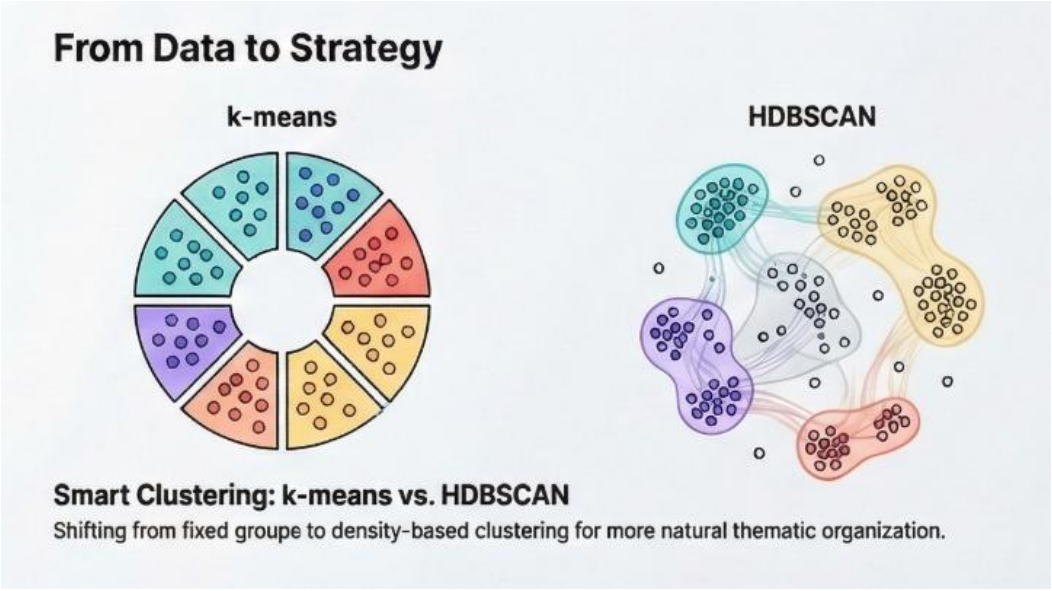


Fig. 2. Smart Clustering: k-means vs. HDBSCAN – Shifting from Fixed Groups to Density-Based
Source: Own elaboration, 2026

3.4. SEO Applications and Identified Research Gaps

Matošević et al. (2021) demonstrate the application of machine learning for web page classification, without, however, focusing on the construction of semantic keyword cores. Kar (2023) and Vyawahare (2022) address the broader impact of artificial intelligence on digital marketing and acknowledge the need for NLP model localization, without offering implementations for specific languages. Xu et al. (2021) explicitly note the limited effectiveness of multilingual models for morphologically rich languages. Research targeting sectors with descriptive terminology – such as services and fast-moving consumer goods – is particularly scarce, representing a concrete and unfilled research niche.

4. Discussion

The literature review reveals that most automated keyword selection methods rely on quantitative indicators such as frequency and search volume, without exploring the semantic interrelations between terms. Vinutha and Padma (2023), Srivastava (2023), and Matošević et al. (2021) highlight the strength of NLP methods in capturing user intent and thematic structure, but their analyses are predominantly based on English-language data and highly formalised domains. The conclusion that emerges is that semantic grouping requires an integrated approach combining morphologically robust word embeddings with adapted clustering algorithms.

Three research gaps are identified through this review. First, the absence of adapted solutions for morphologically rich languages such as Bulgarian. Second, the insufficient methodological development for sectors with descriptive and dynamic terminology. Third, the limited application of expert-based validation of semantic clustering results – a methodological gap explicitly absent from the majority of the reviewed publications. These gaps collectively justify the development of an algorithm integrating word embedding, clustering, and an evaluation function adapted for Bulgarian text and verified against real data from diverse market contexts.

Based on the analysis the following conceptual model for further analysis is proposed (fig. 1).

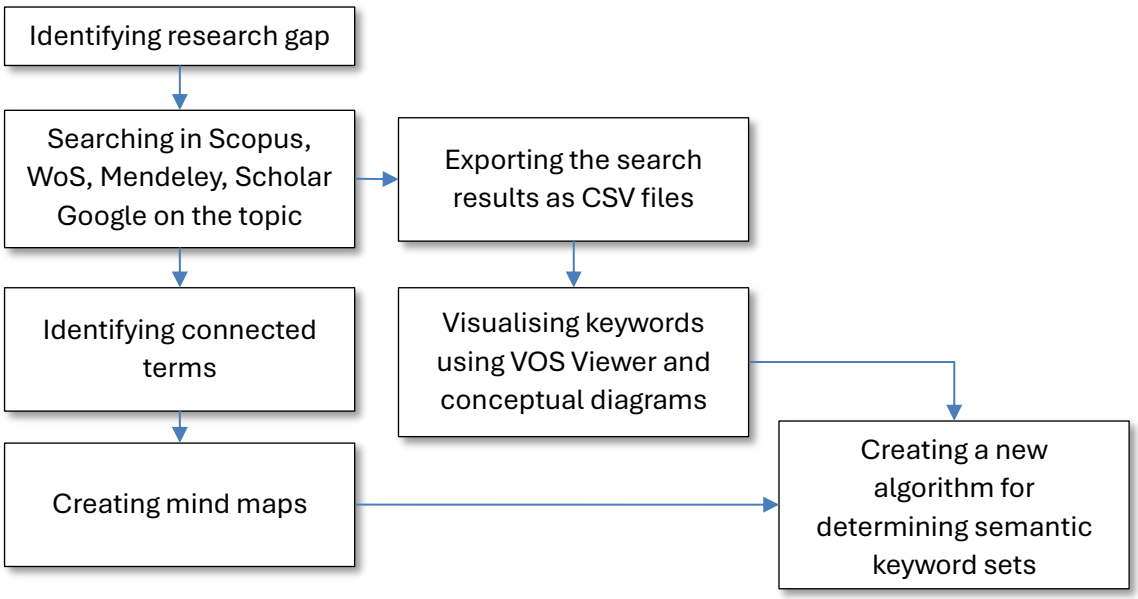


Fig. 1. Conceptual model for further analysis on the topic "algorithms for determining semantic keyword sets"
Source: Own contribution

5. Conclusion

This literature review systematises the development of algorithmic approaches to semantic keyword grouping, from early frequency-based methods to contemporary transformer architectures. Three principal research gaps are identified: the lack of adapted solutions for morphologically complex languages; insufficient methodological development for sectors with descriptive terminology; and limited expert-based verification of clustering outputs. These gaps justify the development of an algorithm integrating word embedding, clustering, and an evaluation function adapted for Bulgarian text and tested on real data from diverse business contexts – a contribution with both scientific and applied value in the field of digital marketing and SEO optimisation.

References

1. Bianchi, T. (2024) *Market share of leading search engines worldwide from January 2015 to January 2024*. Statista. Available at: https://www.statista.com/statistics/1381664/worldwide-all-devices-market-share-of-search-engines/ (Accessed: 5 August 2025).
2. Dean, B. (2024) *Google's 200 Ranking Factors: The Complete List*. Backlinko. Available at: https://backlinko.com/google-ranking-factors (Accessed: 25 August 2025).
3. Guelailia, R. & Bouziane, M. (2024) 'Enhancing Search Engine Optimization through Artificial Intelligence', *Beam Journal of Economic Studies*, 8(2), pp. 497–498.
4. Jindal, A. (2025) 'Generative Engine Optimization (GEO): A New Paradigm for Search in the Age of AI', *International Journal of Computer Applications*, 178(4), pp. 15–27.
5. Kar, S. (2023) 'Impact of Artificial Intelligence on Digital Marketing', *International Journal of Scientific Research in Engineering and Management*, 10(4), pp. 45–52.
6. Matošević, G., Dobša, J. & Mladenić, D. (2021) 'Using Machine Learning for Web Page Classification in Search Engine Optimization', *Future Internet*, 13(1), 9. https://doi.org/10.3390/fi13010009
7. Sancanin, B. & Penjisevic, A. (2022) 'Use of Artificial Intelligence for the Generation of Media Content', *Social Informatics Journal*, 1(1), pp. 1–7.
8. Shawn (2023) *The Impact of AI on Backlink Analysis and Link Building Strategies*. Feature Digital. Available at: https://feature-digital.com/the-impact-of-ai-on-backlink-analysis-and-link-building-strategies/ (Accessed: 25 August 2025).
9. Srivastava, A. (2023) 'AI/ML in search engine optimizer', *International Journal of Innovative Research and Creative Technology (IJIRCT)*, 9(6). ISSN 2454-5988.
10. Vinutha, M.S. & Padma, M.C. (2023) 'Insights into search engine optimization using natural language processing and machine learning', *International Journal of Advanced Computer Science and Applications (IJACSA)*, 14(2), pp. 86–89.
11. Vyawahare, H.R. (2022) 'Machine Learning: A Solution Approach for Complex Problems', *International Journal of Scientific Research in Engineering and Management*, 6(6), pp. 1–7.
12. Ziakis, C., Kyrkoudis, T., Karagkiozidou, M. & Vlachopoulou, M. (2021) 'SEO and Machine Learning Integration for Digital Marketing Strategies', *International Journal of Online Marketing*, 11(1), pp. 65–81.