Integrative Deep Learning and Text Similarity Frameworks for Advanced Keyword Extraction and Semantic Intelligence in Multidomain Text Analytics

[1]Monique L. Duval

[1] Universidad de Buenos Aires, Argentina

Received: 19th Nov 2025 | Received Revised Version: 15th Dec 2025 | Accepted: 30th Jan 2025 | Published: 01th Feb 2026

Volume 08 Issue 02 2025 |

Abstract

The accelerating growth of unstructured textual data across scientific, social, medical, and technological domains has created an unprecedented demand for intelligent systems capable of extracting meaningful, concise, and semantically rich information. Among these tasks, keyword extraction and semantic similarity modeling occupy a central role because they directly enable indexing, retrieval, sentiment analysis, document classification, question answering, and automated knowledge generation. Despite decades of research, the complexity of language, the heterogeneity of domains, and the short and noisy nature of many modern texts continue to challenge both classical and neural approaches. This study develops a comprehensive, integrative research framework that synthesizes classical keyword extraction algorithms, text similarity metrics, and state-of-the-art deep learning architectures to advance the theoretical and empirical understanding of automated semantic intelligence. Drawing strictly upon the referenced literature, this work connects statistical, graph-based, rule-based, and transformer-based approaches into a unified conceptual ecosystem.

The paper begins by positioning keyword extraction as a foundational problem in natural language engineering, grounded in issues of domain specificity, semantic ambiguity, and linguistic variability, as elaborated by Firoozeh et al. (2020) and Miah et al. (2021). Classical unsupervised techniques such as RAKE, SOBEK, and TextRank are explored not merely as algorithms but as epistemic devices that encode assumptions about term relevance, co-occurrence, and discourse structure (Huang et al., 2020; Reategui et al., 2022; Huang & Xie, 2021). These methods are then contrasted with modern deep learning paradigms such as BERT-based architectures and transformer-driven neural taggers, which model language through contextual embeddings and attention mechanisms that capture long-range semantic dependencies (Tang et al., 2019; Martinc et al., 2021).

A central theoretical contribution of this study is the articulation of text similarity as the conceptual bridge between keyword extraction, sentiment analysis, and semantic understanding. By integrating similarity measures such as Jaccard, embedding-based similarity, and semantic alignment, the article demonstrates how relevance, polarity, and conceptual cohesion can be modeled in a single analytical framework (Fernando & Herath, 2021; Mohler et al., 2011; Amur et al., 2023). The methodological section proposes a hybrid pipeline in which neural models generate contextual representations, while symbolic and statistical methods refine and validate extracted keywords against domain knowledge and prior public information (Huang & Xie, 2021; Jain et al., 2022).

The results are described through an extensive comparative analysis of how classical, hybrid, and deep learning approaches perform across domains such as scientific literature, social media, and biomedical texts. The findings show that deep neural architectures excel in capturing semantic nuance, while hybrid methods grounded in similarity metrics and domain knowledge improve precision, interpretability, and robustness to noise (Dang et al., 2020; Imran et al., 2020; Blake & Mangiameli, 2011). The discussion critically examines the implications of these findings for automated question generation, sentiment detection, and knowledge extraction in complex domains such as healthcare and education (Gilal et al., 2022; Alaggio et al., 2022).

By offering an integrated, theory-driven synthesis of keyword extraction, semantic similarity, and deep learning, this article contributes a comprehensive foundation for future research and applied systems. It argues that the future of text analytics lies not in the dominance of any single algorithmic paradigm, but in the strategic orchestration of symbolic, statistical, and neural intelligence into adaptive, explainable, and domain-aware semantic engines.

Keywords: Keyword extraction, semantic similarity, deep learning, text mining, sentiment analysis, transformer models, natural language engineering

Cite This Article: Monique L. Duval. (2026). Integrative Deep Learning and Text Similarity Frameworks for Advanced Keyword Extraction and Semantic Intelligence in Multidomain Text Analytics. The American Journal of Agriculture and Biomedical Engineering, 8(2), 1–9. Retrieved from https://theamericanjournals.com/index.php/tajabe/article/view/7355

Introduction

Aflatoxins The rapid digitalization of human knowledge has fundamentally transformed how information is created, stored, and consumed. Every scientific publication, medical report, patent, social media post, and educational resource contributes to a continuously expanding ocean of textual data. This vast proliferation of unstructured text has made the problem of information overload not merely an inconvenience but a structural barrier to knowledge discovery. Within this context, keyword extraction and semantic analysis have emerged as core technologies for enabling automated systems to understand, organize, and retrieve textual information in ways that approximate human cognition. As Firoozeh et al. (2020) emphasize, keyword extraction is not a peripheral task but a central operation in natural language engineering because keywords act as conceptual anchors that connect documents to queries, topics, and human interpretation.

Historically, keyword extraction was treated as a relatively straightforward problem of term frequency and statistical prominence. However, decades of research have demonstrated that the notion of a “keyword” is deeply semantic, contextual, and domain-dependent. A word or phrase is not a keyword simply because it appears often, but because it represents a concept that is central to the meaning and purpose of a document. Miah et al. (2021), in their study of keyword extraction within the electric double-layer capacitor domain, illustrate how domain specificity dramatically alters what counts as a relevant term. Technical fields employ specialized vocabularies, implicit assumptions, and layered conceptual structures that cannot be captured through surface-level statistics alone. This insight motivates a shift from purely statistical models toward hybrid and neural systems that attempt to model meaning itself.

The emergence of deep learning and transformer-based architectures has further transformed the landscape of text analytics. Models such as BERT and transformer-based neural taggers can represent words not as isolated tokens but as context-dependent embeddings that encode semantic, syntactic, and pragmatic information simultaneously (Tang et al., 2019; Martinc et al., 2021). These models have demonstrated remarkable performance in tasks ranging from sentiment analysis to named entity recognition and keyword identification. Yet, their power introduces new challenges related to interpretability, data dependence, and domain adaptation. As Dang et al. (2020) argue, deep learning models can achieve high accuracy, but their effectiveness depends critically on data quality, annotation consistency, and the complexity of the target task.

At the same time, classical approaches such as RAKE, SOBEK, and TextRank continue to play an important role in practical systems because of their transparency, computational efficiency, and independence from labeled data (Huang et al., 2020; Reategui et al., 2022; Huang & Xie, 2021). These methods operationalize assumptions about how keywords behave in text, such as the tendency of important terms to co-occur or to occupy prominent positions in a discourse network. Although they lack the representational depth of neural models, they offer a form of algorithmic interpretability that is particularly valuable in scientific and legal contexts.

A crucial but often under-theorized dimension of keyword extraction is its relationship to text similarity. Similarity measures provide the mathematical and conceptual glue that links words, sentences, and documents into coherent semantic structures. Fernando and Herath (2021) demonstrate how Jaccard similarity can be used to correlate past and future actions in video data, but the same logic applies to textual units, where overlap and divergence in token sets reflect degrees of conceptual alignment. Mohler et al. (2011) further show that semantic similarity based on dependency graphs and lexical resources can be used to grade short answers, illustrating how meaning can be quantified through relational structures. Amur et al. (2023) extend this insight by reviewing short-text semantic similarity techniques, emphasizing the importance of embedding-based and knowledge-based approaches for handling sparse and ambiguous texts.

The convergence of keyword extraction, sentiment analysis, and semantic similarity is particularly evident in modern applications such as social media analytics and healthcare informatics. Imran et al. (2020) and Jain et al. (2022) show how deep learning models can detect polarity and emotion in tweets by learning complex semantic patterns, while Alaggio et al. (2022) demonstrate the critical importance of precise terminology in the classification of haematolymphoid tumors. In such high-stakes domains, inaccurate or incomplete keyword extraction can have serious consequences for diagnosis, research, and policy.

Despite the richness of existing research, a persistent gap remains in the theoretical integration of these diverse approaches. Many studies evaluate individual algorithms or models in isolation, but few offer a comprehensive framework that explains how classical methods, similarity measures, and deep learning architectures can be combined into a coherent system. This article addresses that gap by synthesizing insights from the provided references into an integrative model of semantic intelligence. Rather than treating keyword extraction, sentiment analysis, and similarity modeling as separate tasks, it conceptualizes them as interdependent components of a single cognitive pipeline that transforms raw text into structured knowledge.

By grounding every claim in the referenced literature and elaborating their theoretical implications, this study aims to provide not only an empirical comparison but a deep conceptual foundation for future research. The central argument is that the most powerful and reliable text analytics systems will be those that strategically combine the interpretability of classical algorithms, the contextual sensitivity of neural models, and the formal rigor of similarity metrics. In doing so, they will move closer to the long-standing goal of artificial systems that can genuinely understand, rather than merely process, human language.

1. Methodology

The methodological foundation of this study is constructed as an integrative analytical framework that synthesizes classical keyword extraction algorithms, text similarity metrics, and deep learning architectures into a unified pipeline for semantic intelligence. Rather than proposing a single new algorithm, the methodology articulates how different methodological traditions, as represented in the referenced literature, can be systematically combined to address the multifaceted challenges of keyword extraction, sentiment analysis, and semantic similarity. This approach is grounded in the recognition that language is a complex, hierarchical, and context-dependent system that cannot be adequately modeled by any single computational paradigm (Firoozeh et al., 2020; Amur et al., 2023).

The starting point of the methodological framework is the representation of textual data. All subsequent operations depend on how text is encoded, segmented, and normalized. Classical approaches such as RAKE and SOBEK assume that text can be decomposed into tokens, phrases, and co-occurrence structures that reflect latent semantic importance (Huang et al., 2020; Reategui et al., 2022). These methods operate primarily on surface forms, using heuristics such as stop-word filtering, phrase delimitation, and frequency weighting to generate candidate keywords. The methodological significance of these steps lies in their ability to reduce linguistic complexity into manageable symbolic units, which can then be manipulated algorithmically.

In contrast, deep learning approaches such as BERT-based models and transformer-based neural taggers represent text in high-dimensional embedding spaces that encode contextual meaning (Tang et al., 2019; Martinc et al., 2021). Each token is mapped to a vector whose position reflects not only its lexical identity but also its relationship to surrounding words. This contextualization is achieved through attention mechanisms that dynamically weight different parts of the input based on their relevance to the prediction task. The methodological implication is that meaning is not pre-defined by rules or frequencies but learned from data as a distributed pattern across millions of parameters.

The integrative methodology proposed in this study treats these two representational paradigms as complementary rather than competing. Classical tokenization and phrase extraction provide a transparent, rule-based scaffold that identifies candidate units of meaning, while deep embeddings provide a rich semantic substrate that captures context and nuance. By aligning these representations through similarity metrics, the system can exploit the strengths of both.

Text similarity serves as the core analytical bridge in this framework. Similarity measures quantify how closely two textual units, such as words, phrases, or documents, are related in meaning. Fernando and Herath (2021) illustrate the power of Jaccard similarity in modeling temporal relationships, while Mohler et al. (2011) demonstrate how semantic similarity based on dependency graphs can align student answers with reference solutions. In the context of keyword extraction, similarity metrics can be used to evaluate how well a candidate keyword represents the central themes of a document by comparing its embedding or token set with that of the full text or with domain-specific reference corpora (Miah et al., 2021; Huang & Xie, 2021).

The methodological pipeline begins with candidate generation using classical or hybrid algorithms such as RAKE, SOBEK, NER-RAKE, and TextRank (Huang et al., 2020; Reategui et al., 2022; Huang & Xie, 2021). These algorithms produce a ranked list of potential keywords based on heuristics, graph centrality, or named entity recognition. This initial list is deliberately inclusive, prioritizing recall over precision in order to avoid prematurely discarding potentially important terms.

Next, deep learning models are employed to compute contextual embeddings for the document and for each candidate keyword. In the case of BERT-based models, these embeddings capture the semantic role of each word or phrase within its specific context (Tang et al., 2019; Jain et al., 2022). Transformer-based taggers such as TNT-KID further refine this process by labeling tokens according to their likelihood of being keywords, effectively integrating supervised or semi-supervised learning into the pipeline (Martinc et al., 2021).

The similarity between candidate keywords and the document representation is then calculated using embedding-based metrics, token overlap measures, or knowledge-based similarity, depending on the domain and data availability (Amur et al., 2023; Mohler et al., 2011). Candidates that exhibit high semantic similarity to the document's core representation are retained and re-ranked, while those with low similarity are filtered out. This step operationalizes the intuition that a true keyword should be a semantic microcosm of the document as a whole.

Domain knowledge and prior public information further enhance this process, particularly in specialized contexts such as patents, scientific literature, and healthcare. Huang and Xie (2021) demonstrate how incorporating prior public knowledge into a TextRank-based model improves patent keyword extraction by aligning extracted terms with established technical concepts. Similarly, Alaggio et al. (2022) highlight the importance of standardized terminology in medical classification systems. In the proposed methodology, domain ontologies, controlled vocabularies, or knowledge bases can be used to validate and refine candidate keywords, ensuring both semantic relevance and terminological accuracy.

Sentiment analysis and polarity detection are integrated into the framework as parallel semantic layers. Deep learning models trained on social media and COVID-19 tweets, as described by Imran et al. (2020) and Dang et al. (2020), demonstrate how embeddings can encode not only topical meaning but also emotional and evaluative dimensions. By linking extracted keywords to sentiment scores or emotion categories, the system can generate richer, more nuanced representations of text that go beyond topical indexing.

The methodological robustness of the framework is grounded in an awareness of data quality and problem complexity. Blake and Mangiameli (2011) show that classification performance is strongly influenced by the interaction between data quality and task difficulty. In keyword extraction and semantic analysis, noisy, short, or domain-mismatched data can degrade the performance of even the most sophisticated models. The hybrid approach mitigates this risk by allowing classical methods to provide stable baselines and by using similarity metrics to detect and correct semantic drift.

In summary, the methodology of this study is not a single algorithm but a layered, modular architecture in which classical extraction, deep representation, similarity measurement, and domain knowledge interact dynamically. This design reflects the theoretical insight that language understanding is a multi-level process, requiring the integration of surface patterns, contextual meaning, and conceptual structure. By articulating this integrative methodology, the study provides a foundation for empirical analysis and practical system design across diverse text-rich domains.

2. Results

The results of the integrative framework described in this study are best understood not as isolated numerical outputs but as a coherent pattern of performance across different types of textual data, algorithmic paradigms, and semantic tasks. Drawing on the empirical and theoretical insights reported in the referenced literature, the results reveal how classical, deep learning, and hybrid approaches differ in their ability to extract meaningful keywords, capture sentiment, and model semantic similarity across domains such as scientific literature, social media, patents, and healthcare texts.

One of the most consistent findings across studies is that deep learning models, particularly those based on transformer architectures, exhibit superior performance in capturing contextual and semantic nuance. Tang et al. (2019) demonstrate that attention-based BERT models can simultaneously classify progress notes and extract keywords with high accuracy, largely because their embeddings encode long-range dependencies and domain-specific usage patterns. Similarly, Martinc et al. (2021) show that the TNT-KID neural tagger can identify keywords by learning token-level representations that reflect both syntactic and semantic importance. These results indicate that deep models excel in environments where meaning is heavily context-dependent, such as medical records or complex scientific texts.

However, the results also reveal important limitations of purely neural approaches. Dang et al. (2020) and Imran et al. (2020) note that deep learning-based sentiment and polarity detection systems are highly sensitive to the quality and representativeness of their training data. When applied to cross-cultural or rapidly evolving domains such as COVID-19 tweets, these models can exhibit bias, overfitting, or reduced generalization. Blake and Mangiameli (2011) further contextualize this problem by demonstrating that data quality and problem complexity interact in ways that can amplify classification errors. In the context of keyword extraction, this means that neural models trained on one domain may misidentify or overlook important terms in another, particularly when domain-specific terminology or rare concepts are involved.

Classical and hybrid keyword extraction methods provide a contrasting pattern of results. Algorithms such as RAKE, SOBEK, and TextRank tend to be more stable across domains because they rely on statistical and graph-based properties of text rather than learned representations (Huang et al., 2020; Reategui et al., 2022; Huang & Xie, 2021). Miah et al. (2021) show that in the highly specialized domain of electric double-layer capacitors, similarity-based keyword extraction techniques can effectively identify relevant technical terms, particularly when augmented with domain knowledge. Reategui et al. (2022) further demonstrate that SOBEK performs well in extracting keywords from diverse datasets, offering a balance between precision and recall that is competitive with more complex models.

Yet, classical methods often struggle with polysemy, synonymy, and contextual variation. A term that is statistically prominent in a document may not be semantically central, and a concept that is crucial to meaning may be expressed through varied linguistic forms that escape frequency-based detection. This is where text similarity measures and deep embeddings play a crucial role. Amur et al. (2023) emphasize that embedding-based similarity metrics are particularly effective for short texts, where traditional co-occurrence statistics are sparse and unreliable. Mohler et al. (2011) similarly show that semantic similarity based on dependency graphs can align student answers with reference solutions even when lexical overlap is low.

The integration of similarity metrics into the keyword extraction pipeline yields a distinctive pattern of results. When candidate keywords generated by classical algorithms are filtered and re-ranked based on their semantic similarity to the document embedding, the resulting keyword sets exhibit higher conceptual coherence and relevance. Huang and Xie (2021) provide empirical evidence for this effect in patent keyword extraction, where the incorporation of prior public knowledge and similarity-based ranking improves both precision and domain alignment. These results suggest that similarity metrics act as a semantic validation layer, ensuring that extracted keywords are not merely frequent but meaningfully representative.

In sentiment analysis and polarity detection, the results show a similar pattern of complementarity. Deep learning models such as BERT-DCNN, as described by Jain et al. (2022), achieve high accuracy by learning complex semantic and affective patterns from large datasets. However, their predictions become more interpretable and robust when linked to keyword-level representations that indicate which terms contribute most strongly to a given sentiment. Imran et al. (2020) demonstrate that cross-cultural emotion detection benefits from this dual representation, as it allows analysts to trace model outputs back to culturally specific expressions and keywords.

In specialized domains such as healthcare, the results underscore the critical importance of terminological precision. Alaggio et al. (2022) show that the classification of haematolymphoid tumors depends on highly specific and standardized terminology. Keyword extraction systems that fail to capture these nuances risk producing misleading or clinically irrelevant outputs. The integration of named entity recognition, as in NER-RAKE (Huang et al., 2020), and domain knowledge, as in TextRank with prior public information (Huang & Xie, 2021), significantly improves the alignment of extracted keywords with established medical vocabularies.

In educational and knowledge generation applications, the results highlight the value of semantic similarity for evaluating and generating content. Gilal et al. (2022) show that automated multiple-choice question generation relies on accurate keyword identification and semantic alignment between questions and answers. Mohler et al. (2011) further demonstrate that grading short answers depends on measuring how closely a student’s response matches the conceptual structure of a reference answer. In both cases, the combination of keyword extraction and similarity modeling enables systems to move beyond surface matching toward genuine semantic assessment.

Taken together, these results reveal a consistent pattern: deep learning models provide unparalleled sensitivity to context and nuance, classical algorithms offer stability and interpretability, and similarity metrics provide the semantic glue that integrates these components into a coherent system. No single approach dominates across all tasks and domains. Instead, the most robust and effective systems are those that orchestrate multiple methods to balance precision, recall, interpretability, and adaptability.

3. Discussion

The results of this integrative analysis invite a deeper theoretical and practical reflection on the nature of keyword extraction, semantic similarity, and deep learning in contemporary text analytics. At a conceptual level, the findings challenge the notion that progress in natural language processing is simply a matter of replacing older algorithms with newer neural architectures. Instead, they support a more nuanced view in which different methodological traditions embody different epistemic assumptions about what language is and how meaning can be computationally represented.

Classical keyword extraction methods such as RAKE, SOBEK, and TextRank are grounded in a view of language as a network of co-occurring symbols whose statistical and structural properties reflect underlying semantic importance (Huang et al., 2020; Reategui et al., 2022; Huang & Xie, 2021). These algorithms implicitly assume that meaning emerges from patterns of usage within a document or corpus. Their strength lies in their transparency and domain independence: they do not require large labeled datasets or complex training procedures, and their outputs can be traced back to observable properties of the text. However, as Firoozeh et al. (2020) point out, these methods struggle with ambiguity, synonymy, and the subtle ways in which context shapes meaning.

Deep learning models, by contrast, embody a radically different epistemology. Models such as BERT and transformer-based taggers represent language as a high-dimensional manifold in which words and phrases are embedded based on their contextual relationships across vast corpora (Tang et al., 2019; Martinc et al., 2021). Meaning is not defined by explicit rules but by the position of a vector in an embedding space. This allows neural models to capture nuanced semantic and syntactic patterns that elude classical algorithms. Yet, this representational power comes at the cost of opacity, data dependence, and potential bias. Dang et al. (2020) and Imran et al. (2020) show that deep models can inadvertently encode cultural, topical, or sampling biases present in their training data, leading to uneven performance across contexts.

The role of text similarity in this landscape is both theoretical and practical. Similarity measures provide a formal way to relate different linguistic units in terms of their semantic proximity. Fernando and Herath (2021) demonstrate this principle in the domain of action anticipation, but the same logic applies to text: words, phrases, and documents can be compared based on overlap, embedding distance, or structural alignment. Mohler et al. (2011) and Amur et al. (2023) show that such measures are essential for tasks that require semantic understanding rather than mere lexical matching.

By integrating similarity metrics into keyword extraction and sentiment analysis pipelines, the framework described in this study effectively reconciles the strengths and weaknesses of classical and neural approaches. Similarity acts as a semantic validation layer that filters out statistically prominent but semantically irrelevant terms and highlights contextually important but infrequent concepts. This has profound implications for domains where precision and interpretability are paramount, such as healthcare, law, and education.

The application of this integrative framework to biomedical texts, as exemplified by Alaggio et al. (2022), illustrates the stakes of semantic accuracy. In the classification of haematolymphoid tumors, the difference between two closely related terms can correspond to radically different diagnoses and treatments. Keyword extraction systems that rely solely on frequency or generic embeddings may fail to capture these distinctions. The incorporation of named entity recognition (Huang et al., 2020) and domain-specific knowledge (Huang & Xie, 2021) ensures that extracted keywords align with established medical ontologies, thereby supporting reliable downstream analysis.

In social media and sentiment analysis, the discussion shifts toward the interpretability and cultural sensitivity of models. Imran et al. (2020) demonstrate that emotions and polarity are expressed differently across cultures, languages, and contexts. Deep learning models can learn these patterns, but without keyword-level explanations and similarity-based validation, their outputs risk being opaque and difficult to audit. By linking sentiment predictions to extracted keywords and their semantic relationships, analysts can better understand why a given text is classified as positive, negative, or emotionally charged.

Educational applications such as automated question generation and grading further highlight the importance of semantic alignment. Gilal et al. (2022) and Mohler et al. (2011) show that these tasks require systems to identify not just topical keywords but the conceptual structure of knowledge. A multiple-choice question or a short answer is meaningful only insofar as it reflects the underlying concepts of a subject. Similarity-based keyword extraction enables systems to map student responses and generated questions onto reference knowledge structures, supporting more accurate and pedagogically sound assessments.

Despite these strengths, the integrative framework also faces limitations that must be acknowledged. The reliance on deep learning models introduces dependencies on large, high-quality datasets that may not be available for all domains or languages. Blake and Mangiameli (2011) remind us that data quality and problem complexity can interact in unpredictable ways, leading to degraded performance even in sophisticated systems. Moreover, the computational cost and energy consumption of training and deploying large transformer models raise practical and ethical concerns.

Future research should therefore focus on developing more efficient, interpretable, and domain-adaptive models that preserve the semantic richness of deep learning while reducing its resource footprint and vulnerability to bias. One promising direction is the increased use of hybrid models that leverage pre-trained embeddings in combination with lightweight, domain-specific classifiers and similarity metrics. Another is the integration of structured knowledge bases and ontologies to ground neural representations in explicit conceptual frameworks.

In theoretical terms, the findings of this study support a view of language understanding as an emergent property of interacting representational systems. Keywords, embeddings, similarity measures, and domain knowledge each capture different aspects of meaning, and their integration yields a more complete and robust model of semantic intelligence. Rather than seeking a single “best” algorithm, researchers and practitioners should focus on designing architectures that orchestrate multiple methods in a principled and context-sensitive manner.

4. Conclusion

This research has presented a comprehensive and theoretically grounded synthesis of keyword extraction, semantic similarity, and deep learning within the broader field of text analytics. Drawing strictly from the provided references, the study has demonstrated that these three pillars of natural language processing are not isolated techniques but deeply interdependent components of a unified semantic intelligence framework. Classical keyword extraction algorithms provide transparent and stable mechanisms for identifying candidate concepts, deep learning models offer rich contextual representations of meaning, and similarity metrics serve as the connective tissue that aligns linguistic units into coherent semantic structures.

The central conclusion of this work is that the future of keyword extraction and text analytics lies in integrative, hybrid systems rather than in the dominance of any single methodological paradigm. Studies such as those by Miah et al. (2021), Reategui et al. (2022), and Huang and Xie (2021) show that classical and similarity-based methods remain highly effective, particularly when augmented with domain knowledge. At the same time, the transformative power of deep learning, as evidenced by Tang et al. (2019), Martinc et al. (2021), and Jain et al. (2022), cannot be ignored, especially in contexts where semantic nuance and contextual understanding are paramount.

By articulating a layered methodological framework and interpreting its results across diverse domains, this article contributes a holistic vision of semantic intelligence that is both theoretically rigorous and practically relevant. It underscores the importance of data quality, domain specificity, and interpretability, as highlighted by Blake and Mangiameli (2011), Imran et al. (2020), and Alaggio et al. (2022). Ultimately, the integration of keyword extraction, similarity modeling, and deep learning offers a pathway toward text analytics systems that are not only more accurate, but also more transparent, adaptable, and aligned with human understanding of language.

References

1. Alaggio, R.; Amador, C.; Anagnostopoulos, I.; Attygalle, A.D.; Araujo, I.B.D.O.; Berti, E.; Bhagat, G.; Borges, A.M.; Boyer, D.; Calaminici, M.; et al. The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Lymphoid Neoplasms. Leukemia 2022, 36, 1720–1748.
2. Amur, Z.H.; Hooi, Y.K.; Bhanbhro, H.; Dahri, K.; Soomro, G.M. Short-Text Semantic Similarity (STSS): Techniques, Challenges and Future Perspectives. Applied Sciences 2023, 13, 3911.
3. Blake, R.; Mangiameli, P. The effects and interactions of data quality and problem complexity on classification. Journal of Data and Information Quality 2011, 2, 1–28.
4. Dang, N.C.; Moreno-García, M.N.; De La Prieta, F. Sentiment analysis based on deep learning: A comparative study. Electronics 2020, 9, 483.
5. Fernando, B.; Herath, S. Anticipating human actions by correlating past with the future with Jaccard similarity measures. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13219–13228.
6. Firoozeh, N.; Nazarenko, A.; Alizon, F.; Daille, B.J. Keyword extraction: Issues and methods. Natural Language Engineering 2020, 26, 259–291.
7. Gilal, A.R.; Waqas, A.; Talpur, B.A.; Abro, R.A.; Jaafar, J.; Amur, Z.H. In Question Guru: An Automated Multiple-Choice Question Generation System. In Proceedings of the 2nd International Conference on Emerging Technologies and Intelligent Systems, ICETIS 2022, Online, 2–3 September 2022; Volume 2, pp. 501–514.
8. Huang, H.; Wang, X.; Wang, H. NER-RAKE: An improved rapid automatic keyword extraction method for scientific literatures based on named entity recognition. Proceedings of the Association for Information Science and Technology 2020, 57, e374.
9. Huang, Z.; Xie, Z. A patent keywords extraction method using TextRank model with prior public knowledge. Complex & Intelligent Systems 2021, 8, 1–12.
10. Imran, A.S.; Daudpota, S.M.; Kastrati, Z.; Bhatra, R. Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related tweets. IEEE Access 2020, 8, 181074–181090.
11. Jain, P.K.; Quamer, W.; Pamula, R.; Saravanan, V. Employing BERT-DCNN with semantic knowledge base for social media sentiment analysis. Journal of Ambient Intelligence and Humanized Computing 2022.
12. Martinc, M.; Škrlj, B.; Pollak, S. TNT-KID: Transformer-based neural tagger for keyword identification. Natural Language Engineering 2021, 28, 409–448.
13. Miah, M.S.U.; Sulaiman, J.; Bin Sarwar, T.; Zamli, K.Z.; Jose, R. Study of keyword extraction techniques for electric double-layer capacitor domain using text similarity indexes: An experimental analysis. Complexity 2021, 2021, 8192320.
14. Mohler, M.; Bunescu, R.; Mihalcea, R. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; pp. 752–762.

15. Reategui, E.; Bigolin, M.; Carniato, M.; dos Santos, R.A. Evaluating the Performance of SOBEK Text Mining Keyword Extraction Algorithm. In Proceedings of the Machine Learning and Knowledge Extraction: 6th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2022, Vienna, Austria, 23–26 August 2022; pp. 233–243.

16. Tang, M.; Gandhi, P.; Kabir, M. Progress notes classification and keyword extraction using attention-based deep learning models with BERT. arXiv 2019, arXiv:1910.05786.