Uluslararası İleri Doğa
Bilimleri ve Mühendislik
Araştırmaları Dergisi
Sayı 9, S. 240-246, 7, 2025
© Telif hakkı IJANSER'e aittir
Araştırma Makalesi

IJANSER
https://as-proceeding.com/index.php/ijanser
ISSN:2980-0811

International Journal of Advanced
Natural Sciences and Engineering
Researches
Volume 9, pp. 240-246, 7, 2025
Copyright © 2025 IJANSER
Research Article

Towards Hybrid Strategies for Classifying Subjects in Scientific Publications: Comparing Traditional, ML and Model-Driven Methods

Ayberk BALTACI[1], Arzum KARATAŞ[2*]

[1] *Computer Engineering Department/Bandırma Onyedi Eylul University, Balıkesir-Türkiye*
[2] *Computer Engineering Department/Bandırma Onyedi Eylul University, Balıkesir-Türkiye*

* *akaratas@bandirma.edu.tr*

(Received: 16 July 2025, Accepted: 21 July 2025)

(5th International Conference on Scientific and Innovative Studies ICSIS 2025, July 15-16, 2025)

ATIF/REFERENCE: Baltacı, A. & Karataş, A. (2025). Towards Hybrid Strategies for Classifying Subjects in Scientific Publications: Comparing Traditional, ML and Model-Driven Methods, *International Journal of Advanced Natural Sciences and Engineering Researches*, 9(7), 240-246.

Abstract – In today's rapidly evolving research landscape, the exponential growth of digital content has rendered the subject classification of academic publications a critical and increasingly complex task. This study systematically examines three predominant methodologies employed in subject classification for scholarly articles: traditional, machine learning-based, and model-driven methods. Traditional methods, grounded in manual curation and expert-driven evaluation, offer nuanced understanding of complex subject matter but face limitations in scalability and objectivity. Machine learning-based methods provide automated processing and efficient handling of large-scale datasets, enabling greater consistency and adaptability. Meanwhile, model-driven methods—particularly those leveraging natural language processing and deep neural architectures—offer enhanced capability in detecting latent patterns within high-dimensional text data.

By integrating these methods, the study proposes a comprehensive and balanced classification framework that synthesizes the interpretive depth of manual systems, the algorithmic efficiency of machine learning, and the advanced analytical potential of deep learning models. This synergy facilitates broader thematic coverage and improved classification accuracy, highlighting how each methodology uniquely contributes to the overall effectiveness of content analysis. Ultimately, the study not only guides researchers in selecting suitable classification strategies for academic literature, but also emphasizes how methodological integration can enhance both the precision and productivity of research workflows. Such hybrid approaches are essential for navigating the increasing complexity of academic knowledge in a digitally-driven era.

Keywords – Review, Scientific Document Analysis, Subject Classification, Traditional Model, Machine Learning-Based Subject Classification, Model-Driven Subject Classification.

I. INTRODUCTION

In today's digitally enriched research landscape, scholars frequently engage in extensive online searches to identify academic articles aligned with their interests. While digital platforms offer access to vast repositories—ranging from scientific journals to e-books—the sheer volume and diversity of available resources pose challenges in locating thematically relevant content. This information overload necessitates the development of effective classification systems capable of organizing scholarly outputs into meaningful categories.

Conventional indexing mechanisms employed by journals and conference proceedings—typically structured by year, volume, and issue—are often inadequate for thematic exploration. Moreover, many publications are not indexed in widely used academic search engines such as Google Scholar, thereby requiring researchers to navigate directly to journal websites and manually browse archival structures, resulting in significant time investment [1].

To address these limitations, automated classification methods have gained prominence, aiming to enhance the accessibility and interpretability of academic content. Some journals request authors to specify article topics at submission, yet restrictive classification schemes—often limited to a single thematic label—fail to reflect the multidimensional nature of scholarly work [2].

Article titles and keywords serve as initial proxies for thematic content, but more sophisticated approaches are required to uncover underlying structures. Traditional topic modeling methods, such as Latent Dirichlet Allocation (LDA), treat each topic as a distribution over words and enable coarse-grained categorization [3]. However, the requirement to predefine topic numbers, reliance on abstract-only content, and the labor-intensive nature of manual validation limit their effectiveness [4]. Studies such as Griffiths et al. [2] and Makagonov et al. [5] show that topic modeling based solely on abstracts yields suboptimal results compared to full-text analysis.

This study presents a comparative analysis of three prevailing classification paradigms: *traditional methods*, *machine learning-based methods*, and *model-driven methods* leveraging advanced natural language processing. By examining their theoretical underpinnings, operational strengths, and domain applicability, we highlight how each approach contributes to the evolving demands of academic knowledge organization. Ultimately, the study offers a comprehensive guide to selecting and integrating suitable classification strategies for contemporary research workflows.

II. CLASSIFICATION METHODS

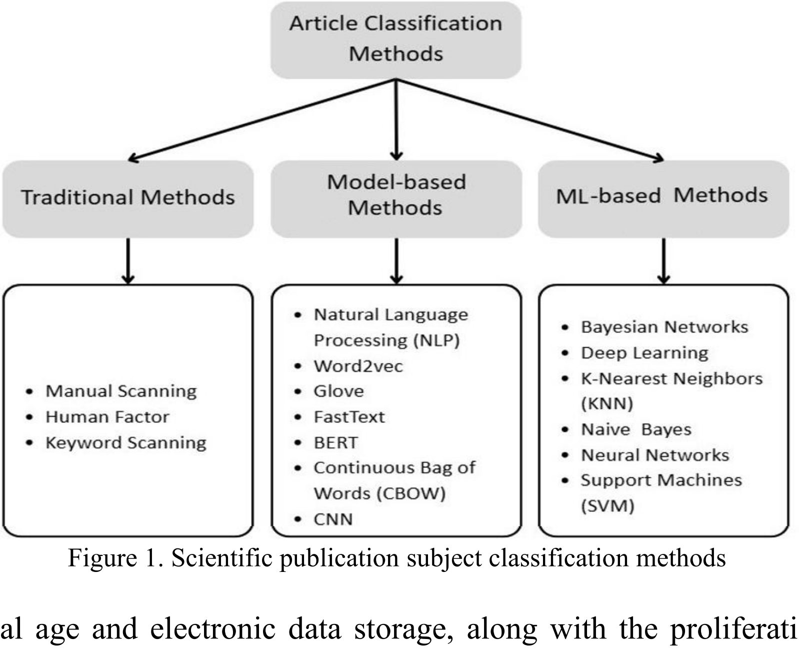

Figure 1. Scientific publication subject classification methods

The rise of the digital age and electronic data storage, along with the proliferation of internet-based services, has paved the way for innovative systems in the field of subject classification in academic writing. This development has also brought new challenges, particularly due to the increasing popularity of interdisciplinary and multidimensional articles. Throughout history, libraries, encyclopedias, and journal publishers have developed specialized classification systems for articles in various fields. Major academic databases such as Web of Science and Scopus have long adopted a journal-centric classification method, meaning that an article's subject is determined based on the journal in which it is published. This approach has been criticized for various accuracy deficiencies [7].

Both existing databases and newly launched databases now offer the possibility of article content-based subject classification using more modern methods such as machine learning. These methods enable target-oriented academic searches through content-based scanning, focusing on article content in addition to journal or publication classification. In the research model of this study, we focused on three important topic classification methods: *traditional methods*, *machine learning-based methods*, and *model-driven methods* (see Figure 1) [7].

A. Traditional Methods

Traditional classification methods, long utilized in academic research and library science, rely on manual evaluation of article content, author intent, and journal scope [8]. Systems such as the Dewey Decimal and Library of Congress Classifications organize broad information resources [9, 10], while academic journals often implement their own thematic categorizations. Although effective in interpreting nuanced content, these methods face limitations in scalability, objectivity, and applicability to interdisciplinary research.

Hybrid approaches combining statistical analysis with illustrative examples are increasingly adopted to support quantitative findings [11]. Baum's study demonstrates the use of varied media sources to examine war coverage. Keyword search techniques offer quick access but depend heavily on dataset familiarity; overly specific or broad queries risk missing relevant items or generating false positives [12]. Notably, platforms like Google intentionally limit comprehensive topic retrieval.

In political science, databases structured by human-defined taxonomies facilitate longitudinal analysis of events such as legislation and media attention [13, 14]. For instance, Adler and Wilkerson [15] leveraged subject-tagged congressional bills to streamline committee jurisdiction research, reducing manual effort substantially. However, systems like the THOMAS Legislative Index Dictionary face challenges such as "subject drift," where evolving classifications disrupt historical consistency [8, 13]. This phenomenon impedes cross-temporal comparisons, as seen in the post-1994 addition of "women's rights" without retroactive tagging. A summary of studies employing traditional classification methods for literature review is presented in Table 1.

The studies summarized in Table 1 underscore the enduring relevance of traditional scientific publication classification methods. Techniques such as content analysis, document review, and systematic literature evaluation facilitate nuanced exploration of thematic structures across diverse academic texts. These approaches offer researchers a broad analytical lens, enabling comprehensive topic delineation and promoting greater classification accuracy. As foundational tools in scholarly categorization, traditional methods continue to enhance both the precision and efficiency of research workflows.

Table 1. Studies that classify subject scientific publications using traditional methods

Study	**Year**	**Methods**
Matematiksel Dil ile İlgili Makalelerin İncelenmesi: Bir İçerik Analizi [16]	2023	Qualitative research methods, document review, criterion sampling, article classification form, descriptive content analysis
Mimarlık ve Edebiyat İlişkisine Dair Yapılmış Akademik Çalışmaların Bir Sınıflandırması [17]	2020	Screening conducted through the National Thesis Center, content analysis, analysis of postgraduate theses in the fields of architecture and literature
Türkiye'de Yapay Zekâ Alanında Yazılmış Yüksek Lisans Tezlerinin İncelenmesi [18]	2023	Qualitative research design, survey model, YÖK National Thesis Center database search, content analysis
Türkçe Eğitimi Alanında Yenilenmiş Bloom Taksonomisini Temel Alarak Yapılan Akademik Çalışmaların İncelenmesi [19]	2022	Qualitative research approach, document review, criterion sampling, content analysis, categorical content analysis
2010-2020 Yılları Arasında Mobil Öğrenme Çalışmalarının İçerik Analiz Yöntemi ile Değerlendirilmesi: Türkiye Örneği [20]	2020	Qualitative analysis, situation assessment study, content analysis, Journal Park platform scan, use of publication classification form
Girişimcilik Alanında Yazılan Akademik Makalelerin Kategorik Olarak Değerlendirilmesi: Girişimciliğin Türkiye'deki Akademik Örüntüsü [21]	2019	Systematic literature review, ULAKBİM journal search, categorical evaluation, data analysis

B. Machine Learning-Based Methods

Machine learning methods for topic classification in academic articles streamline data processing and enable scalable analysis of large and complex corpora, surpassing traditional manual approaches in efficiency and adaptability [22]. The classification workflow comprises key steps: data preprocessing (e.g., cleaning and structuring texts), feature extraction using techniques such as TF-IDF and word embeddings [23], model training via algorithms like Naive Bayes, SVM, Decision Trees, Random Forests, and deep learning architectures (CNN, RNN, LSTM) [24], followed by model optimization, validation (e.g., cross-validation), and final evaluation through accuracy, precision, recall, and F1-score metrics [25].

While ML methods offer powerful automation capabilities, their performance depends on data quality, feature relevance, and parameter tuning. Their "black box" nature may hinder interpretability, suggesting that hybrid approaches integrating traditional insights with ML techniques can yield more balanced classification outcomes.

Several ML algorithms are widely used in academic screening and classification tasks. Bayesian Networks model probabilistic relationships but are challenged by high-dimensional data [26]. Naive Bayes, despite its simplicity and computational efficiency, may falter when feature independence assumptions are violated [27]. Logistic Regression provides flexible, probabilistic modeling for binary outcomes but requires extensive data for stability [28]. Decision Trees and Random Forests support non-linear analysis and ensemble robustness, though they risk overfitting and may be resource-intensive [29]. Support Vector Machines (SVM) excel in high-dimensional contexts with kernel versatility, though training speed varies [30]. K-Nearest Neighbor (KNN) offers intuitive classification but is sensitive to noise and optimal 'k' selection [30]. Finally, Neural Networks simulate complex relationships and offer adaptive learning potential, but may be computationally demanding in large-scale applications [30].

Table 2. Studies that classify subjects scientific publications using machine learning

Study	Year	Methods
Eğitimde Makine Öğrenmesi: Araştırmalardaki Güncel Eğilimler Üzerine İnceleme [31]	2021	Web of Science database search, analysis of articles by algorithms, methods, sample profiles and educational fields, examination of experimental and applied studies.
Doğal Doğal İşleme ile Akademik Metinlerin Kümelenmesi [32]	2022	Text cleaning, tokenization, lemmatization, stemming, TF-IDF and word vector representations, K-Means, K-Medoids, OPTICS and Affinity Propagation clustering methods
Natural language processing for the Turkish Academic texts in the engineering field and development of a decision support system: the case of TUBITAK project proposals [33]	2023	Key term extraction, similarity detection, subject classification, Naïve Bayes classification approach
Kısa Metinleri Yazıldıkları Dile Göre Sınıflandırma ve Farklı Öznitelik Seçim Yöntemlerinin Uygulanması [34]	2021	Language recognition, different classification methods such as Naïve Bayes, Linear Regression, K-Nearest Neighbor, Support Vector Machines, TF-IDF feature selection etc.

C. Model-Driven Methods

Natural language is inherently rich, flexible, and complex, making it challenging for computers to interpret without structural logic and predefined mathematical representations [35]. The field of Natural Language Processing (NLP) addresses this challenge by developing algorithms and systems that enable machines to understand and process human language through rigid linguistic rules, facilitating tasks like spam detection and news classification [36]. Key mechanisms in NLP—such as meaning extraction, named entity recognition, topic discovery, and word embeddings—allow for advanced text comprehension. In particular, topic discovery refers to identifying latent thematic structures in text corpora without relying on predefined dictionaries [35].

To achieve this, textual data must be converted into numerical vectors via word embedding models (e.g., Word2vec, GloVe, fastText, BERT), which enable operations like semantic similarity and vector arithmetic

[37–39]. These models leverage neural networks or context-based matrices to produce context-aware representations, improving downstream tasks such as document retrieval and summary generation.

In modern classification systems, deep learning architectures—notably Convolutional Neural Networks (CNNs)—enhance performance in text analysis by capturing hierarchical features and reducing input dimensionality through pooling layers [40]. Such systems process vast document collections efficiently and support autonomous decision-making in domains ranging from medical diagnostics to social media analytics. While topic classification organizes documents rapidly, topic discovery emulates human cognition by extracting overarching themes from large-scale unstructured data.

III. COMPARISION OF CLASSIFICATION METHODS

Traditional, machine learning-based, and model-driven methods employed in the subject classification of academic articles each offer distinct advantages and limitations. A comparative evaluation of these methodologies enables researchers and domain experts to select the most appropriate strategy tailored to specific research objectives and data characteristics.

Traditional classification techniques have long been utilized in academia and library science, typically relying on manual assessment of article content, keywords, and author expertise. These methods benefit from human evaluators' capacity to interpret contextual cues, tone, and underlying intent. However, they are time-consuming, prone to subjective bias, and lack scalability—particularly when handling large or complex datasets. Moreover, the proliferation of interdisciplinary and multidisciplinary studies increasingly challenges the rigid boundaries of conventional categorization frameworks.

Machine learning approaches, by contrast, facilitate the automated and efficient processing of extensive document collections. This classification pipeline involves stages such as data preparation, feature extraction, model selection and training, followed by optimization and validation. ML methods outperform traditional techniques in speed and adaptability, though they are sensitive to data quality, feature selection, and parameter tuning. Additionally, the opaque, "black box" nature of many algorithms may hinder interpretability, especially in critical decision-making contexts.

Model-driven techniques, leveraging natural language processing (NLP) and deep learning architectures, provide advanced capabilities for uncovering latent semantic structures and key thematic patterns within textual data. NLP tools—such as topic discovery and named entity recognition—capitalize on linguistic regularities to enhance computational analysis. Despite their efficacy, these methods often involve technical complexity and implementation challenges, as well as potential risks associated with emerging technologies and evolving algorithmic paradigms.

Table 3. Advantages and limitations of scientific publication classification methods

Method	**Advantages**	**Limitations**
Traditional Methods	Human understanding and in-depth analysis. Nuanced assessment of complex issues.	Time consuming and subjective. Cannot scale with large data sets.
Machine Learning-Based Methods	Fast processing of large data sets Automatic and efficient analysis.	Difficulty in interpreting decisions. Dependence on data and model parameters.
Model-Driven Methods	Advanced classification and analysis. Ability to extract main ideas from large text data.	Technical complexity and information requirements. Risks of emerging technologies.

IV. CONCLUSION

This study presents a comprehensive examination of traditional, machine learning, and model-based approaches employed for subject classification in academic articles. Each methodology is critically analyzed in terms of its strengths, limitations, and contribution to academic research, as well as its responsiveness to the evolving demands of contemporary scientific and technological contexts.

Traditional methods, grounded in expert judgment and manual assessment, have long served academic and bibliographic communities by enabling nuanced interpretation of scholarly content. In contrast, machine learning approaches automate data processing and analysis, offering substantial gains in efficiency and scalability—particularly for large datasets. Model-based techniques, leveraging natural language processing (NLP) and deep learning frameworks, provide enhanced capabilities for extracting and interpreting complex semantic patterns embedded in textual data.

The findings of this study suggest that an integrated application of these methodologies yields a more balanced and effective strategy for topic classification. The interpretive depth of traditional methods, the operational efficiency of machine learning algorithms, and the advanced analytical precision of model-based approaches together offer researchers a broader lens and more accurate classification outcomes across diverse research domains.

REFERENCES

[1] Hanyurwimfura, D., et al., Topics and Search Based Classification of Scientific Publications. Journal of Computational and Theoretical Nanoscience, 2015. 12(12): p. 5210-5222.

[2] Griffiths, T.L. and M. Steyvers, Finding scientific topics. Proceedings of the National academy of Sciences, 2004. 101(suppl_1): p. 5228-5235.

[3] Anupriya, P. and S. Karpagavalli. LDA based topic modeling of journal abstracts. in 2015 International Conference on Advanced Computing and Communication Systems. 2015. IEEE.

[4] Sing, D.C., L.N. Metz, and S. Dudli, Machine learning-based classification of 38 years of spine-related literature into 100 research topics. Spine, 2017. 42(11): p. 863-870.

[5] Makagonov, P., Alexandrov, M. & Gelbukh, A. (2004). Clustering abstracts instead of full texts. Metin, Konuşma ve Diyalog, LNCS 3206, 129–135, Springer.

[6] Singh, P., et al., Revisiting subject classification in academic databases: A comparison of the classification accuracy of web of science, scopus & dimensions. Journal of Intelligent & Fuzzy Systems, 2020. 39(2): p. 2471-2476.

[7] Bornmann, L., Field classification of publications in Dimensions: A first case study testing its reliability and validity. Scientometrics, 2018. 117: p. 637-640.

[8] Hillard, D., S. Purpura, and J. Wilkerson, Computer-assisted topic classification for mixed-methods social science research. Journal of Information Technology & Politics, 2008. 4(4): p. 31-46.

[9] Scott, M.L. and M.L. SCOTT, Dewey decimal classification. Libraries Unlimited, 1998.

[10] Jenkins, C., et al., Automatic classification of Web resources using Java and Dewey decimal classification. Computer Networks and ISDN Systems, 1998. 30(1-7): p. 646-648.

[11] Baum, M.A., Soft news goes to war: Public opinion and American foreign policy in the new media age. 2011: Princeton University Press.

[12] Carneiro, P., et al., "Identify-to-reject": A specific strategy to avoid false memories in the DRM paradigm. Memory & cognition, 2012. 40(2): p. 252-265.

[13] Baumgartner, F.R., B.D. Jones, and J.D. Wilkerson, Studying policy dynamics. Policy dynamics, 2002: p. 29-46.

[14] Segal, J.A. and H.J. Spaeth, The Supreme Court and the attitudinal model revisited. 2002: Cambridge University Press.

[15] Adler, E.S. and J.D. Wilkerson, Intended consequences: Jurisdictional reform and issue control in the US House of Representatives. Legislative Studies Quarterly, 2008. 33(1): p. 85-112.

[16] Kabakçı-Alyeşil, D., Yitmez, B. G., & Faydaoğlu, Ş. (2023). Matematiksel dil ile ilgili makalelerin incelenmesi: Bir içerik analizi. *Muş Alparslan Üniversitesi Eğitim Fakültesi Dergisi, 3*(1), 1–24.

[17] Serin Güner, A. P., & Gökmen, H. (2020). Mimarlık ve edebiyat ilişkisine dair yapılmış akademik çalışmaların bir sınıflandırması. *İDEALKENT, 11*(31), 1722–1763.

[18] Alkan, A., & Sevli, O. (2023). Türkiye'de yapay zekâ alanında yazılmış yüksek lisans tezlerinin incelenmesi. *Osmaniye Korkut Ata Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 6*(1), 931–947.

[19] Güler, M., & Mert, O. (2022). Türkçe eğitimi alanında yenilenmiş Bloom taksonomisini temel alarak yapılan akademik çalışmaların incelenmesi. *Bayburt Eğitim Fakültesi Dergisi, 17*(35), 1089–1118.

[20] Altunçekiç, A. (2020). 2010-2020 yılları arasında mobil öğrenme çalışmalarının içerik analiz yöntemi ile değerlendirilmesi: Türkiye örneği. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi, 40*(3), 1087–1104.

[21] Gözüm, A. G. (2019). Girişimcilik alanında yazılan akademik makalelerin kategorik olarak değerlendirilmesi: Girişimciliğin Türkiye'deki akademik örüntüsü. *Ufuk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 8*(15), 367–385.

[22] Kaynar, O., et al. Makine öğrenmesi yöntemleri ile Duygu Analizi. in International Artificial Intelligence and Data Processing Symposium (IDAP'16). 2016.

[23] Yuan, H., et al., A detection method for android application security based on TF-IDF and machine learning. Plos one, 2020. 15(9): p. e0238694.

[24] Raschka, S., Model evaluation, model selection, and algorithm selection in machine learning. arXiv preprint arXiv:1811.12808, 2018.

[25] Novaković, J.D., et al., Evaluation of classification models in machine learning. Theory and Applications of Mathematics & Computer Science, 2017. 7(1): p. 39.

[26] Grossman, D. and P. Domingos. Learning Bayesian network classifiers by maximizing conditional likelihood. in Proceedings of the twenty-first international conference on Machine learning. 2004.

[27] Berrar, D., Bayes' theorem and naive Bayes classifier. Encyclopedia of bioinformatics and computational biology: ABC of bioinformatics, 2018. 403: p. 412.

[28] Rymarczyk, T., et al., Logistic regression for machine learning in process tomography. Sensors, 2019. 19(15): p. 3400.

[29] Charbuty, B. and A. Abdulazeez, Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends, 2021. 2(01): p. 20-28.

[30] Mahesh, B., Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 2020. 9(1): p. 381-386.

[31] Tosunoğlu, E., Yılmaz, R., Özeren, E., & Sağlam, Z. (2021). Eğitimde makine öğrenmesi: Araştırmalardaki güncel eğilimler üzerine inceleme. Ahmet Keleşoğlu Eğitim Fakültesi Dergisi, 3(2), 178–199.

[32] Taşkıran, F., & Kaya, E. (2022). Doğal dil işleme ile akademik metin kümeleme. Konya Mühendislik Bilimleri Dergisi, 10(2022), 41–51.

[33] Kat, B. (2023). Natural language processing for the Turkish academic texts in the engineering field and development of a decision support system: The case of TÜBİTAK project proposals. Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi, 38(3), 1879-1892. https://doi.org/10.17341/gazimmfd.1132053

[34] Kat, B. (2023). Natural language processing for the Turkish academic texts in the engineering field and development of a decision support system: The case of TÜBİTAK project proposals. Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi, 38(3), 1879-1892. https://doi.org/10.17341/gazimmfd.1132053

[35] Lezama-Sánchez, A.L., M. Tovar Vidal, and J.A. Reyes-Ortiz, An Approach Based on Semantic Relationship Embeddings for Text Classification. Mathematics, 2022. 10(21): p. 4161.

[36] Ramos, F. and J. Vélez, Integración de técnicas de procesamiento de lenguaje natural a través de servicios web. Universidad Nacional del Centro de la provincia de Buenos Aires, 2016.

[37] Lezama-Sánchez, A.L., M. Tovar Vidal, and J.A. Reyes-Ortiz, Integrating Text Classification into Topic Discovery Using Semantic Embedding Models. Applied Sciences, 2023. 13(17): p. 9857.

[38] Athiwaratkun, B., A.G. Wilson, and A. Anandkumar, Probabilistic fasttext for multi-sense word embeddings. arXiv preprint arXiv:1806.02901, 2018.

[39] Devlin, J., et al., Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[40] Mete, B.R. and T. Ensari. Flower classification with deep CNN and machine learning algorithms. in 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). 2019. IEEE.