applied sciences

MDPI

Article

WordMap: Text Mining Application of Enhanced Corpus Segmentation and Semantic Topic Recognition

Zhijian Wei, Huiwen Zou, Patrick Cheong-Iao Pang *, Penny Wong-On Chao and Benjamin K. Ng

Faculty of Applied Sciences, Macao Polytechnic University, Macao, China; p2311393@mpu.edu.mo (Z.W.); p2413253@mpu.edu.mo (H.Z.); onpenny@gmail.com (P.W.-O.C.); bng@mpu.edu.mo (B.K.N.)
* Correspondence: mail@patrickpang.net

Abstract: This study presents WordMap, an integrated text mining application developed to enhance the efficiency and usability of text analysis over a network. As unstructured text data continues to grow across domains, effective tools for segmentation and topic modeling have become increasingly essential for extracting insightful information. However, most existing solutions depend on multiple disconnected tools, and these often compromise workflow efficiency and user experience. Unlike traditional tools, WordMap combines corpus segmentation, topic modeling, and result visualization into a unified workflow for both Chinese and English languages, thereby reducing workflow fragmentation and lowering the user threshold. To assess usability and user acceptance, this research adopts the Technology Acceptance Model (TAM). WordMap employs PKUSEG and NLTK for bilingual corpus segmentation, utilizes BERTopic for dynamic topic modeling, and integrates interactive visualization to enable intuitive analysis. The PLS-SEM result shows that the perceived ease of use (PEOU) has a significant impact on both perceived usefulness (PU) and user attitude (ATT), while ATT strongly predicts behavioral intention (BI) ($\beta = 0.674$, $p < 0.001$). The results indicate that integrating core text mining processes into a user-centered design significantly boosts user satisfaction and adoption. By combining key processes and empirically validating user perceptions, the proposed framework facilitates the development of efficient and accessible text mining tools. It offers both theoretical and practical insights for future advancement and deployment in the field of text mining.

Keywords: text mining; corpus segmentation; topic recognition; topic modeling; user acceptance; TAM; PLS-SEM

check for updates

Academic Editor: Douglas O'Shaughnessy

Received: 9 May 2025
Revised: 2 June 2025
Accepted: 10 June 2025
Published: 12 June 2025

Citation: Wei, Z.; Zou, H.; Pang, P.C.-I.; Chao, P.W.-O.; Ng, B.K. WordMap: Text Mining Application of Enhanced Corpus Segmentation and Semantic Topic Recognition. *Appl. Sci.* **2025**, *15*, 6632. https://doi.org/10.3390/app15126632

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The rapid expansion of unstructured text data in digital environments poses significant challenges for information processing and knowledge extraction. The exponential growth of text content from platforms like online communities [1] and news reports [2] as well as user reviews [3] not only stimulates the demand for efficient data processing but also drives advancements in text mining techniques. Functioning as a robust analytical tool, text mining has significantly reshaped how textual information is processed and utilized [4]. By combining natural language processing (NLP) and data mining techniques, text mining facilitates the extraction of meaningful insights and relevant knowledge from large-scale text corpora; this enables essential operations like text classification, semantic recognition, and topic analysis.

Nevertheless, traditional text mining approaches meet several challenges when dealing with complex semantic structures and large-scale textual data. Many conventional methods

rely heavily on manual rule-writing and method maintenance, which are not only time-consuming but also lack flexibility and scalability [5]. These constraints present significant challenges in terms of efficiency and accuracy, limiting their applicability in large-scale and dynamic text environments [6]. Moreover, existing text analysis workflows typically require the use of separate tools for text segmentation, topic modeling, and visualization. This causes users to switch environments and requires certain programming skills. This fragmented process seriously affects the efficiency of text segmentation, especially for users with limited technical background.

To address these challenges, it is crucial to develop more adaptable and scalable methods to improve the efficiency and accuracy of text mining. Consequently, this study developed an integrated text mining application, WordMap, to solve this problem by integrating different functions into a unified and easy-to-access application, and this can improve the efficiency of text mining and the overall user experience. Firstly, WordMap integrates automatic text corpus segmentation and employs PKUSEG (for Chinese) and NLTK (for English) to support bilingual word segmentation tags. Secondly, it combines BERTopic for dynamic topic modeling to improve the accuracy of topic detection by capturing context and semantic nuances. Thirdly, WordMap also adds interactive visualization tools that allow users to intuitively explore text task results and topic relationships. By simplifying these processes into an integrated framework, it helps users reduce workflow fragmentation, improves usability, and lowers the technical barrier for conducting text analysis. The system is available at https://github.com/mpu-patrick-lab/wordmap (accessed on 11 June 2025).

The organization of this paper is as follows: Section 2 provides an overview of previous studies on text mining techniques, including traditional text corpus segmentation and topic modeling methods. Section 3 outlines WordMap's system architecture, core methods, and evaluation approach. Section 4 presents the experimental results of specific applications, including functional demonstration and questionnaire results. Section 5 discusses the advantages of the integrated framework and its contributions to various application scenarios. Section 6 summarizes the research and proposes directions for future improvements.

2. Literature Review

Integrating text mining with deep learning has emerged as a critical research frontier for analyzing sophisticated unstructured data across domains. Early text segmentation methods mainly depended on rule-based strategies along with language-related features like syntactic structure [7], while traditional topic models like Latent Dirichlet Allocation (LDA) [8] and Non-Negative Matrix Factorization (NMF) [9] dominated semantic analysis. LDA modeled latent topics by using Dirichlet distributions, whereas NMF discovered themes through Non-Negative Matrix Factorization of term–document matrices [10]. However, traditional segmentation and topic modeling approaches exhibit limitations in adapting to domain-specific language and capturing nuanced semantics, thereby constraining their scalability in specialized contexts. These shortcomings have motivated the emergence of more advanced neural and hybrid models in recent years.

From 2021 onward, hybrid approaches combining neural networks and topic modeling emerged. Researchers have integrated Top2vec for topic modeling and RoBERTa for sentiment analysis to process over one hundred thousand news reports, which identified education, economy, and sports as key themes across five countries [11]. Recent research uses text mining for analyzing patient experience using large language models; however, it does not produce statistical results and visualizations automatically [12]. By conducting a large-scale news analysis, this process offers a noteworthy reference for understanding the coverage focus of different countries and provides data support for news communica-

tion research. Liu et al. [13] proposed a user recommendation framework that combines Collaborative Topic Modeling (CTM) with collaborative filtering and generative adversarial networks to improve personalization recently. Their model effectively captures users' potential interests from text content and improves the accuracy of recommendations.

In other domains, such as research into topic models and Artificial Intelligence, scholars have adopted the LDA topic model and Gibbs sampling technology for research [14]. These methods have been employed for semantic extraction and latent topic discovery. They have also successfully analyzed relevant comments in medical forums and implemented topic classification and valid sentiment identification. Visvam Devadoss explored the application of Artificial Intelligence and NLP in news platforms [15], conducting classification tasks and generating relevant content by analyzing social media trends and applying data mining methods. The developed system was able to generate news on similar topics through NLP analysis technology, which can effectively reduce the workload of journalists. In the social domain, other researchers have conducted studies in social media analytics research to make valuable predictions about public opinion and behavior by analyzing trends and sentiment on social media [16].

As topic modeling techniques have advanced, researchers have begun to evaluate the performance of traditional and modern models across different domains. A recent comparative study by Gan et al. empirically evaluated LDA, Top2Vec, and BERTopic across several text datasets. Their results show that BERTopic consistently outperformed the other two models in terms of topic coherence and diversity, particularly in multilingual and semantically rich corpora [17]. In addition, BERTopic has shown greater flexibility than models such as CTM in adapting to specific tasks. Albarrak et al. introduced U-BERTopic, which incorporated urgency signals and contextual embeddings to detect emerging cybersecurity threats [18]. The model achieved a higher accuracy and faster response compared to conventional topic models, and illustrated BERTopic's superior extensibility in dynamic, real-world applications.

Inspired by these developments, our work proposes a text mining application that enhances corpus segmentation and semantic topic recognition. By integrating text segmentation, topic analysis, visual knowledge graphs, and other features, we utilize the BERTopic topic model and NLP technology to process unstructured text data in a unified framework, providing users with a convenient text application. The aim is to solve the limitations of traditional text mining methods through data preprocessing, enhanced corpus segmentation, and semantic topic recognition, especially in processing unstructured text data.

3. Research Design and Assessment Model

3.1. Research Design

This study adopts a dual design of system development and questionnaire evaluation, which involves developing a text mining application with an integrated framework and utilizes TAM to conduct questionnaire evaluation to measure user acceptance.

3.1.1. System Process

Unstructured text mining seeks to uncover and analyze latent topics within large-scale unstructured textual data. This process entails effective preprocessing techniques to extract meaningful semantic information and enhance the accuracy of subsequent analyses through multi-module collaboration. In response to these needs, WordMap has been designed and implemented as an efficient and user-friendly text mining system, integrating key functionalities such as file uploading, text segmentation, topic modeling, and result visualization. The system delivers analytical results through an intuitive interface, thereby supporting a wide range of text data processing scenarios. The systematic design of

WordMap (as depicted in Figure 1) follows a user-centric approach, ensuring that text processing is both accurate and efficient.

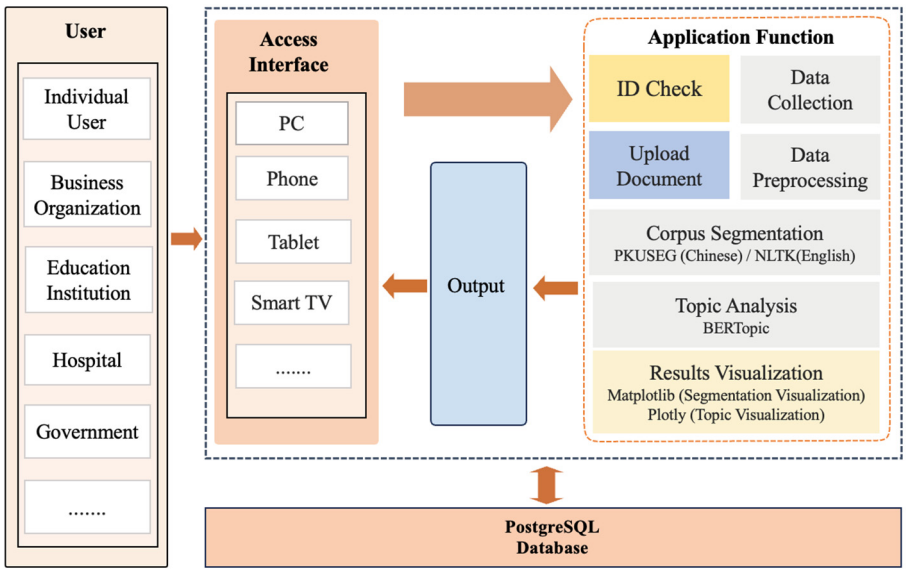


Figure 1. System process.

3.1.2. Modular Architecture

The WordMap system implements a modular layered architecture (as illustrated in Figure 2), which enhances system maintainability and future compatibility through a low-coupling design. Each functional module functions independently and supports parallel processing, whereby the word segmentation module and the visualization module can be invoked synchronously during text analysis to form a collaborative workflow. The architecture is also extensible and allows developers to incorporate additional modules at any layer in the future based on user requirements.

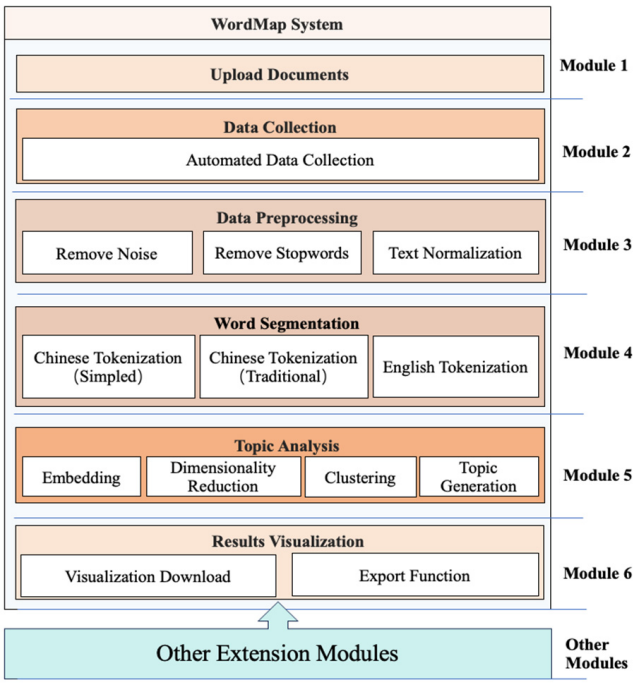


Figure 2. Modular architecture.

3.1.3. User Interface

The user interface (UI) design of the WordMap system emphasizes interactivity, usability, and simplicity of functions, with the goal of enabling users to complete text mining tasks efficiently and effectively. Users can seamlessly perform a series of processes. When the user logs in, the system authenticates through a secure interface and assigns access rights based on the user's role. Subsequently, the WordMap system supports uploading files in Excel format. After successful upload, the system automatically processes the text data and presents the text segmentation structure to the user. Users can query the segmentation results promptly on the results page and proceed to apply the topic analysis function. The UI caters to both technical and non-technical users (as illustrated in Figure 3), emphasizing intuitive layout, real-time system feedback, and interactive controls to guarantee that the text mining process is efficient and accessible.

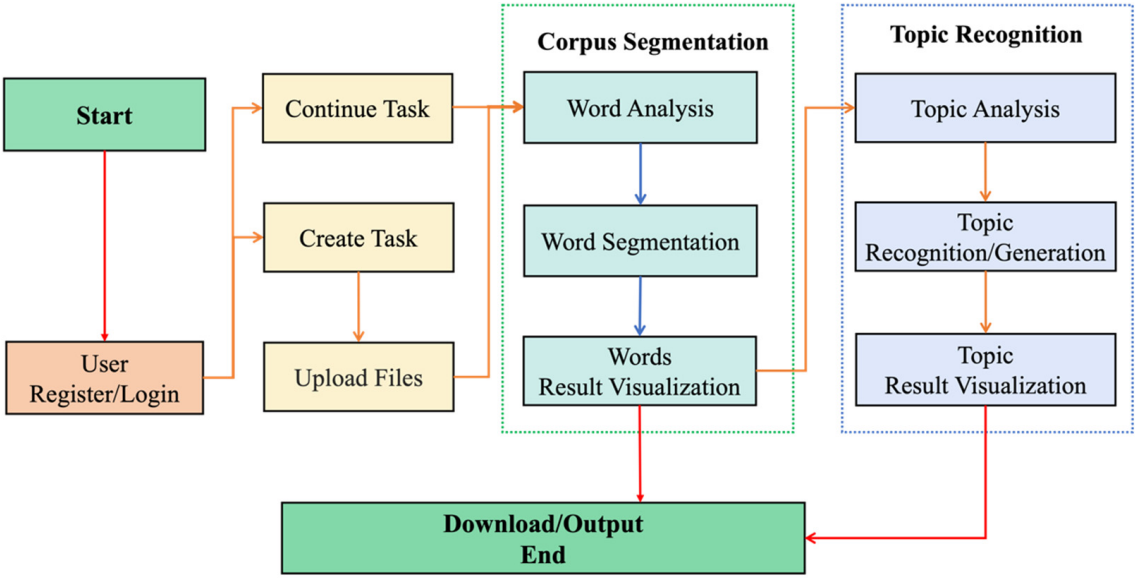


Figure 3. User interface flowchart.

3.2. Module Implementation

To improve the processing efficiency of text mining tasks, WordMap combines three key functional modules to build a unified integration framework. This section elaborates on the implementation of three core modules, including text segmentation, topic modeling, and visualization.

3.2.1. Corpus Segmentation

Corpus segmentation is essential for multilingual contexts involving Chinese and English. Chinese segmentation is challenging due to its semantic complexity. Traditional approaches are divided into three categories: rule-based, statistical, and machine learning-based methods [19]. Rule-based word segmentation methods rely on predefined linguistic rules and knowledge bases to match text [20]. This method has difficulty in accurately segmenting multiple-meaning and out-of-dictionary (OOD) words [21]. Statistical models [22], on the other hand, employ probabilistic techniques, such as Hidden Markov Models (HMMs) [23], Maximum Entropy Models (MEMs) [24], and Conditional Random Fields (CRFs) [25]. HMMs regard word segmentation as a sequence labeling task [26], but they are heavily reliant on training data. MEMs are based on the principle of information entropy and have high requirements for feature engineering [27]. CRFs have high computational complexity and inefficiency during the process of large-scale texts [28]. Therefore, these

word segmentation methods all have obvious limitations. With the advancement of neural networks such as LSTM [29], machine learning-based methods have demonstrated superior effectiveness in sequence labeling tasks [30]. Therefore, this system employs PKUSEG [31] for Chinese text segmentation. For English, NLTK [32] is used to implement morphological normalization and stop-word filtering. This bilingual pipeline ensures high-quality segmentation as a foundation for subsequent analytical tasks.

3.2.2. Topic Modeling

The system incorporates BERTopic (outperforms traditional methods such as NMF, Top2Vec) [33], and LDA by combining BERT embeddings with a class-based Term Frequency–Inverse Document Frequency (c-TF-IDF) algorithm [34]. BERT encodes contextual semantics into dense vector representations, while c-TF-IDF emphasizes term significance across topic clusters [35]. Uniform Manifold Approximation and Projection (UMAP) is employed for dimensionality reduction, preserving local structure, and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) automatically identifies optimal topic numbers [36]. The multilingual capability and contextual modeling of BERTopic enable precise topic clustering in unstructured text corpora.

3.2.3. Result Visualization

To present the analysis results in an intuitive and interpretable approach, the system integrates visualization modules including Highchart, knowledge graphs, heatmaps, and hierarchical clustering diagrams. These tools allow users to explore topic associations, keyword distributions, and topic similarity. Highchart supports frequency-based dynamic scaling and parallel comparisons, while topic networks visualize semantic relationships through variations in node size and edge thickness. This visualization layer enhances user interpretability and supports exploratory text mining and trend discovery [37].

3.3. Function Demonstration

The WordMap system offers a range of result visualization features designed to help users interpret text analysis outcomes intuitively. It provides various graphical representations and interactive views to facilitate users' understanding. Since the frequency of word occurrences varies across different text corpus, the system utilizes Highcharts to visualize the relative significance of each word in relation to its frequency. Words with higher frequencies are displayed with more prominent visual effects.

As shown in Figure 4, users can easily identify the key terms and main themes of the text through the word cloud, which aids in the comprehension and analysis of the text's topics.

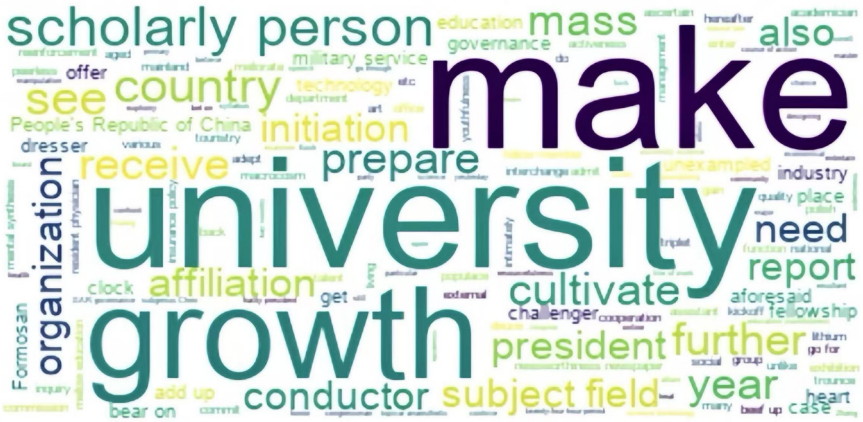


Figure 4. Word cloud diagram for visualization of keywords.

The word distribution in topics diagram visually represents the results of the text analysis through a histogram, displaying the distribution of words for each grouped topic. This facilitates an intuitive comparison of the word distribution across different topics. WordMap groups words into topics based on contextual and semantic similarity, which enable users to identify coherent thematic clusters. As illustrated in Figure 5, the words are clustered into four distinct topic groups. The word distributions are calculated based on the raw term frequency, which provides an interpretable and accessible representation for non-technical users.

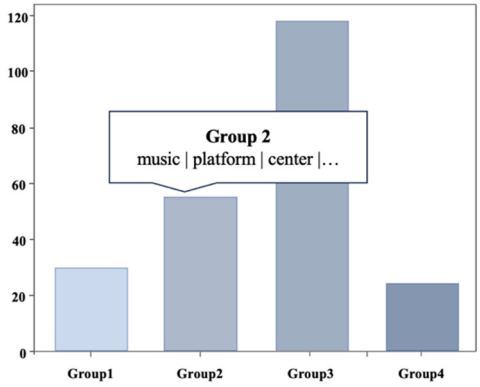


Figure 5. Word distribution in topics diagram for word frequency and intuitive comparison across topics.

An inter-topic distance map is a valuable visual tool in topic modeling. It effectively illustrates the proximity or divergence between distinct topics in the results of topic modeling processes. Specifically, this graphical map represents the inter-topic distances of topic modeling outcomes generated by using the BERTopic method. By analyzing the distance graph between topics generated by BERTopic, users can identify topics with similarities and group them into the same topic group. In Figure 6, the distribution of topics in the semantic space and the division of each topic group allows users to quickly locate the main topic groups.

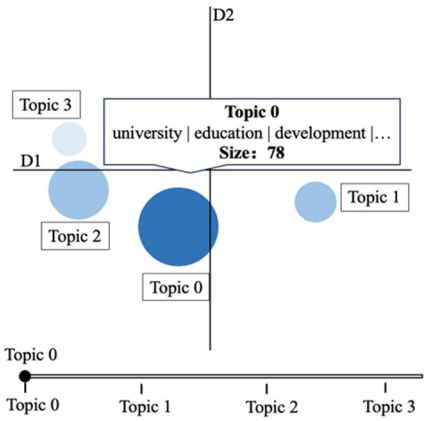


Figure 6. Inter-topic distance map for identification of the main topic clusters.

This Topic Probability Distribution represents how frequently each topic appears and how it is spread across the text dataset. The diagram illustrates both the comparative proportions and frequency distributions of different topics within the text corpus. With the topic distribution graph generated by BERTopic, users can intuitively see the relative proportion of different topics in the entire text dataset. This also supports interactive operations, and when the user clicks on a specific item, the complete information will be

displayed. Figure 7 shows the details of Topic 0–3, and this allows users to understand the topic distribution of text data.

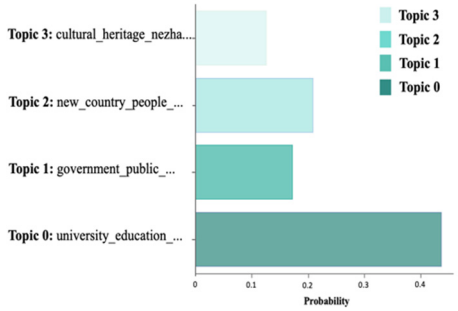


Figure 7. Topic distribution for the frequency and proportions of different topic groups.

The hierarchy chart serves as a visual representation that conveys the hierarchical structure among various topics. Within the chart, topics are organized based on their semantic similarity, and more similar topics are aggregated into an upper node, while more different topics exist as distinct nodes. It reflects the relationship between topics and sub-topics, while also assisting users in identifying underlying hierarchical patterns within text data, aiding in the analysis of complex text corpora. In Figure 8, Topic 0 and Topic1 are similar, so they are aggregated into one upper node. Topic 2 and Topic 3 are different nodes. By analyzing these nodes, users can better understand the semantic associations of different topics in the text and identify the main content and related types in the text. As shown in Figure 9, the Topic Word Scores Graph allows users to understand the connotations and characteristics of each topic. This visualization is based on c-TF-IDF, which identifies the most representative keywords for each topic, and serves as a deeper layer of topic analysis.

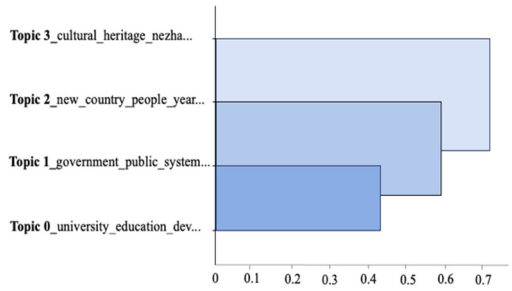


Figure 8. Hierarchical clustering graph for visualization of semantic associations and hierarchical structure.

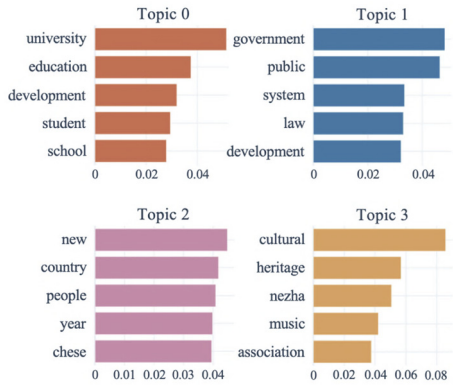


Figure 9. Topic word score for illustration of the proportions and scores of topics within each group.

The similarity matrix represents the semantic similarity across all topics in the form of a 2D heat map, with both the rows and columns associated with individual topics. Each cell

indicates the similarity score between a pair of topics. As the value increases, the semantic similarity between the corresponding topics strengthens. Users can quickly identify the similarity between topics by the color depth. As illustrated in Figure 10, a higher similarity between topics is represented by deeper blue shades, whereas lower similarity is indicated by lighter green tones. The relatively low similarity between Topic 1 and Topic 2 reflects substantial semantic divergence. This visualization enables users to intuitively interpret inter-topic relationships.

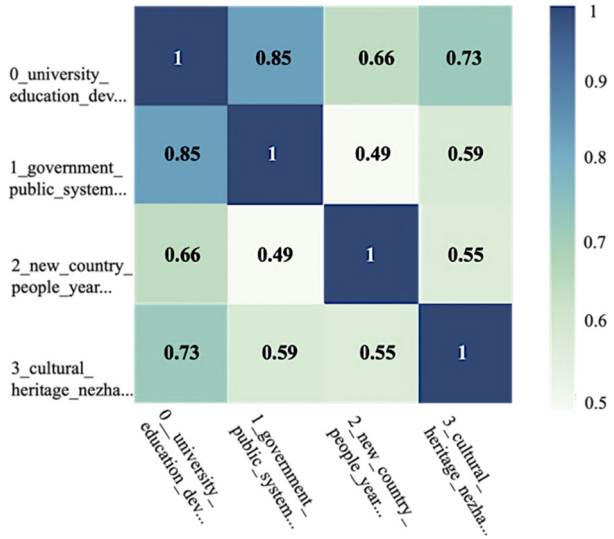


Figure 10. Heat map for visualization of semantic similarity and topic relationships by color depth.

3.4. Assessment Model

A questionnaire was used to assess the effectiveness of the system through the Wenjuanxing Survey Platform, which yielded 147 valid responses. Testing is essential to determine whether the WordMap application satisfies user needs. The survey assessment measures user acceptance, confirms the system's performance, and assesses the system's overall impact [38]. This survey examines the system from various perspectives, including external factors, PU, PEOU, ATT, and BI. Demographic data such as age, gender, and occupation were collected to ensure the representativeness of the sample. Additionally, external factors influencing usage were evaluated, and the users' perceptions of the system's utility and PEOU were measured. Based on the survey, this study proposes the following four hypotheses to explore the relationships between users' perceptions of the system and their behavioral intentions:

H1. *PU positively influences ATT.*

H2. *PEOU positively influences ATT.*

H3. *PEOU positively influences PU.*

H4. *ATT positively influences BI.*

This study employs SEM (see Figure 11) to build a model that validates users' acceptance and intention to use the WordMap system. Through a comprehensive analysis of the path-based associations derived from the TAM framework, the SEM model (as shown in Figure 11) could effectively expose how PEOU and PU influence ATT, which, in turn, affects BI. Consequently, the model can validate the system's usability and user acceptance.

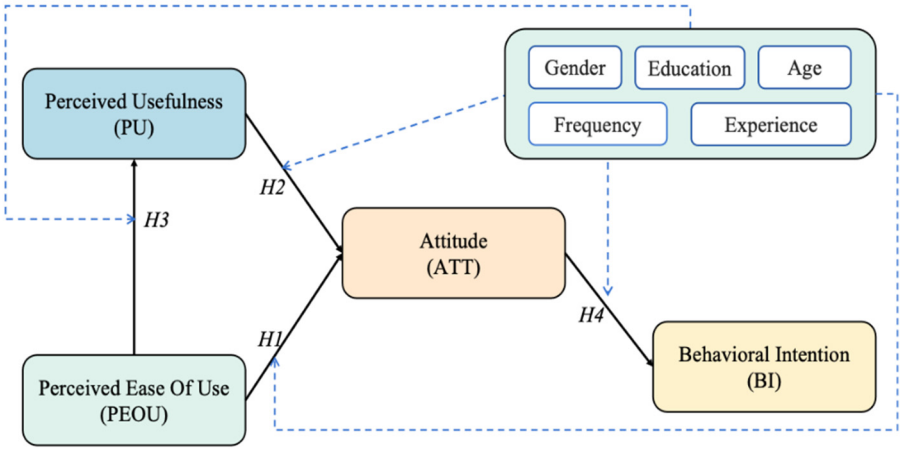


Figure 11. SEM structure.

4. Results

4.1. Results of T-Test

A paired samples *t*-test was conducted to evaluate the task completion time between the traditional text analysis method and the proposed method. The participants were instructed to complete three tasks: (1) upload a file and perform corpus segmentation; (2) conduct topic analysis based on the segmented content; and (3) use a visualization tool to examine the results. These tasks were carried out by using both a traditional text mining application and the WordMap under equivalent experimental conditions. Figure 12 illustrates the average time spent on text analysis using traditional applications versus WordMap.

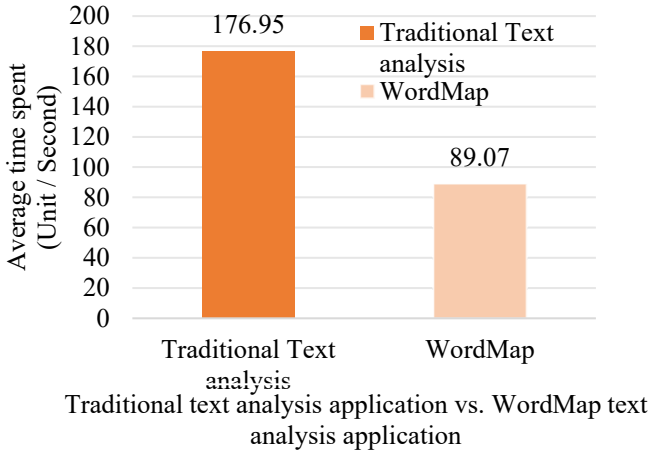


Figure 12. The results of paired *t*-test.

The results showed that the participants completed the task significantly faster by using the proposed method (M = 87.88, SD = 34.15), t (146) = 31.20, $p < 0.001$. The 95% confidence interval for the mean difference was [82.31, 93.44], demonstrating a robust advantage of the proposed method in terms of efficiency (see Table 1).

Table 1. Paired samples *t*-test.

	Paired Differences						
	Mean	SD	95% CI		t	df	Sig.
			LLCI	ULCI			
Time1-Time 2	87.88	34.15	82.31	93.44	31.20	146	<0.001

4.2. Sample Characteristics

The sample characteristics of the participants provide valuable context for evaluating the applicability of the WordMap system. From the sample of 147 individuals, 55.7% were male and 44.2% were female. A substantial proportion, 75.4%, were between the ages of 21 and 30. Regarding the participants' educational backgrounds, 34.6% were undergraduates, 40.1% held a master's degree, and 25.1% held a doctoral degree. Remarkably, 60.5% of respondents indicated familiarity with text mining. Among them, 44.2% of the respondents indicated occasional use of text mining tools, while 22.4% reported no prior usage. A proportion of 53% of the participants had prior experience with different text mining tools, suggesting a strong understanding of related applications. Concerning their views on text mining knowledge, 37.4% showed a desire for further learning, 39.4% felt no urgent need to acquire additional knowledge, and 23.1% were unsure. Detailed sample characteristics are provided in Table 2.

Table 2. Sample characteristics.

Demographic Information		Frequency	Percent (%)
Gender	Male	82	55.7
	Female	65	44.2
Age in year	18–20	36	24.4
	21–25	57	38.7
	26–30	54	36.7
Education	Undergraduate	51	34.6
	Master	59	40.1
	Doctor	37	25.1
Past experience (Text mining-related)	Yes	89	60.5
	No	58	39.4
Frequency (Text mining tools)	Always	65	44.2
	Sometime	49	33.3
	Never	33	22.4
Used other tools before	Yes	78	53
	No	69	46.9
Attitude toward text mining knowledge	Yes, Need knows more	55	37.4
	No	58	39.4
	Not sure	34	23.1

4.3. Results of SEM-PLS

This study assessed the reliability and validity of the proposed variables within the measurement model [39]. Based on the composite reliability and Cronbach's alpha values, each measure surpassed the general threshold of 0.6, confirming the internal consistency reliability. Additionally, the average variance extracted (AVE) surpassed the 0.50 cutoff [40], reinforcing the convergent validity (see Table 3).

The Fornell–Larcker criterion was used to evaluate the discriminant validity (see Table 4), where the square roots of the AVE for each construct were greater than their correlations with other constructs, satisfying the required standard.

Regarding hypothesis testing, all of the structural associations highlight their importance and significance through the magnitude of standardized values. The results from the SEM analysis are presented in Figure 13. This study used SEM to test the hypotheses, and the findings supported the extension of TAM theory, validating the model's applicability to the field of text mining [41].

Table 3. Critical indicators of CFA results.

Construct	Item	Factor Loading	Cronbach's Alpha	AVE	Composite Reliability
PU	PU1	0.751	0.696	0.521	0.812
	PU2	0.750			
	PU3	0.618			
	PU4	0.758			
PEOU	PEOU1	0.709	0.689	0.517	0.810
	PEOU2	0.758			
	PEOU3	0.646			
	PEOU4	0.758			
ATT	ATT1	0.744	0.691	0.520	0.812
	ATT2	0.786			
	ATT3	0.676			
	ATT4	0.672			
BI	BI1	0.719	0.649	0.586	0.809
	BI2	0.745			
	BI3	0.828			

Table 4. Analysis results of discriminant validity.

	ATT	BI	PEOU	PU
ATT	**0.721**			
BI	0.674 **	**0.766**		
PEOU	0.545 **	0.546 **	**0.719**	
PU	0.514 **	0.385 **	0.431 **	**0.722**

Note: The values on the diagonal represent square roots of AVE; ** $p < 0.01$.

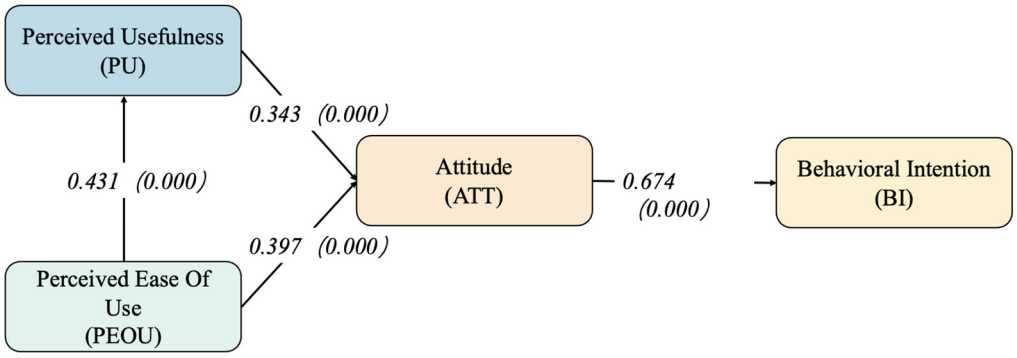


Figure 13. SEM results.

Table 5 presents the detailed statistical results of the hypothesis testing, including path coefficients, *T*-values, *p*-values, and significance levels for all structural relationships identified in the model.

Table 5. Results of hypothesis testing.

Hypothesis (Path)	Coeff.	T-Value	*p*-Value	Conclusion
H1 (PU → ATT)	0.343	4.599	$p < 0.001$	Supported
H2 (PEOU → ATT)	0.397	5.246	$p < 0.001$	Supported
H3 (PEOU → PU)	0.431	5.688	$p < 0.001$	Supported
H4 (ATT → BI)	0.674	13.767	$p < 0.001$	Supported

Note: Coeff. = Path Coefficients.

Based on the results shown in Table 5, all of the proposed hypotheses were supported by statistically significant path relationships:

H1 (PU → ATT). *Supported. PU had a significant positive effect on ATT ($\beta = 0.343$, $t = 4.599$, $p < 0.001$).*

H2 (PEOU → ATT). *Supported. PEOU also had a significant positive influence on ATT ($\beta = 0.397$, $t = 5.246$, $p < 0.001$).*

H3 (PEOU → PU). *Supported. PEOU had a positive impact on PU ($\beta = 0.431$, $t = 5.688$, $p < 0.001$).*

H4 (ATT → BI). *Supported. ATT had a strong positive impact on BI ($\beta = 0.674$, $t = 13.767$, $p < 0.001$).*

Overall, these results provide strong empirical support for the proposed TAM in this study. They confirm the significance of each hypothesized relationship and the practical relevance of the TAM in explaining the users' behavioral intentions.

5. Discussion

5.1. Evaluation and Implications

This study has developed the WordMap system by using a highly integrated framework, aiming to evaluate both its performance and user acceptance. This section discusses the performance and user feedback of the WordMap system, as well as the benefits offered by its integrated framework.

In theoretical terms, this study verified the applicability of WordMap text mining applications among users through the TAM and contributed to advancements in text mining research. The empirical results indicated that PEOU has a direct positive impact on both PU and ATT, indicating that an intuitive system can effectively enhance users' perception of its value and generate a positive attitude towards use. Furthermore, PU significantly influences ATT, and ATT has a significant predictive effect on BI. This means that users' perception of the system's practicality can be transformed into a positive attitude, which serves as the core driving force for users to continue using WordMap. In practical terms, the intuitive interactive interface and integrated framework of WordMap provide both usability for processing text tasks and convenience for non-technical users. WordMap demonstrates strong potential in specific application scenarios. Specifically, we evaluated the system using real-world news datasets, which included articles from various domains such as education, technology, and art. These evaluations confirmed WordMap's ability to perform corpus segmentation, topic analysis, and visualization across distinct subject areas. WordMap's capacity to efficiently process large-scale unstructured text suggests strong potential for applications in areas such as sentiment analysis (e.g., [42]) and social media data-driven analysis (e.g., [43]). Moreover, the system's extensible framework makes it well-suited for cross-domain text mining tasks, including financial analysis, healthcare analytics, marketing research, and legal document processing.

This study validated the effectiveness and user acceptance of the WordMap system while demonstrating its adaptability to real-world applications. By integrating a user-friendly interface with powerful analytical capabilities, WordMap represents a meaningful advancement in making text mining more accessible, practical, and relevant across various domains.

5.2. Limitations

While the WordMap system has demonstrated strong performance and practical value in processing large volumes of unstructured text, several limitations were identified during testing that warrant attention. One of the limitations is that while BERTopic generally performs well in common domains, the system may struggle to generate coherent or

meaningful topics when applied to small corpora, due to insufficient word frequency and limited contextual richness. Another limitation concerns the handling of noisy or domain-specific text. Although standard preprocessing steps have been implemented, manual cleaning and adjustment are often still necessary to ensure accurate results. However, to maintain ease of use for non-technical users, the current system does not support manual intervention during preprocessing.

These limitations suggest that the future development of WordMap should focus on enhancing its adaptability to specialized domain language, increasing robustness in noisy data environments, and improving the stability of large-scale data processing. Addressing these issues will be essential for further optimizing the functionality and reliability of the WordMap system in practical applications.

5.3. Future Work

Based on the current limitations, future improvements of the WordMap system will focus on enhancing its adaptability, usability, and robustness across diverse application scenarios. Firstly, to improve topic modeling performance on small-scale corpora, we plan to explore techniques such as data augmentation and few-shot learning. These methods can help compensate for limited word frequency and contextual richness, which currently hinder topic coherence in smaller datasets. Secondly, for noisy or domain-specific text, future versions of WordMap will consider incorporating optional manual preprocessing modules. This would allow technically proficient users to make targeted adjustments to input data without compromising the overall simplicity and usability of the system for non-technical users. Thirdly, to further improve the clarity and interpretability of topic visualization, we will experiment with alternative dimensionality reduction methods, such as t-SNE and PCA. These techniques may offer enhanced topic separation and reduce visual overlap in the two-dimensional projection space. We also plan to integrate additional analytical functionalities, such as sentiment analysis modules, to extend the system's capabilities and better address diverse user needs.

6. Conclusions

The WordMap system establishes a comprehensive and automated workflow for text mining by integrating corpus preprocessing, topic modeling, and result visualization into a modular and scalable framework. This networked application supports multilingual processing of simplified Chinese, traditional Chinese and English texts, ensuring data quality through cleaning, segmentation, and normalization. By employing an improved BERTopic model, the system enables accurate extraction of latent topics and reveals hierarchical semantic structures, facilitating in-depth topic exploration. The visualization module includes topic networks, word clouds, and statistical charts, which enhance interpretability through interactive and multi-level displays.

The highly integrated design of the system significantly improves user satisfaction and acceptance. The questionnaire results demonstrate the stability of the system when dealing with large-scale unstructured text, the intuitiveness of the interface, and the efficiency of the analysis process. The WordMap provides a practical and extensible solution for unstructured text analysis, offering valuable support for academic research and data-driven decision-making. Its modular design and automated pipeline reduce manual intervention, allowing users to efficiently extract valuable information from large-scale text corpora. User feedback indicates high levels of system acceptance, with its features highly valued for supporting exploratory analysis and decision-making.

In conclusion, the proposed integrated text mining framework enhances both the efficiency and usability of text mining processes. The WordMap application offers a

practical and scalable platform for unstructured text analysis, enabling researchers to rapidly uncover topic distribution patterns in large corpora and providing actionable data insights for informed decision-making. Future work will focus on further performance optimization and functional expansion to meet increasingly diverse analytical demands.

Author Contributions: Conceptualization, Z.W. and P.C.-I.P.; Methodology, Z.W.; Software, Z.W.; Validation, Z.W. and H.Z.; Formal analysis, Z.W. and H.Z.; Investigation, Z.W. and P.C.-I.P.; Resources, P.C.-I.P.; Data curation, Z.W.; Writing—original draft, Z.W.; Writing—review & editing, H.Z., P.C.-I.P. and P.W.-O.C.; Visualization, Z.W.; Supervision, P.C.-I.P., P.W.-O.C. and B.K.N.; Project administration, P.C.-I.P. and B.K.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the Macao Science and Technology Development Fund (funding ID: 0048/2021/APD) and Macao Polytechnic University research grant (project code: RP/FCA-10/2022).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of Macao Polytechnic University (project code HEA002-FCA-2024; approved on 27 May 2024).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The dataset is publicly available at https://github.com/mpu-patrick-lab/wordmap (accessed on 11 June 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kim, H.; Cho, I.; Park, M. Analyzing genderless fashion trends of consumers' perceptions on social media: Using unstructured big data analysis through Latent Dirichlet Allocation-based topic modeling. *Fash. Text.* **2022**, *9*, 6. [CrossRef]
2. Cano-Marin, E.; Mora-Cantallops, M.; Sanchez-Alonso, S. The power of big data analytics over fake news: A scientometric review of Twitter as a predictive system in healthcare. *Technol. Forecast. Soc. Change* **2023**, *190*, 122386. [CrossRef]
3. Blasco-Arcas, L.; Lee, H.-H.M.; Kastanakis, M.N.; Alcañiz, M.; Reyes-Menendez, A. The role of consumer data in marketing: A research agenda. *J. Bus. Res.* **2022**, *146*, 436–452. [CrossRef]
4. Pichiyan, V.; Muthulingam, S.; G, S.; Nalajala, S.; Ch, A.; Das, M.N. Web Scraping using Natural Language Processing: Exploiting Unstructured Text for Data Extraction and Analysis. *Procedia Comput. Sci.* **2023**, *230*, 193–202. [CrossRef]
5. Lee, S.; Song, J.; Kim, Y. An Empirical Comparison of Four Text Mining Methods. *J. Comput. Inf. Syst.* **2010**, *51*, 1–10.
6. Antons, D.; Grünwald, E.; Cichy, P.; Salge, T.O. The application of text mining methods in innovation research: Current state, evolution patterns, and development priorities. *RD Manag.* **2020**, *50*, 329–351. [CrossRef]
7. Xuan Bach, N.; Le Minh, N.; Shimazu, A. A Reranking Model for Discourse Segmentation using Subtree Features. In Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Seoul, Republic of Korea, 5–6 July 2012; Lee, G.G., Ginzburg, J., Gardent, C., Stent, A., Eds.; Association for Computational Linguistics: Seoul, Republic of Korea, 2012; pp. 160–168. Available online: https://aclanthology.org/W12-1623/ (accessed on 3 January 2025).
8. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 933–1002.
9. Lee, D.; Seung, H.S. Algorithms for Non-negative Matrix Factorization. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation: Denver, CO, USA, 2000; Volume 13, Available online: https://proceedings.neurips.cc/paper_files/paper/2000/hash/f9d1152547c0bde01830b7e8bd60024c-Abstract.html (accessed on 3 January 2025).
10. Lee, J.-H.; Park, S.; Ahn, C.-M.; Kim, D. Automatic generic document summarization based on non-negative matrix factorization. *Inf. Process. Manag.* **2009**, *45*, 20–34. [CrossRef]
11. Ghasiya, P.; Okamura, K. Investigating COVID-19 News Across Four Nations: A Topic Modeling and Sentiment Analysis Approach. *IEEE Access Pract. Innov. Open Solut.* **2021**, *9*, 36645–36656. [CrossRef]
12. Zhu, Q.; Chen, R.; Pang, P.C.-I.; Li, J.; Mao, C. Identifying Kidney Stone Risk Factors Through Patient Experiences with a Large Language Model: Text Analysis and Empirical Study. *J. Med. Internet Res.* **2025**, *27*, e66365. [CrossRef]
13. Liu, D.-R.; Huang, Y.; Jhao, J.-J.; Lee, S.-J. News recommendations based on collaborative topic modeling and collaborative filtering with generative adversarial networks. *Data Technol. Appl.* **2024**, *58*, 24–41. [CrossRef]
14. Jelodar, H.; Wang, Y.; Orji, R.; Huang, S. Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 2733–2742. [CrossRef] [PubMed]

15. Devadoss, A.K.V.; Thirulokachander, V.R.; Devadoss, A.K.V. Efficient daily news platform generation using natural language processing. *Int. J. Inf. Technol.* **2019**, *11*, 295–311. [CrossRef]
16. Song, M.; Hu, C.; Yuan, J.; Zhang, A.; Liu, X. Toward an ecological civilization: Exploring changes in China's land use policy over the past 35 years using text mining. *J. Clean. Prod.* **2023**, *427*, 139265. [CrossRef]
17. Gan, L.; Yang, T.; Huang, Y.; Yang, B.; Luo, Y.Y.; Richard, L.W.C.; Guo, D. Experimental Comparison of Three Topic Modeling Methods with LDA, Top2Vec and BERTopic. In *Artificial Intelligence and Robotics. ISAIR 2023*; Lu, H., Cai, J., Eds.; Communications in Computer and Information Science; Springer: Singapore, 2024; Volume 1998. [CrossRef]
18. Albarrak, M.; Pergola, G.; Jhumka, A. U-BERTopic: An Urgency-Aware BERT-Topic Modeling Approach for Detecting CyberSecurity Issues via Social Media. In Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security, Lancaster, UK, 29–30 July 2024; Mitkov, R., Ezzini, S., Ranasinghe, T., Ezeani, I., Khallaf, N., Acarturk, C., Bradbury, M., El-Haj, M., Rayson, P., Eds.; International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security: Lancaster, UK, 2024; pp. 196–211. Available online: https://aclanthology.org/2024.nlpaics-1.22/ (accessed on 5 January 2025).
19. Cui, M.; Huang, R.; Hu, Z.; Xia, F.; Xu, X.; Qi, L. Semantic rule-based information extraction for meteorological reports. *Int. J. Mach. Learn. Cybern.* **2024**, *15*, 177–188. [CrossRef]
20. Palmer, D.D. A Trainable Rule-Based Algorithm for Word Segmentation. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics, Madrid, Spain, 7–12 July 1997; pp. 321–328. [CrossRef]
21. Vo, B.-K.H.; Collier, N. Twitter Emotion Analysis in Earthquake Situations. *Int. J. Comput. Linguist. Appl.* **2013**, *4*, 159–173.
22. Beeferman, D.; Berger, A.; Lafferty, J. Statistical Models for Text Segmentation. *Mach. Learn.* **1999**, *34*, 177–210. [CrossRef]
23. Khare, R.; An, Y. An empirical study on using hidden markov model for search interface segmentation. In Proceedings of the 18th ACM Conference on Information and Knowledge Management, Hong Kong, China, 2–6 November 2009; pp. 17–26. [CrossRef]
24. Zhang, L.-Y.; Qin, M.; Zhang, X.-M.; Ma, H.-X. A Chinese word segmentation algorithm based on maximum entropy. *Int. Conf. Mach. Learn. Cybern.* **2010**, *3*, 1264–1267. [CrossRef]
25. Vemulapalli, R.; Tuzel, O.; Liu, M.-Y.; Chellapa, R. Gaussian Conditional Random Field Network for Semantic Segmentation. 2016, pp. 3224–3233. Available online: https://openaccess.thecvf.com/content_cvpr_2016/html/Vemulapalli_Gaussian_Conditional_Random_CVPR_2016_paper.html (accessed on 9 January 2025).
26. Xue, N.; Shen, L. Chinese Word Segmentation as LMR Tagging. In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, Sapporo, Japan, 11–12 July 2003; pp. 176–179. [CrossRef]
27. Zhao, Y.; Fu, G. A MEMs-based Labeling Approach to Punctuation Correction in Chinese Opinionated Text. In Proceedings of the International Conference on Artificial Intelligence (ICAI), Las Vegas, NV, USA, 22–25 July 2013; pp. 329–335.
28. Wang, Y.; Shi, C.; Xiao, B.; Wang, C.; Qi, C. CRF based text detection for natural scene images using convolutional neural network and context information. *Neurocomputing* **2018**, *295*, 46–58. [CrossRef]
29. Koshorek, O.; Cohen, A.; Mor, N.; Rotman, M.; Berant, J. Text Segmentation as a Supervised Learning Task. *arXiv* **2018**, arXiv:1803.09337. [CrossRef]
30. Iosifov, I.; Iosifova, O.; Sokolov, V. Sentence Segmentation from Unformatted Text using Language Modeling and Sequence Labeling Approaches. In Proceedings of the 2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T), Kharkiv, Ukraine, 6–9 October 2020; pp. 335–337. [CrossRef]
31. Luo, R.; Xu, J.; Zhang, Y.; Zhang, Z.; Ren, X.; Sun, X. PKUSEG: A Toolkit for Multi-Domain Chinese Word Segmentation. *arXiv* **2022**, arXiv:1906.11455. [CrossRef]
32. Yogish, D.; Manjunath, T.N.; Hegadi, R.S. Review on Natural Language Processing Trends and Techniques Using NLTK. In *Recent Trends in Image Processing and Pattern Recognition*; Santosh, K.C., Hegadi, R.S., Eds.; Springer: Berlin/Heidelberg, Germany, 2019; pp. 589–606. [CrossRef]
33. Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv* **2022**, arXiv:2203.05794. [CrossRef]
34. Albanese, N.C. Topic Modeling with LSA, pLSA, LDA, NMF, BERTopic, Top2Vec: A Comparison; Towards Data Science: 2022. Available online: https://towardsdatascience.com/topic-modeling-with-lsa-plsa-lda-nmf-bertopic-top2vec-a-comparison-5e6ce4b1e4a5/ (accessed on 15 January 2025).
35. Lalitha, T.B.; Sreeja, P.S. Title-Based Topic Modeling on E-learning Web Content Titles Using BERTopic Model. In *Fifth International Conference on Computing and Network Communications*; Thampi, S.M., Chaudhary, V., Pathan, A.-S.K., Li, K.C., Krishnaswamy, D., Eds.; Springer Nature: Singapore, 2025; pp. 559–580. [CrossRef]
36. Rachel J., J.L.; Bhuvaneswari, A.; Kumudha, M. Topic Modeling Based Clustering of Disaster Tweets Using BERTopic. In Proceedings of the 2024 MIT Art, Design and Technology School of Computing International Conference (MITADTSoCiCon), Pune, India, 25–27 April 2024; pp. 1–6. [CrossRef]
37. Kirsh, I. Visualizing Web Users' Attention to Text with Selection Heatmaps. In *Web Engineering*; Brambilla, M., Chbeir, R., Frasincar, F., Manolescu, I., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; pp. 517–520. [CrossRef]

38. Wallace, L.G.; Sheetz, S.D. The adoption of software measures: A technology acceptance model (TAM) perspective. *Inf. Manag.* **2014**, *51*, 249–259. [CrossRef]
39. Hair, J.F.; Risher, J.J.; Sarstedt, M.; Ringle, C.M. When to use and how to report the results of PLS-SEM. *Eur. Bus. Rev.* **2019**, *31*, 2–24. [CrossRef]
40. Cheung, G.W.; Wang, C. Current Approaches for Assessing Convergent and Discriminant Validity with SEM: Issues and Solutions. *Acad. Manag. Proc.* **2017**, *2017*, 12706. [CrossRef]
41. Venkatesh, V.; Morris, M.G.; Davis, G.B.; Davis, F.D. User Acceptance of Information Technology: Toward a Unified View. *MIS Q.* **2003**, *27*, 425–478. [CrossRef]
42. Guo, J.; Yang, X.; Wang, Y.; Pang, P.C.-I.; Im, S.-K.; Li, J.; Yang, Y. Performance Evaluation and Application Potential of Small Large Language Models in Complex Sentiment Analysis Tasks. *IEEE Access* **2025**, *13*, 49007–49017. [CrossRef]
43. Si, Y.-W.; Sun, L.; Pang, P.C.-I. Roles of Information Propagation of Chinese Microblogging Users in Epidemics: A Crisis Management Perspective. *Internet Res.* **2021**, *31*, 540–561. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.