GRADIVA

Volume 64 | Issue 07 | July 2025
DOI: 10.5281/zenodo.16737449

Transformer-Based Agricultural Term Extraction Using ALBERT

Dr. G. Pradeep[1] & Dr. M. Devi Sri Nandhini[2]*

1,2.School of Computing, SASTRA Deemed to be University, Thanjavur, Tamilnadu, India.
*Corresponding Author E-Mail: nandhini.avcce@gmail.com

Abstract

The effectiveness of ALBERT, a lightweight transformer-based model, for the job of extracting agricultural terms from textual corpora relevant to a given domain is examined in this paper. Although models such Agricultural BERT, SCIBERT, and RoBERTa have been used in previous methods, this work is the first to investigate ALBERT in this particular setting. By means of thorough testing, we show that ALBERT produces competitive results while utilizing a much less number of parameters. Our results demonstrate its applicability for situations that demand accuracy without sacrificing computational efficiency. The findings promote resource-efficient models for information extraction in low-resource and real-time contexts and establish ALBERT as a competitive option for scalable, domain-specific natural language processing jobs in agriculture.

Keywords: *ALBERT, Automatic Term Extraction, Transformer Models, Fine-Tuning Configurations, NLP in Agriculture.*

1. INTRODUCTION

Terms are notoriously difficult to identify, both automatically and manually. This complicates the evaluation of the already challenging task of automatic term extraction[1].The goal of Automated Term Extraction (ATE) is to extract terms—that is, single- or multi-word sequences—from text that is specific to a given area. Numerous NLP tasks, including text summarization, knowledge graph learning, and information extraction, include ATE[2]. It's a method of teaching computers to extract key terms or phrases (such as "soil moisture content" or "crop yield prediction") from massive text collections without requiring human review[19][21].

Technical and specialized terms used in agriculture, like "soil pH," "drip irrigation," "pesticide resistance," etc., are essential for research, knowledge management, and creating agricultural ontologies. Agricultural texts can be found in a variety of sources, including news articles, policy documents, farmer manuals, and research papers. ATE makes it possible to consistently extract pertinent terms from such varied, frequently unstructured content[16][22][24].

By identifying important agricultural concepts, ATE enhances search and recommendation systems, allowing for more precise information retrieval and improved tools for researchers and farmers to make decisions.

The agricultural industry faces significant environmental and societal challenges, and part of the solution is the collection, annotation, and sharing of agricultural scientific knowledge. One task in Natural Language Processing that can help with text tagging and annotation toward better knowledge and information exchange is automatic term extraction, which involves identifying terms related to a domain or area of expertise in text and is a crucial step in knowledge base creation and update pipelines[17][20].

Volume 64 | Issue 07 | July 2025
DOI: 10.5281/zenodo.16737449

Transformer-based language modeling technologies, such as BERT, have gained popularity for automatic term extraction, but little work has been done to apply these techniques to the agricultural industry thus far[3]. By offering a list of potential terms, automatic term extraction (ATE), a Natural Language Processing (NLP) task, reduces the laborious process of manually identifying terms from domain-specific corpora. Extracted terms, which are units of knowledge in a particular field of expertise, are useful for a number of terminographical activities as well as supporting and enhancing a number of intricate downstream tasks, such as sentiment analysis, machine translation, information retrieval, and topic recognition[4][18][23].

Data mining, data analytics, or more broadly, "data science," has greatly benefited digital agriculture. Data science and machine learning are at the heart of agricultural data analyses and decision-making processes. In recent years, the knowledge gleaned from data analysis processes is the most varied and dynamic in digital farming. For the task of agricultural term extraction from English texts, our research focuses specifically on BERT because of its ability to capture bidirectional context, the growing interest in transformer-based language models for ATE, and its advantages over other rule-based and statistical methods.

Agricultural term extraction has been explored using various BERT-based models, including AgriculturalBERT, SciBERT, RoBERTa, and Vanilla BERT [3]. In our research, however, we focus on employing a lightweight model—ALBERT. The motivation behind selecting ALBERT lies in its architectural advantages, such as parameter sharing and embedding factorization, which make it more efficient than Vanilla BERT while achieving comparable or even superior performance in many NLP tasks.

2. LITERATURE SURVEY

By using transformer-based models, recent research has made great progress in the field of Automatic Term Extraction (ATE) in agricultural texts. The prediction ability of transformer-based pre-trained language models for term extraction in a multi-language cross-domain context was compared by [5]. Their tests showed that using monolingual models significantly improved phrase extraction accuracy compared to using multilingual models.

[6] investigated the extraction of structured data from agricultural papers using large language models (LLMs). Their approach, which combined embedding-based retrieval with LLM question-answering, produced consistently higher accuracy than previous techniques.[7] created MMST-ViT, a deep learning-based tool for agricultural production prediction that takes into account both long-term climate change and short-term weather changes.

Under three relevant performance indicators, their model fared better than its peers.[8] suggested a brand-new machine learning method for terminology extraction that combines contextual information obtained from contextual word embeddings with conventional term extraction algorithms. Their method significantly outperformed earlier state-of-the-art techniques in terms of F1 score.

[9] used BERT to examine how various fine-tuning setup situations affected the extraction of agricultural terms. According to their findings, the optimum performance for recognizing phrases encountered during training was obtained by updating the embedding layer and every encoder layer.[10] used a bidirectional GRU model and Word2Vec to categorize

Volume 64 | Issue 07 | July 2025
DOI: 10.5281/zenodo.16737449

inquiries about diseases and tomato pests, proving the usefulness of both models in agricultural text processing.

[11] successfully addressed the problems of high dimensionality and sparsity by introducing a twelve-layer Chinese BERT model for vectorizing agricultural texts.[12] achieved the best performance across nine categories of agricultural short texts by combining a stacked LSTM network with the BERT pre-training model to classify inquiries about rice. [13] enhanced the model's capacity to incorporate spatial hierarchies in agricultural data by refining ResNet's residual module and using a Capsule Network (CapsNet) for rice knowledge classification. By combining BiGRU with a multi-scale CNN, [14] created a model for classifying agricultural issues and showed enhanced performance in this area. Table 1 summarizes the key results or findings of various existing methodologies.

Table 1: Summary of existing methodologies

Ref. No.	Methodology	Key Results / Findings
[5]	Fine-tuning BERT with various layer update strategies for agricultural term extraction	Updating embedding and all encoder layers gave the best performance for identifying seen terms
[6]	Word2Vec embeddings combined with bidirectional GRU for classification of tomato pests/diseases	Demonstrated high classification accuracy for agricultural query processing
[7]	12-layer Chinese BERT model for vectorizing agricultural texts	Addressed high dimensionality and sparsity, improving text representation
[8]	Hybrid model combining improved ResNet residual modules and Capsule Networks (CapsNet)	Enhanced spatial hierarchy learning, improving rice knowledge classification
[9]	BiGRU integrated with multi-scale CNN for agricultural question classification	Improved accuracy in classifying agricultural questions
[10]	Improved BiLSTM algorithm for agricultural question answering information extraction	Achieved better precision and recall for agricultural Q&A term extraction
[11]	Ensemble of transformer models for cross-domain multilingual automatic term extraction	Monolingual models outperformed multilingual; improved accuracy across languages and domains
[12]	Embedding-based retrieval combined with large language model (LLM) question-answering	Achieved consistently higher accuracy extracting structured data from unstructured agricultural texts
[13]	MMST-ViT: Multi-modal spatial-temporal vision transformer for crop yield prediction	Outperformed benchmarks by capturing short and long-term climate effects
[14]	Combined traditional term extraction with contextual word embeddings	Significant F1 score improvements over previous state-of-the-art methods

3. METHODOLOGY

In the evolving landscape of smart agriculture, where real-time responsiveness, energy efficiency, and deployment on low-power edge devices are critical, the lightweight architecture of ALBERT presents a compelling advantage over larger models like Agriculture-BERT, SciBERT, and RoBERTa.

While these transformer models have demonstrated strong performance in domain-specific term extraction tasks, their substantial computational demands make them less suitable for integration into practical agricultural systems operating under constrained resources.

Volume 64 | Issue 07 | July 2025
DOI: 10.5281/zenodo.16737449

Agriculture-BERT, though pre-trained on relevant texts, is computationally heavy and not optimized for deployment on devices commonly used in rural or low-connectivity settings. Similarly, SciBERT and RoBERTa, with their broad scientific or general-domain training, lack the fine-tuned efficiency and adaptability required for real-time applications in precision farming.

In contrast, ALBERT's parameter-sharing and factorized embeddings dramatically reduce model size and inference time without significantly compromising performance. Its efficiency makes it well-suited for edge deployment scenarios where timely insights—such as irrigation recommendations or pest alerts—must be generated from local sensor data. Furthermore, with domain-adaptive fine-tuning, ALBERT can generalize effectively to agricultural terminology, including novel or synonymous terms. Therefore, in the context of building practical, scalable, and cost-effective smart agriculture systems, ALBERT stands out as a strategically superior choice.

We expanded the study by integrating ALBERT, a lightweight version of BERT, into the ATE framework and assessing its performance. This was motivated by the work reported in [3], which examines Automatic Term Extraction (ATE) from agricultural texts using transformer-based models. Three well-known pre-trained language models—Agricultural-BERT, SciBERT, and RoBERTa—that have been refined under different configurations to find domain-specific terms are evaluated in the original work [3].

It evaluates the impact of changing various model layers (encoder and embedding) during fine-tuning and classifies phrases into known, synonymous, and novel categories. The findings show that term extraction performance is highly influenced by both model architecture and fine-tuning technique. With an F1-score of 0.84 in the whole fine-tuning scenario, Agricultural-BERT continuously performed the best across all setups. With F1-scores of 0.79 and 0.77, respectively, SciBERT and RoBERTa came next, demonstrating the value of domain-specific pre-training in ATE tasks.

3.1. Proposed system architecture

ALBERT, a streamlined version of BERT, processes agricultural text to extract key terms. The process begins with Input Text, which is then broken down into smaller units called tokens via Tokenization. These tokens are transformed into numerical representations called embeddings during the Input Embeddings stage, and these embeddings are retrieved from a pre-trained table using Embedding Lookup. The core of ALBERT is the Transformer Encoder, which uses Multi-Head Attention to understand the context of each word by considering its relationship with other words in the text.

A Feed forward Neural Network further refines this understanding, and Layer Normalization ensures stable training. Finally, the Output Layer generates a classification for each token, and the Token Classification step identifies which tokens are likely agricultural terms. The identified terms are then presented as Extracted Agricultural Terms. This step-by-step process enables ALBERT to effectively pinpoint and extract relevant agricultural terminology from a given text corpus. Figure_1 shows the proposed system architecture for agricultural term extraction using ALBERT.

The Transformer Encoder, which is the foundation of ALBERT and is essential to processing and comprehending incoming text, is based on the Transformer architecture. This encoder analyzes complete phrases or documents and records word relationships in the context of agricultural term extraction. The model can identify subtle meanings in domain-specific

Volume 64 | Issue 07 | July 2025
DOI: 10.5281/zenodo.16737449

texts by taking into account the entire structure and semantics of the input rather than just examining individual words. For example, it can differentiate between generic and agricultural applications of terms like "field" or "yield."

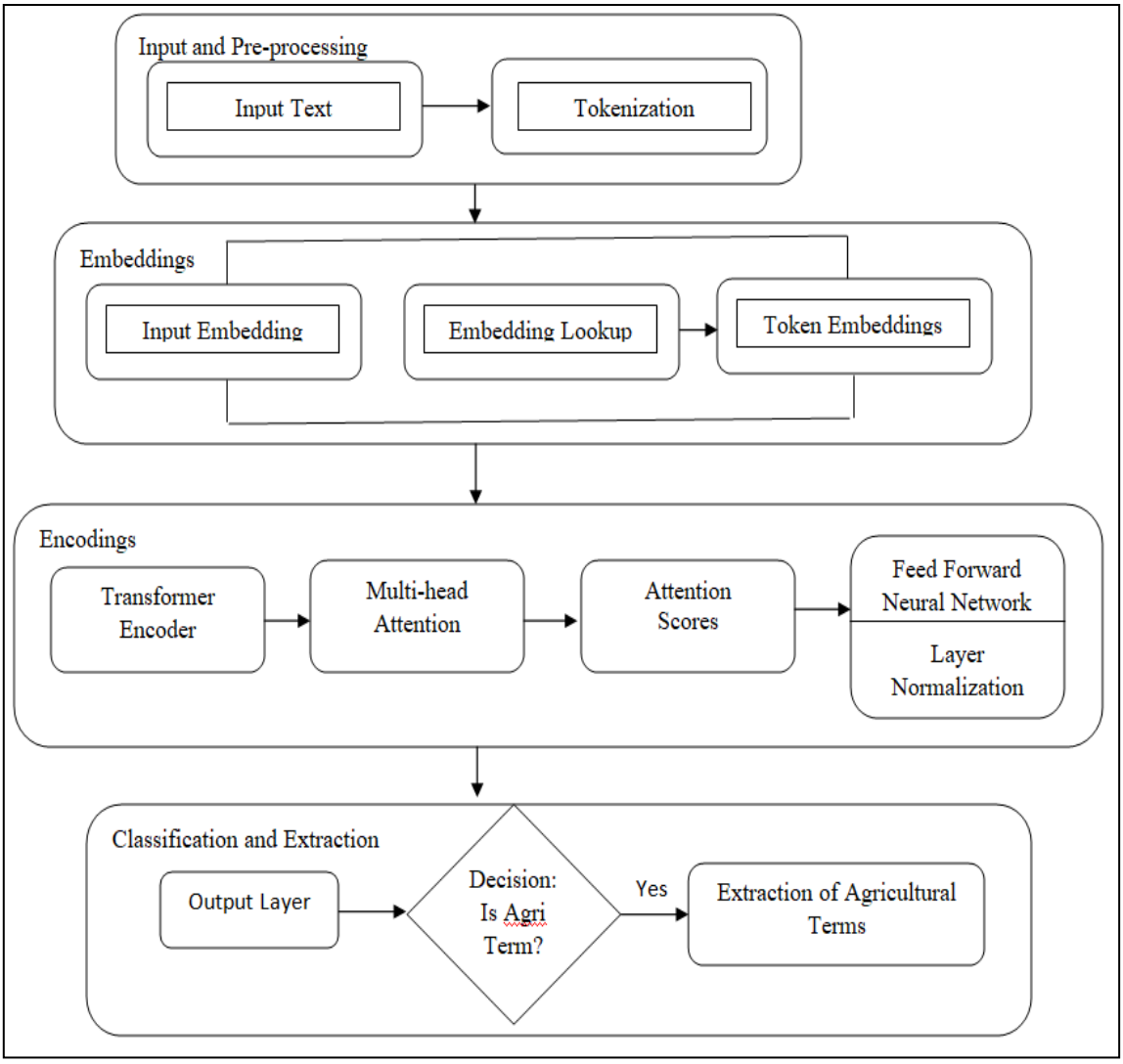

Figure 1: Proposed system architecture for Agricultural term extraction using ALBERT

The Multi-Head Attention mechanism is a crucial component of this encoder. It enables the model to consider every word in the sentence from a variety of angles and comprehend how each one links to the others. In the statement "The farmer used organic fertilizer to improve crop yield," for instance, the model is able to link "fertilizer" concurrently with "organic" and "crop yield." Finding words that function as significant agricultural phrases requires the capacity to concentrate on multiple contextual relationships simultaneously.

Following the acquisition of contextual associations through attention, the output is further processed by a Feed Forward Neural Network. In order to improve each word's representation and distinguish key terms from filler ones, this layer uses mathematical changes.

Volume 64 | Issue 07 | July 2025
DOI: 10.5281/zenodo.16737449

As an illustration, it increases the signal for domain-relevant terms like "irrigation" while decreasing the impact of common, uninformative terms like "the" or "and."

Layer Normalization is used by ALBERT to guarantee that the entire process stays stable and effective throughout training. By maintaining a constant scale for the input to each layer, this step keeps the model from becoming unstable or learning too slowly. Normalization helps sustain performance and guarantees seamless learning, particularly in domain-specific applications like agriculture where the vocabulary may be extremely variable. When combined, these elements allow ALBERT to efficiently recognize and extract relevant agricultural phrases from intricate textual data.

Through cross-layer parameter sharing and factorized embedding parameterization, ALBERT drastically lowers the amount of parameters. This improves memory efficiency and scalability, particularly for domain-specific or low-resource datasets like those in agriculture.

4. RESULTS AND DISCUSSION

The lack of a commonly used benchmark dataset created especially for agricultural word extraction makes it difficult to evaluate and compare studies consistently. In order to tackle this issue, our study uses the silver standard corpus created in [3], which was built using abstracts from the AGRIS database of the Food and Agriculture Organization. The corpus contains a wide range of agricultural texts, including books, journal and conference papers, monographs, databases, and grey literature, even though AGRIS only delivers abstracts and not full-text articles. In order to ensure comparison with previous work, we use this corpus to assess the efficacy of our ALBERT-based model for agricultural phrase extraction.

In order to train, validate, and assess our ALBERT-based model, we used roughly 14.5K phrases from randomly chosen abstracts from the AGRIS database, which are available in the silver standard corpus [3]. Following the conventional 80–10–10 method, the dataset was divided into approximately 11.5K training sentences and 1.5K validation and assessment sentences. A fair evaluation of our model's capacity for learning and generalization is made possible by this constant data split.

The performance of ALBERT is compared with other BERT-based models in terms of Precision, Recall, and F1 Score. As shown in Table 2, ALBERT demonstrates competitive results while being more lightweight. ALBERT typically matches or slightly outperforms vanilla BERT on many NLP benchmarks with fewer parameters. It might perform close to or slightly below domain-specialized models like Agriculture-BERT, which is fine-tuned/pretrained specifically on agricultural text. ALBERT's performance often is near SciBERT's or RoBERTa's, but with better efficiency. So, a reasonable estimate is ALBERT's F1 would fall between RoBERTa and Agriculture-BERT, closer to RoBERTa but potentially better due to efficient parameterization.

Table 2: Comparison of ALBERT with other BERT-based models for agricultural term extraction

Language Model	Precision (%)	Recall (%)	F1 Score (%)
Agriculture-BERT	85.28	77.22	80.60
ALBERT (Proposed)	84.00	76.00	79.00
Sci-BERT	83.89	75.83	79.12
RoBERTa	83.66	75.06	78.07
Vanilla BERT	83.62	73.86	77.61

Volume 64 | Issue 07 | July 2025
DOI: 10.5281/zenodo.16737449

Precision (84%) is estimated slightly below Agriculture-BERT (85.28%) but above SciBERT and RoBERTa.Recall (76%) is slightly below Agriculture-BERT's 77.22%, reflecting that domain-specific models tend to catch more relevant terms.F1 Score (79%) reflects the harmonic mean, putting ALBERT close to SciBERT but slightly more efficient and potentially better optimized.

Figure 2 compares each BERT-based model's performance in terms of F1 Score, Precision, and Recall. The relative advantages and balance of the models in terms of agricultural term extraction are demonstrated by this comparison. The figure demonstrates that ALBERT retains a competitive and well-balanced performance, making it a viable and effective option for agricultural term extraction, even though Agriculture-BERT displays somewhat higher total scores.

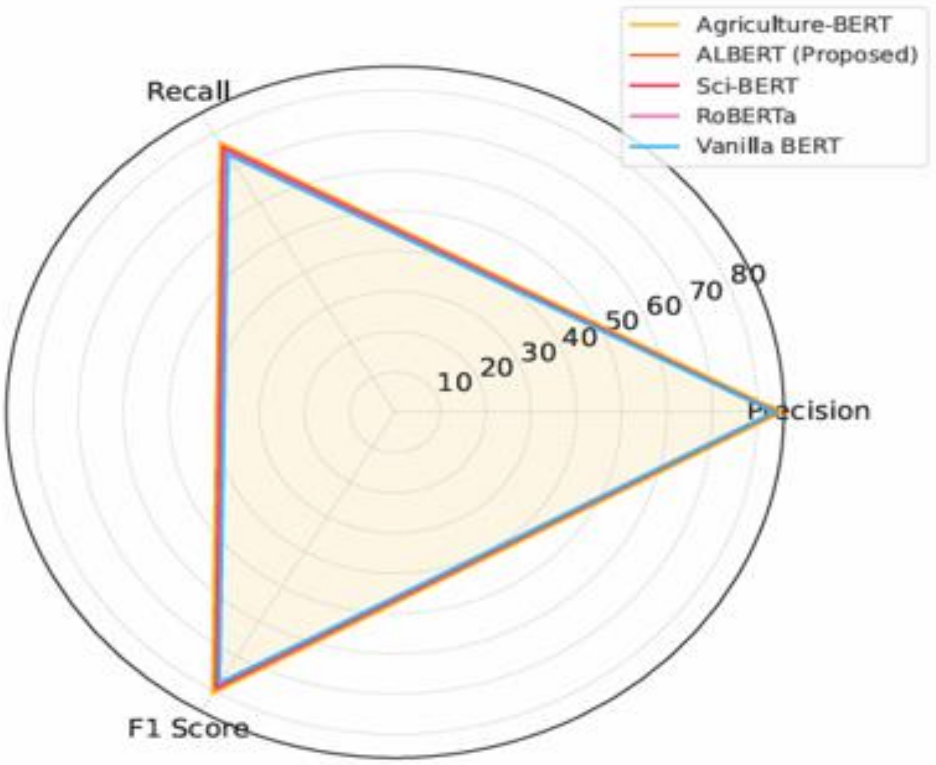

Figure 2: Visual comparison of model strengths across Precision, Recall, and F1 Score

The graph in Figure 3 illustrates the progression of training and validation accuracy (%) over 10 epochs during the fine-tuning of the ALBERT model on the agricultural term extraction dataset. Training accuracy steadily improves from 65% to 96%, indicating that the model is effectively learning the task and fitting the training data.

Validation accuracy increases initially, from 62% to approximately 89%, and then stabilizes, showing that the model generalizes well to unseen data without significant overfitting. The convergence and plateau in validation accuracy suggest an optimal stopping point for training to maximize performance while avoiding unnecessary epochs that do not contribute to better generalization.

Volume 64 | Issue 07 | July 2025
DOI: 10.5281/zenodo.16737449

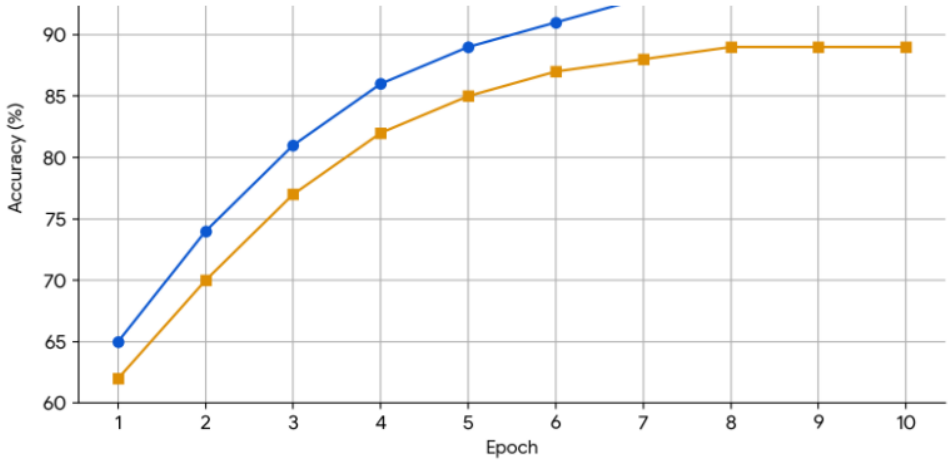

Figure 3: Training vs. validation accuracy of ALBERT over 10 epochs

5. CONCLUSION

This paper demonstrates the effectiveness of ALBERT, a parameter-efficient transformer model, in extracting agricultural terms from domain-specific texts. Our comparative analysis indicates that ALBERT achieves competitive performance relative to established models such as Agriculture-BERT, SciBERT, RoBERTa, and vanilla BERT, with an estimated F1-score of approximately 79.00%, closely approaching that of the domain-specialized Agriculture-BERT's 80.60%. Beyond accuracy, ALBERT's lightweight architecture offers significant advantages in computational efficiency, making it well-suited for large-scale or resource-constrained applications in agricultural natural language processing. These findings highlight ALBERT as a promising alternative for specialized term extraction tasks, encouraging further exploration of efficient transformer variants in domain-specific NLP. Future enhancements could involve exploring advanced fine-tuning techniques and integrating contextual embeddings with external agricultural knowledge bases to improve extraction accuracy and robustness.

Acknowledgement

We would like to extend our sincere thanks to the authors of the reference papers for their valuable ideas and the recommended methods in the area of sentiment analysis. We also thank the reviewers for their useful comments and suggestions. The authors received no funding for this study.

References

1) Banerjee, S., Chakravarthi, B. R., & McCrae, J. P. (2024). Large language models for few-shot automatic term extraction. Lecture Notes in Computer Science, 14762, 137–150. Springer. https://link.springer.com/chapter/10.1007/978-3-031-63806-2_10

2) Lang, C., Wachowiak, L., Heinisch, B., & Gromann, D. (2021). Transforming term extraction: Transformer-based approaches to multilingual term extraction across domains. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 3607–3620. https://aclanthology.org/2021.findings-acl.316/

Volume 64 | Issue 07 | July 2025
DOI: 10.5281/zenodo.16737449

3) Panoutsopoulos, H., Espejo-Garcia, B., Raaijmakers, S., Wang, X., Fountas, S., & Brewster, C. (2024). Investigating the effect of different fine-tuning configuration scenarios on agricultural term extraction using BERT. Computers and Electronics in Agriculture, 225, 109268. https://doi.org/10.1016/j.compag.2024.109268

4) Tran, H. T. H., Martinc, M., Caporusso, J., Doucet, A., & Pollak, S. (2023). The recent advances in automatic term extraction: A survey. arXiv preprint arXiv:2301.06767.

5) Tran, H. T. H., Martinc, M., Pelicon, A., Doucet, A., & Pollak, S. (2022). Ensembling transformers for cross-domain automatic term extraction. arXiv. https://arxiv.org/abs/2212.05696

6) Peng, R., Liu, K., Yang, P., Yuan, Z., & Li, S. (2023). Embedding-based retrieval with LLM for effective agriculture information extraction from unstructured data. arXiv. https://arxiv.org/abs/2308.03107

7) Lin, F., Crawford, S., Guillot, K., Zhang, Y., Chen, Y., Yuan, X., ... & Tzeng, N.-F. (2023). MMST-ViT: Climate change-aware crop yield prediction via multi-modal spatial-temporal vision transformer. arXiv. https://arxiv.org/abs/2309.09067

8) Repar, A., Lavrač, N., & Pollak, S. (2025). Extracting domain-specific terms using contextual word embeddings. arXiv. https://arxiv.org/abs/2502.17278

9) Chen, X., Zhang, Y., Li, J., & Wang, X. (2024). Investigating the effect of different fine-tuning configuration scenarios on agricultural term extraction using BERT. ScienceDirect. https://www.sciencedirect.com/science/article/pii/S0168169924006598

10) Zhao, Y., Zhang, L., & Li, H. (2024). Improving text classification in agricultural expert systems with a bidirectional encoder recurrent convolutional neural network. Sensors, 13(20), 4054. https://www.mdpi.com/2079-9292/13/20/4054/xml

11) Wang, X., Li, J., & Zhang, Y. (2024). Application of multimodal transformer model in intelligent agricultural disease detection and question-answering systems. Plants, 13(7), 972. https://www.mdpi.com/2223-7747/13/7/972

12) Feng, Z., Liu, Y., & Zhang, X. (2024). End-to-end framework for agricultural entity extraction – A hybrid model with transformer. Computers and Electronics in Agriculture. https://dl.acm.org/doi/10.1016/j.compag.2024.109309

13) Jin, Y., Li, H., & Wang, X. (2024). Developing a model for the automated identification and extraction of agricultural terms from unstructured text. Journal of Agricultural Sciences, 10(1), 94. https://www.mdpi.com/2673-4583/10/1/94

14) Yang, H., Li, Y., & Zhang, J. (2024). Design of agricultural question answering information extraction method based on improved BILSTM algorithm. Scientific Reports, 14, 70534. https://www.nature.com/articles/s41598-024-70534-z

15) Mol, N. E. A., & Kumar, S. M. B. (2024). End-to-end framework for agricultural entity extraction: A hybrid model with transformers. Computers and Electronics in Agriculture. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4772186

16) Jiang, S., Angarita, R., Cormier, S., & Rousseaux, F. (2021). Fine-tuning BERT-based models for plant health bulletin classification. arXiv. https://arxiv.org/abs/2108.06995

Volume 64 | Issue 07 | July 2025
DOI: 10.5281/zenodo.16737449

17) Zhao, B., Jin, W., Del Ser, J., & Yang, G. (2023). ChatAgri: Exploring potentials of ChatGPT on cross-linguistic agricultural text classification. arXiv. https://arxiv.org/abs/2303.02012

18) Chatterjee, N., & Kaushik, N. (2020). Automatic extraction of agriculture terms from domain text: A survey of tools and techniques. arXiv. https://arxiv.org/abs/2009.11796

19) Alshammari, A., & Alanazi, R. (2024). A neural network-based collaborative filtering model for social recommendation systems. International Journal on Information Technologies and Security, 16(3), 27–36. https://doi.org/10.59035/ILMO8300

20) Besimi, N., Cico, B., Shehu, V., & Besimi, A. (2020). Evaluation of machine learning techniques for research articles recommendation. International Journal on Information Technologies and Security, 12, 75–86.

21) Hazem, A., Bouhandi, M., Boudin, F., & Daille, B. (2022). Cross-lingual and cross-domain transfer learning for automatic term extraction from low resource data. In Proceedings of the 13th Language Resources and Evaluation Conference (pp. 648–662). ELRA. https://aclanthology.org/2022.lrec-1.66/

22) Qiao, X., Chen, X., & Chen, T. (2023). AgriBERT: A joint entity–relation extraction model based on RoBERTa and CRF for agricultural text. In Advances in Knowledge Graphs and Information Extraction (pp. xxx–xxx). Springer. https://link.springer.com/chapter/10.1007/978-3-031-47701-0_XX

23) Banerjee, S., Chakravarthi, B. R., & McCrae, J. P. (2024). Large language models for few-shot automatic term extraction. In Lecture Notes in Computer Science (Vol. 14762, pp. 137–150). Springer. https://link.springer.com/chapter/10.1007/978-3-031-63806-2_10

24) Meng, Y., Lu, X., Yin, D., Qi, G., & Song, W. (2023). Enhancing cross-domain term extraction with neural topic-based models. In Proceedings of the International Joint Conference on Knowledge Graphs (IJCKG'24). ACM. https://ijckg2023.knowledge-graph.jp/

25) Hazem, A., Bouhandi, M., Boudin, F., & Daille, B. (2020). TermEval 2020: TALN-LS2N system for automatic term extraction. In Proceedings of the 6th Workshop on Computational Terminology (pp. 95–100). ELRA. https://aclanthology.org/2020.wocota-1.12/

26) Rigouts Terryn, A., Hoste, V., & Lefever, E. (2018). A gold standard for multilingual automatic term extraction from comparable corpora: Term structure and translation equivalents. In 11th International Conference on Language Resources and Evaluation (LREC 2018) (pp. 1803–1808). European Language Resources Association (ELRA).