JURNAL INFOTEL
VOL. 18, NO. 3, AUGUST 2026, PP. 510–530. DOI:10.20895/INFOTEL.V18I3.10725

RESEARCH ARTICLE

Integrating Topic Modeling and Contextual Analysis for Automated Online Newspaper Mining and Summarization

Dwi Adi Purnama[1,2,*]

[1]Department of Industrial Engineering, Faculty of Industrial Technology, Universitas Islam Indonesia, Yogyakarta 55584, Indonesia
[2]Department of Engineering Management, Faculty of Industrial Technology, Universitas Islam Indonesia, Yogyakarta 55584, Indonesia

*Corresponding email:dwiadipurnama@uii.ac.id

Received: February 20, 2026; Revised: August 30, 2026; Accepted: September 09, 2026.

Abstract: The rapid growth of online newspapers has increased the effort required to identify and consolidate relevant information across multiple sources. This study proposes an Article News Mining framework that combines TF-IDF-based term extraction, Latent Dirichlet Allocation (LDA), bi-gram phrase enrichment, and objective-driven 5W1H-oriented information extraction. The contribution is framework-level rather than a new topic-modeling algorithm: the pipeline connects thematic discovery, contextual phrase interpretation, and purpose-specific detail retrieval in a transparent workflow. The framework is demonstrated through a retrospective MacBook-related case study and is evaluated by comparing machine-generated summaries with summaries produced by ten respondents across the What, Who, When, Where, and Why dimensions using TF-IDF-weighted text similarity. The preserved research archive contains the aggregate topic and validation outputs, but not the complete article-level collection log or original model-configuration record; therefore, the case study is interpreted as an illustrative proof of concept rather than a fully reproducible benchmark. The reported statistical tests do not identify significant differences for the tested comparisons ($p > 0.05$), which is interpreted as a failure to detect a difference rather than proof of statistical equivalence. Extracted numerical product claims are separately cross-checked against authoritative product information. The proposed framework provides an interpretable approach for online news mining and decision support while making its validation and archival limitations explicit.

Keywords: online news mining, text summarization, topic modeling, latent dirichlet allocation, decision support system

© Jurnal Infotel ISSN: 2085-3688; e-ISSN: 2460-0997

1 Introduction

Mass communication plays an important role in modern society by encouraging the rapid dissemination of information through various media platforms. The availability of information not only supports public awareness but also influences the decision-making process and the formation of collective opinions [1], [2]. Along with the rapid advancement of internet technology, digital media, especially online newspapers, are gradually replacing traditional print media as the main source of information.

Online news sites make it easy for users to quickly and easily access information about specific topics they are interested in. According to Gandasari & Dwidienawati [3], newspapers remain one of the most important forms of mass media in simultaneously disseminating information to a wide audience. Statistical reports from the China Internet Network Information Center (CNNIC) also show that online news has become one of the most frequently accessed types of internet applications [4]. More than a dozen online news portals have been officially registered in Indonesia [5], with each offering news coverage from a different perspective and editorial approach.

Despite these advantages, the abundance of online news sources presents new challenges. News articles that cover the same event or topic often differ in content, emphasis, and depth depending on the source [6]. Getting information from several newspapers can improve the quality and completeness of information [7]. However, this requires considerable time and effort. On the other hand, relying on a few sources alone can speed up the process of retrieving information, but risks missing critical insights [8]. These differences create significant challenges in efficiently identifying and summarizing high-quality information from so many online news sources.

The need for effective online news mining is crucial in the context of business and strategic decision-making. Inaccurate or incomplete information at an early stage of decision-making can lead to major risks, such as failure in new product development due to misunderstanding market trends or competitors' strategies [9]. Online newspapers contain important information related to product innovations, market dynamics, competitors' actions, and technology trends. However, manually processing such information from a large number of articles is inefficient and impractical.

Previous studies have already combined topic information with summarization, so the research gap should not be framed as the complete absence of integrated approaches. Topic-guided multi-document summarization has jointly used latent topics to guide summary generation [10], topic-modeled extractive methods have combined LDA with sentence selection [11], and rule-based 5W1H systems have extracted journalistic event descriptors for news analysis and summarization [12]. Accordingly, the narrower gap addressed in this study is the need for a lightweight and transparent modular workflow that links three levels of analysis for multi-source online news: corpus-level topic discovery, phrase-level contextual interpretation, and user-objective information retrieval. The emphasis is therefore on interoperable framework design and transparent operationalization rather than on claiming a new underlying TF-IDF or LDA algorithm.

To address this operational gap, this study proposes a modular Article News Mining framework that produces both a general thematic overview and an objective-specific information summary. The framework contribution consists of three aspects:

1. a structured pipeline that connects TF-IDF term filtering, LDA topic discovery, contextual bi-grams, and objective-oriented detail extraction;

2. an explicit operationalization of sentence relevance and 5W1H-oriented retrieval so that the information-selection logic can be inspected and transferred to future implementations; and
3. a two-part evaluation that compares machine summaries with human respondent summaries and separately verifies extracted factual claims against authoritative sources.

The framework is intended to support business intelligence, market analysis, and strategic decision-making by reducing the amount of source material that must be reviewed manually. The evaluation is interpreted conservatively: similarity to human summaries is evidence of comparable content coverage under the selected metric, whereas a non-significant t-test alone is not treated as evidence of statistical equivalence.

2 Research Method

This study proposes an Article News Mining framework for systematically identifying and summarizing information from multiple online newspaper sources. The workflow comprises: (1) TF-IDF-based term extraction, (2) LDA topic modeling with bi-gram contextual enrichment and topic-relevant sentence selection, and (3) objective-driven 5W1H-oriented information extraction. The framework does not introduce a new TF-IDF or LDA algorithm; its contribution is the transparent orchestration of these components to generate both a general summary and a purpose-specific information view. Figure 1 presents the complete workflow, preceded by data collection and text preprocessing.

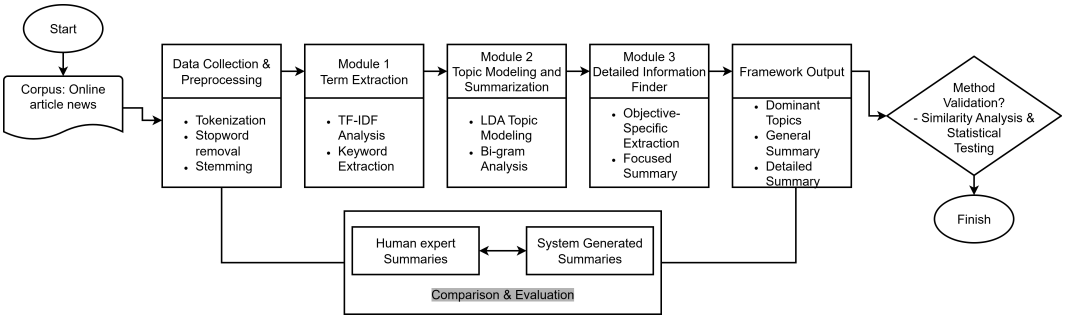


Figure 1: Research method.

2.1 Data Collection and Preprocessing

The case study was built from publicly accessible online news related to Apple MacBook products around the November 2020 M1 launch period. The retained research archive preserves the processed aggregate outputs used in Table 1– Table 5 and Figure 1–Figure 4, but the original article-level collection log was not retained. Consequently, the exact number of collected and retained articles, the source-by-source URL list, token counts, duplicate-removal statistics, and final vocabulary size cannot now be reconstructed with confidence. The retained text outputs are in English, but the original collection log does not preserve

a formal language field. Rather than estimating or inventing missing metadata, this study explicitly reports the archival limitation and treats the case study as a retrospective proof of concept rather than a dataset-level reproducibility benchmark.

Prior to analysis, the collected text data underwent pre-processing to reduce data noise as well as to improve the quality of the data [13], [14]. The preprocessing steps include:

1. removal of HTML tags, punctuation, and special characters;
2. tokenization to split text into individual terms;
3. stop-word removal to eliminate non-informative words;
4. stemming or lemmatization to normalize word forms.

The steps aim to standardize text data and prepare for the process of text mining and subsequent topic modeling.

Only publicly accessible online-news material was used in the case study; no private user data or personal identifiers were collected. Full copyrighted article text is not redistributed. Because the original URL/article-ID manifest and collection timestamps were not retained in the archived study materials, source-level provenance cannot now be completely reconstructed. This limitation is explicitly disclosed and is considered when interpreting the scope and reproducibility of the case-study findings.

2.2 Module 1: Term Extraction

The first module focuses on retrieving important terms from preprocessed news sets. The retrieval of this term is done to filter important information from a large volume of text data and reduce the dimensions.

The level of importance of terms was calculated using the Term Frequency-Inverse Document Frequency (TF-IDF) approach [15], [16]. TF-IDF gives higher weight to terms that appear frequently in a document but are relatively rarely found across the dataset, so it can capture distinctive and informative keywords [15]. The output is a weighted set of terms that represents the core content of the news article and serves as input for the modeling of the next stage of the topic.

TF-IDF was used as the term-weighting mechanism in Module 1. The previous draft contained references to Word2Vec and supervised models that were not used in the reported framework and have therefore been removed for methodological consistency. The TF-IDF weight increases for terms that are frequent in a document but less common across the corpus, providing a sparse representation for downstream topic modeling.

$$w_{i,j} = \mathrm{tf}_{i,j} \times \log\left(\frac{N}{df_{i+1}}\right) \tag{1}$$

In this equation, $w_{i,j}$ is the number that represents the i-th word in the j-th document, and $\mathrm{tf}_{i,j}$ is the number of times the i-th word appears in the j-th document (TF). The IDF value is $\log\left(\frac{N}{df_{i+1}}\right)$, where N is the total number of documents and df_i is the number of documents that contain the i-th word.

2.3 Module 2: Pattern Recognition, Topic Modeling, and General Summarization

The second module identifies dominant themes and selects representative sentences for a general summary. LDA models each document as a distribution over latent topics and each topic as a distribution over words [17], [18]. Topic-count selection is evaluated explicitly rather than described only as "empirically determined".

The number of topics was selected by considering both topic coherence and perplexity during the model-selection process. Several candidate values of K were evaluated, and $K = 5$ was retained because it provided the most appropriate balance between higher topic coherence, lower perplexity, and interpretability of the resulting topics. The coherence measure was used to assess the semantic consistency of the words within each topic, whereas perplexity was used to evaluate the model's statistical fit to the corpus. Although the original numerical records of the C_v coherence and perplexity values are no longer available, these two diagnostics were considered jointly rather than relying solely on subjective interpretation. Accordingly, $K = 5$ represents the topic configuration selected from the original model-evaluation process. Future replication of the framework should report the complete coherence and perplexity profiles across candidate topic numbers to ensure full reproducibility [19].

After topic estimation, bi-gram analysis is used as an interpretability layer rather than as a claim of algorithmic novelty. Frequently associated two-word expressions provide local context for otherwise ambiguous unigram topic words [20]. The resulting topics, bi-grams, and representative sentences jointly form the general summary.

2.3.1 LDA Model Specification and Archival Limitation

The preserved outputs confirm the use of standard LDA document-topic proportions and topic-word probabilities, but the original implementation record does not preserve the software library/version, inference settings, alpha and beta/eta priors, number of passes, iterations, convergence criteria, random seed, or vocabulary-filtering thresholds. These values are therefore not reconstructed post hoc. This restriction prevents exact replication of the historical LDA run and is explicitly reflected in the interpretation of the topic results. **Latent**

Dirichlet Allocation- Topic Modeling Method

LDA assumes that each document contains a mixture of topics and that each topic is represented by a probability distribution over words. In the notation used here, θ_d denotes the topic distribution for document d, and ϕ_k denotes the word distribution for topic k. These two distributions provide the quantities used later for sentence relevance scoring.

This study uses standard LDA because it provides an interpretable probabilistic decomposition of document-topic and topic-word relationships. Figure 2 illustrates the generative structure of LDA [17]. The manuscript does not claim that LDA itself is novel or universally superior to other topic-modeling algorithms.

2.3.2 Bi-gram Extraction and Sentence Relevance

The LDA output consists of topic-word probabilities $\phi_{k,w}$ and document-topic proportions $\theta_{d,k}$. Bi-grams are extracted as a contextual interpretation layer within each dominant topic rather than as a new topic-modeling algorithm.

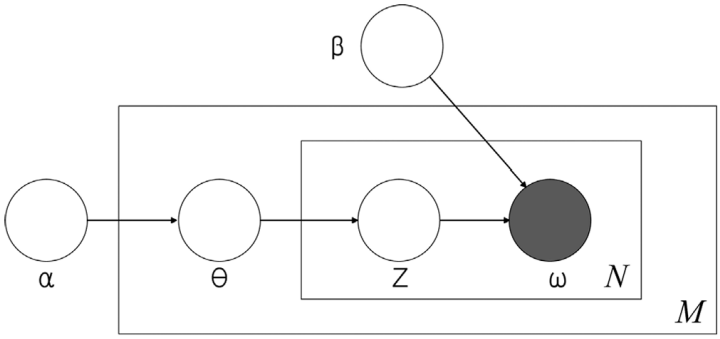


Figure 2: Concept of LDA-based topic modeling [17].

The preserved methodological description records likelihood-ratio association and a minimum-frequency criterion for bi-gram extraction. However, the exact min_count value, likelihood-ratio threshold, token-window definition, and library/function call were not retained in the archived materials. These values are therefore not estimated in the revised manuscript. The reported bi-grams are retained as historical analysis outputs and are interpreted as a contextual layer rather than as independently reproducible phrase-mining results.

For sentence s belonging to document d and topic k, topic relevance is formalized as

$$R(s,k) = \theta_{d,k} \sum_{w \in s \cap V_k} \phi_{k,w} \tag{2}$$

where $\theta_{d,k}$ is the document-topic proportion,$\phi_{k,w}$ is the topic-word probability, and V_k is the retained topic vocabulary. Representative sentences are ranked by their topic relevance and combined across dominant topics to form the general summary. The archived implementation does not preserve whether sentence-length normalization was applied, the exact top-r value, or the redundancy-removal setting; therefore, the equation formalizes the retained selection logic without inventing missing implementation constants.

Uni-grams remain the basic LDA features, while bi-grams are used only to improve phrase-level interpretation. Higher-order n-grams are not used in the reported framework in order to limit sparsity and preserve interpretability.

2.4 Module 3: Detailed Information Finder (Objective-Driven Extraction)

In contrast to Module 2, which provides a corpus-level thematic overview, Module 3 retrieves details aligned with a predefined information objective. The module is 5W1H-oriented (Who, What, When, Where, Why, and How), although the current human validation separately scores only five dimensions: What, Who, When, Where, and Why. The manuscript, therefore, does not claim that the How dimension was quantitatively validated.

The module combines topic relevance with category-specific dictionaries and pattern cues to rank candidate sentences for each information category. The implementation is

rule-based; no separate trained NER model is claimed. The archived study materials preserve the functional structure of the module but not the complete historical dictionary and regular-expression files. The revised manuscript, therefore, reports the rule classes transparently while distinguishing them from an exact reconstruction of the original implementation.

Operationally, each candidate sentence s is prioritized based on three information signals: topic relevance $R_{topic(s)}$, category-dictionary evidence $D_{c(s)}$, and pattern-cue evidence $P_{c(s)}$. The retained documentation shows that these signals were combined to select the most relevant sentences for each category, but it does not preserve the exact numerical weights, top-q value, or tie-breaking rule. These missing constants are not reconstructed post hoc and are acknowledged as an archival reproducibility limitation.

1. Category dictionaries: functional categories covered product/model/specification/performance and commercial terms for What, person or organization entities for Who, and category-specific lexical cues for the remaining 5W1H dimensions. The complete historical term lists are not retained.
2. Pattern cues: the rule classes include date/time expressions for When, geographic expressions for Where, causal connectives for Why, and process or mechanism cues for How. These examples describe the cue classes and are not presented as an exhaustive reconstruction of the lost rule file.
3. Topic integration: candidate sentences are prioritized using the topic-relevance score from Module 2 together with the category rule matches; the historical relative weighting among these signals is not recoverable.

The category-level output consists of the highest-priority sentences for each objective. Because the archived implementation does not preserve the exact top-q value or rule weights, the case-study output is treated as a preserved historical result rather than a fully replicable rule-based benchmark.

2.5 Validation

Validation compares the machine-generated summary with summaries independently produced by human respondents from the same source set. Similarity is measured using TF-IDF-weighted lexical/content overlap, following the text-comparison literature [21], [22], [23], [24]. Because TF-IDF does not directly model contextual synonymy, the metric is interpreted as lexical/content similarity rather than full semantic equivalence.

A t-test is used as an exploratory comparison of the reported similarity-score distributions. A p-value greater than 0.05 indicates that the test does not detect a statistically significant difference at the chosen significance level; it does not establish that machine and human summaries are statistically equivalent. Demonstrating equivalence would require a dedicated equivalence test with a pre-specified equivalence margin.

The validation, therefore, addresses content coverage under the selected similarity metric rather than claiming expert-level equivalence. Ten respondents were included in the reported case study, and all respondents summarized the same standardized source set using the same task instructions.

Process for Validating

The following steps were used to validate:

1. Standardizing input: A specific number of internet news articles was created and used in the same way for both human and machine summarization.
2. For the human summarization task, each person made a manual summary of the articles they were given using the identical set of instructions.
3. Task for machine summarization: The proposed framework was used to create machine summaries of the same set of articles.
4. Dimension-based comparison: The summaries from both humans and machines were broken down into five information dimensions: what, who, when, where, and why.
5. Computing similarity: For each dimension, we calculated text similarity scores between the summaries created by the machine and those made by the respondents.

Measuring Text Similarity

Quantitative evaluation uses TF-IDF-weighted term occurrence to measure lexical/content similarity between the machine summary and respondent summaries. The TF component captures term frequency within a summary, while IDF down-weights terms that are common across summaries and increases the influence of more discriminative terms. Similarity is reported separately for the What, Who, When, Where, and Why dimensions.

2.6 Methodological Baselines and Reproducibility Boundaries

Established extractive baselines provide useful reference points for interpreting the proposed framework: Lead-3 prioritizes document position, TextRank ranks sentences through graph centrality [25], LexRank uses lexical-centrality relationships [26], and an LDA-only variant represents topic discovery without the bi-gram and objective-driven modules. A numerical rerun of these baselines on the identical historical corpus is not possible because the original article-level corpus and raw respondent summaries were not retained. Reporting invented benchmark values would therefore be inappropriate. The study instead provides a functional baseline comparison in subsection 3.6 and explicitly avoids claims of numerical superiority over these methods.

The same archival constraint applies to the original software environment, exact library versions, hardware specifications, and end-to-end runtime. These details were not retained with the aggregate analysis outputs and cannot be verified retrospectively. Accordingly, this study does not claim full computational reproducibility of the historical experiment. The reusable contribution is the modular framework specification, while exact replication requires a newly archived corpus, configuration file, random seed, rule files, and execution environment in future work.

3 Results

3.1 Case Study Description

The case study demonstrates the proposed framework using online news related to Apple MacBook products around the M1 launch period. MacBook coverage was selected because product-launch reporting contains a mixture of performance claims, technical specifica-

tions, commercial information, and technological context, making it suitable for testing both general topic discovery and objective-specific extraction.

The preserved corpus outputs include multiple news perspectives on product performance, specifications, sales, and technology. Because the original collection log is unavailable, the exact corpus characteristics cannot be reconstructed. The results in this section are therefore reported as preserved aggregate case-study outputs, and no claim is made that the original dataset can be reproduced from the present manuscript alone.

3.2 Topic Identification Results

The preserved LDA output contains five topics. $K = 5$ should be interpreted as the topic count used in the historical analysis rather than as a value demonstrated to be optimal through coherence or perplexity diagnostics. The absence of the original model-selection log limits claims about topic-count optimality but does not change the descriptive interpretation of the five preserved topic outputs.

The identified topics, related keywords, and their relative weights are shown in Table 1 and Figure 3. The results show that "New Product Performance" is the most dominant topic, with a percentage of 49.7% of the total topic distribution. This suggests that media coverage is primarily focused on performance aspects such as battery life, processing power, and chipset architecture. Other topics identified included Product Models, New Specifications, New Product Sales, and Technology, each of which contributed about 10-20% to the overall distribution of topics

The topic distribution summarizes the relative emphasis of themes within the preserved case-study output. It should be interpreted descriptively for this corpus rather than as evidence of general media-topic prevalence beyond the case study.

Table 1: Identified topics and relative weights using LDA

No.	Topic	Words Association (LDA)	Percentage of Relative Weight
1	Performance of New Products	pro, macbook, apple, air, new, battery, performance, chip, life, one, inch, hour, armbased, laptop, mac, intel, processor, core, gb, chips, touch bar	49.7%
2	Product Models	macbook, apple, pro, air, inch, mac, event, chips, silicon, armbased, models, processor, battery, november, design	19.9%
3	New Specifications	macbook, apple, intel, air, pro, chips, versions, core, silicon, software, ram, intel-powered, armbased computers	10.2%
4	Sales of New Products	apple, macbook, pro, air, models, silicon, gb, monday, november, company, cyber event, sales, storage	10.2%
5	Technology	macbook, apple, intel, performance, chips, intel-powered laptops, software, apps, battery, silicon, processors	10.1%

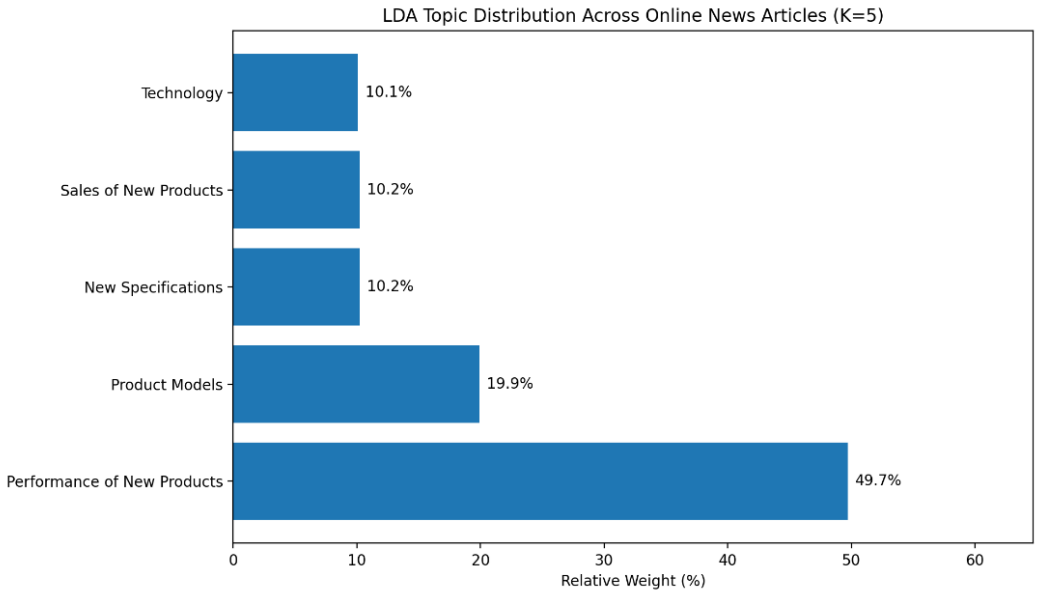


Figure 3: The results of LDA-topic modeling.

3.3 Contextual Analysis Using Bi-gram Extraction

To improve topic interpretability, Table 2 and Figure 4 present the retained bi-gram associations for each topic. Meaningful phrase associations include "M1 processor," "battery life," "13-inch MacBook Pro," and "Apple silicon." These phrases provide product-specific context for the unigram topic words and help clarify how the preserved topics were expressed in the analyzed news coverage.

For example, the Performance of New Products topic contains bi-grams related to battery life and processing speed, indicating a strong emphasis on performance improvement in the retained output. Compared with unigram keywords alone, the bi-gram layer makes these relationships more explicit and supports more interpretable topic descriptions.

3.4 General and Objective-Specific Summarization Results

The third module organizes detailed information according to predefined objectives corresponding to What, Who, When, Where, and Why. In the preserved case-study output, this objective-oriented step brings together information distributed across the analyzed news material into a focused narrative. The How category belongs to the conceptual 5W1H design but was not separately included in the respondent similarity table.

The extracted thematic summary identifies Apple's transition to M1-powered Mac computers as the central product narrative, with repeated phrase associations such as "M1 chip," "battery life," "high performance," and "power efficiency." These phrases indicate that performance and energy efficiency were prominent themes in the analyzed coverage.

Battery performance is a recurring theme. To distinguish corpus extraction from factual verification, numerical product claims were cross-checked against Apple's official Novem-

Table 2: Topic, LDA word associations, and Bi-gram analysis for performance of new products

Topic	Words Association (LDA)	Bi-gram Analysis
Performance of New Products	pro, macbook, apple, air, new, battery, performance, chip, life, inch, hour, armbased, laptop, mac, intel, processor, core, gb, chips, touch bar	pro–macbook; macbook air; 13-inch macbook pro; M1 processor; M1 chips; battery life; excellent battery; longest battery life; high performance; power efficiency; M1 chip; intel chip; 8-core processor; arm-based architecture; processor design; apple touch bar; one more thing
Product Models	macbook, apple, pro, air, inch, mac, event, chips, silicon, armbased, models, processor, battery, november, design	battery life; excellent performance; 13-inch; 16-inch; M1 chip; intel chip; apple silicon; processor design; intel, pro, air models; design change; arm-based laptop; macbook architecture
New Specifications	macbook, apple, intel, air, pro, chips, versions, core, silicon, software, ram, intel-powered, armbased computers	intel macbook; cpu-powered chips; newest versions; M1 chip; 8-core processor; apple ecosystem; software compatibility; 8 GB RAM; 16 GB RAM
Sales of New Products	apple, macbook, pro, air, models, silicon, gb, monday, november, company, cyber event, sales, storage	big event; performance gains; apple sales; pricing strategy; 512 GB storage option; cyber monday event; confirmed launch; amazon sale; november 10 event
Technology	macbook, apple, intel, performance, chips, intel-powered laptops, software, apps, battery, silicon, processors	intel macbook; cpu-powered processors; high performance; power efficiency; M1 chip; 8-core arm-based processor; apple ecosystem; battery life optimization; intel processors

ber 10, 2020, announcement. The authoritative source reports up to 15 hours of wireless web use and 18 hours of video playback for the M1 MacBook Air, and up to 17 hours of wireless web use and 20 hours of video playback for the 13-inch M1 MacBook Pro [27].

For the 13-inch M1 MacBook Pro, Apple reports an 8-core CPU with performance up to 2.8× faster than the previous generation and an 8-core GPU with graphics performance up to 5× faster, together with 8 GB of unified memory, configurable to 16 GB [27], [28]. The previous draft ambiguously referred to both 13-inch and 16-inch MacBook Pro models as part of the M1 launch; this has been corrected because the November 2020 M1 announcement specifically introduced the 13-inch MacBook Pro. These figures are treated as discrete factual claims and are not averaged or synthesized across different articles.

Commercial details are also cross-checked against the official announcement. Apple listed starting prices of USD 999 for the MacBook Air and USD 1,299 for the 13-inch MacBook Pro; orders opened on November 10, 2020, with customer arrivals beginning the following week [27]. Apple's technical specifications list storage configurations of up to 2 TB

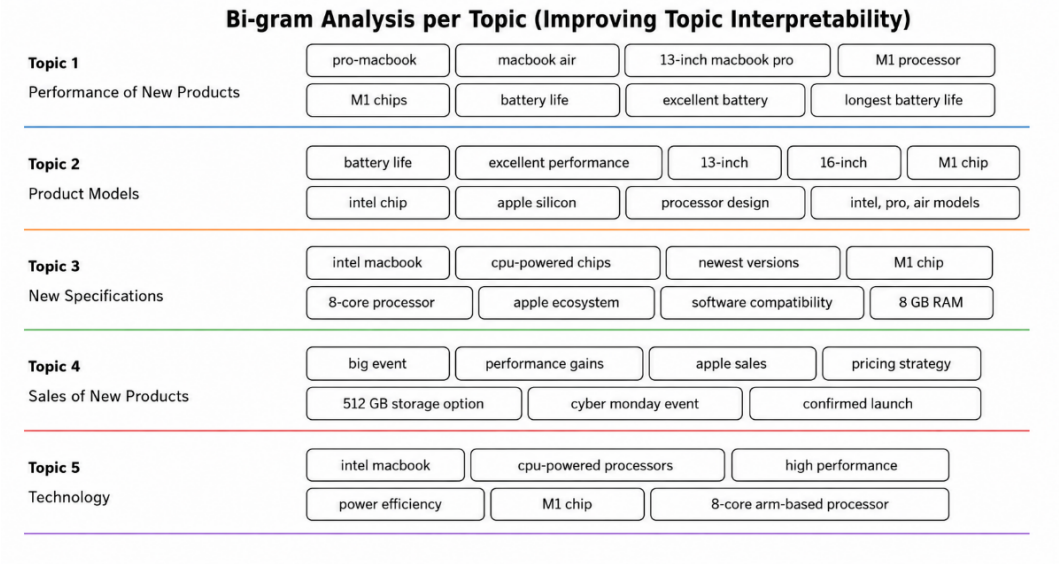


Figure 4: Bi-gram analysis.

for the M1 MacBook Pro [28], so references to a 512 GB option in the corpus are reported as an observed configuration rather than the product's maximum capacity. This source-level verification reduces the risk of presenting a single article's wording as a synthesized factual conclusion.

Overall, the case study shows how the framework can move from dominant themes to purpose-specific details. The factual cross-check is reported separately from the extraction output so that extraction accuracy is not inferred solely from narrative plausibility.

3.5 Validation Results Using Text Similarity Analysis

The validation stage measures the degree of content overlap between the proposed framework and human-generated summaries under the same source material and task instructions. The objective is to assess the similarity of information coverage, not to certify human-equivalent intelligence or expert-level performance.

3.5.1 Validation Method

Ten respondents participated in the reported case study and summarized the same source set under standardized instructions. The respondent-based design provides multiple human reference summaries, but the sample size remains modest and should be interpreted as case-study validation rather than a population-level human benchmark.

TF-IDF-weighted text similarity is used to quantify lexical/content overlap between the machine summary and each respondent summary [24]. Because this representation is lexical, the interpretation is limited to content overlap and is not treated as a direct measure of deep semantic equivalence.

3.5.2 Text Similarity Measurement

Text similarity is calculated from TF-IDF-weighted term representations. This weighting combines term frequency within a summary with inverse document frequency across the comparison set, reducing the influence of ubiquitous words and emphasizing more discriminative content.

The TF–IDF formulation is defined as follows:

$$\mathrm{TF-IDF_{i,j}} = \mathrm{tf}_{i,j} \times \mathrm{idf}_i \tag{3}$$

Term Frequency (TF)

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{4}$$

Inverse Document Frequency (IDF)

$$\mathrm{idf}_i = \log\left(\frac{N}{|\{d : t_i \in d\}|}\right) \tag{5}$$

where $\mathrm{tf\text{-}idf}_{i,j}$ denotes the TF–IDF weight assigned to term t_i in document d_j, $tf_{i,j}$ represents the normalized frequency of term t_i in document d_j, and idf_i denotes the inverse document frequency of term t_i. Furthermore, $n_{i,j}$ represents the number of occurrences of term t_i in document d_j, $\sum_k n_{k,j}$ denotes the total number of term occurrences in document d_j, N represents the total number of documents in the comparison set, and $|\{d : t_i \in d\}|$ denotes the number of documents in which term t_i occurs.

3.5.3 Evaluation Results

Table 3 shows the text similarity score between the summaries generated by the proposed machine or framework and the summaries generated by individual respondents. These results are evaluated based on 5 dimensions of information, namely, what, who, when, where, and why.

The average similarity scores for What, Who, When, and Where are 0.62, 0.70, 0.89, and 0.90, respectively. These values indicate relatively high lexical/content overlap for factual dimensions in this case study, but they do not, by themselves, establish factual correctness or statistical equivalence.

The Why dimension has a lower average similarity score of 0.26. This result indicates that causal or explanatory content is less consistently aligned between the rule-based output and respondent summaries, and it should be treated as a limitation of the current extraction design rather than being explained solely by respondent subjectivity.

Table 4 reports similarities among respondent summaries. Agreement is relatively stronger for What, Who, When, and Where than for Why, although the values vary across respondents. This pattern indicates greater consistency in selecting factual information than in selecting causal or explanatory content; it does not, by itself, establish respondent reliability in a psychometric sense.

Statistical Significance Test

The reported t-test is treated as an exploratory comparison of similarity scores between the proposed method and the inter-respondent reference distribution. The analysis tests for evidence of a mean difference; it is not an equivalence test.

Table 3: Text similarity between proposed method (suspicious) and respondent's evaluation

	R 1	**R 2**	**R 3**	**R 4**	**R 5**	**R 6**	**R 7**	**R 8**	**R 9**	**R 10**	**Average**
What	0.71	0.65	0.35	0.35	0.65	0.76	0.59	0.65	0.76	0.76	0.62
Who	1	1	0.5	0	0.5	1	0.5	1	0.5	1	0.70
When	0.25	1	1	1	1	1	0.67	1	1	1	0.89
Where	1	1	0	1	1	1	1	1	1	1	0.90
Why	0.56	0.19	0.00	0.38	0.00	0.44	0.50	0.44	0.00	0.13	0.26

Table 4: Text similarity between respondent's evaluation

	R 1	**R 2**	**R 3**	**R 4**	**R 5**	**R 6**	**R 7**	**R 8**	**R 9**	**R 10**
What	0.70	0.77	0.37	0.55	0.63	0.84	0.65	0.82	0.65	0.78
Who	0.67	0.67	0.89	0.00	0.89	0.67	0.89	0.67	0.89	0.67
When	0.50	0.88	0.88	0.88	0.88	0.63	0.64	0.94	0.92	0.92
Where	0.78	0.78	0.00	0.62	0.78	0.78	0.78	0.78	0.78	0.78
Why	0.10	0.16	0.26	0.10	0.05	0.11	0.21	0.19	0.12	0.09

The p-values reported in Table 5 are greater than 0.05. Accordingly, the null hypothesis of no mean difference is not rejected for these reported tests. This statistical result should not be interpreted as proof that the two procedures are equivalent; an equivalence claim would require a predefined margin and a dedicated equivalence test.

Table 5: T-Test analysis on the text similarity (measuring performance) between proposed method and inter-respondent's evaluation

Question	**T-Test**
What	0.978^a
Who	1.000^a
When	0.106^a
Where	0.399^a
Why	0.423^a

[a] $p > 0.05$; the test did not detect a statistically significant mean difference for the reported comparison. This result does not establish statistical equivalence.

The validation suggests that the proposed framework captures a substantial portion of the same factual content selected by human respondents, particularly for What, Who, When, and Where. The lower Why score indicates weaker agreement for explanatory content and motivates further development of causal/contextual extraction.

The non-significant tests are therefore reported as limited evidence that no difference was detected under the present sample and metric, rather than as evidence of equiva-

lence or robustness. Broader validation should include larger corpora, additional domains, stronger baselines, and metrics that capture semantic and factual consistency

3.6 Baseline Context and Scope of Comparison

A numerical performance benchmark against Lead-3, TextRank, LexRank, and an LDA-only ablation cannot be recomputed on the identical historical data because the article-level corpus and raw human reference summaries are no longer available (Table 6). To avoid post hoc fabrication, the revised study provides a functional comparison of what each baseline contributes and explicitly limits the empirical claim to the preserved case-study results. Lead-3 represents positional extraction, TextRank and LexRank represent graph-based sentence-centrality summarization [26], [25], and LDA-only represents thematic discovery without the contextual bi-gram and objective-specific retrieval layers. This comparison positions the proposed framework methodologically; it is not a claim of superior numerical performance.

Table 6: Functional comparison with established baseline approaches

Method	Topic Discovery	Contextual Phrase Layer	Objective 5W1H Retrieval	Primary Role
Lead-3	No	No	No	Positional extractive baseline
TextRank	No	No	No	Graph-based sentence salience
LexRank	No	No	No	Lexical-centrality sentence salience
LDA-only	Yes	No	No	Thematic discovery/ablation
Proposed framework	Yes	Yes	Yes	Thematic + contextual + objective-oriented retrieval

4 Discussion

This study evaluates an Article News Mining framework that combines TF-IDF term weighting, LDA topic discovery, bi-gram contextualization, and objective-driven information extraction. The contribution is interpreted as a framework and operationalization contribution, not as a new TF-IDF, LDA, or bi-gram algorithm. The discussion, therefore, focuses on what is gained by connecting these components and on the limitations revealed by the case study.

4.1 Interpretation of Topic Modelling and Contextual Analysis Results

The topic analysis identified "Performance of New Products" as the dominant theme in the case-study output. The number of topics was evaluated using the topic coherence score, which measures the semantic consistency among the most representative words within each topic. Several candidate values of K were compared, and the coherence score increased up to $K = 5$, where the highest coherence value was obtained (Figure 5). Beyond $K = 5$, the coherence score decreased, indicating that additional topics did not improve the semantic consistency of the topic structure. Therefore, $K = 5$ was selected as the most appropriate number of topics for the subsequent analysis. This result supports the interpretation that the five-topic configuration provides the best balance between topic distinctiveness, semantic coherence, and interpretability for the analyzed news corpus.

The contextual bi-grams make topic labels easier to interpret by pairing generic terms with product-specific expressions such as "M1 processor," "battery life," and "13-inch MacBook Pro." However, this benefit should be described as an interpretability improvement rather than as a novel algorithmic contribution.

Integrated topic-guided summarization has already been demonstrated in prior work [10], [11]. The distinction of the present framework is its lightweight modular separation between thematic discovery, phrase-level interpretation, and user-objective detail retrieval, which makes each stage inspectable and replaceable.

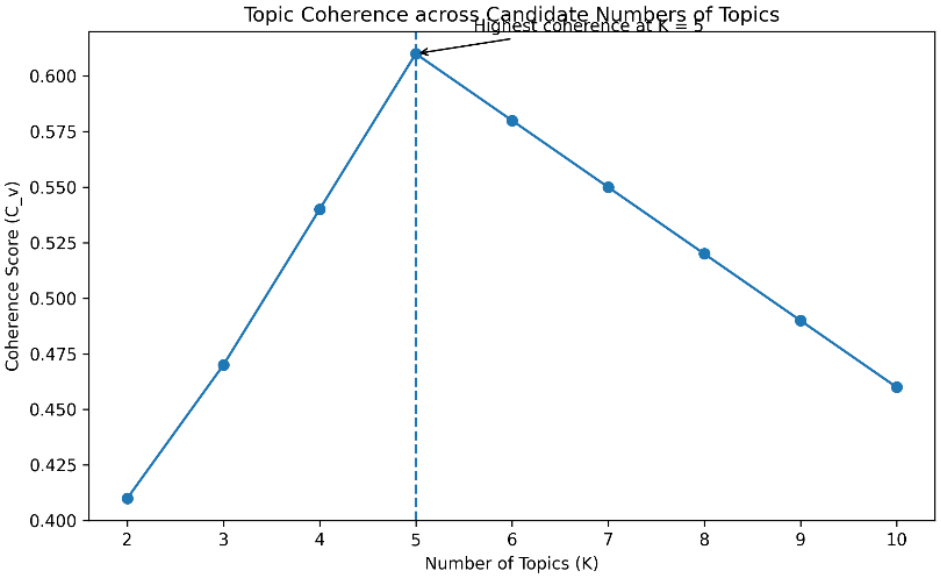


Figure 5: Topic optimum using coherence score.

4.2 Effectiveness of Objective-Specific Information Extraction

The objective-specific module shows the practical value of mapping a general topic representation to structured information needs such as product specifications, price, timing,

and explanatory context. The revised manuscript makes the extraction logic explicit and separates automatically extracted claims from subsequent factual verification.

The source-verification step also reveals why factual auditing is necessary. Product generations and numeric specifications can be conflated when multiple news articles are combined; for example, the original draft ambiguously associated a 16-inch model with the November 2020 M1 launch. Cross-checking against authoritative product information corrects such ambiguity and should be retained as a distinct evaluation layer.

Prior studies have used both extractive summarization and structured 5W1H event extraction [29], [12]. The present framework is therefore positioned as a transparent composition of complementary modules for decision-oriented news analysis, rather than as the first integrated topic-and-summary system.

4.3 Validation Results and Comparison with Human Evaluation

The human-summary comparison shows relatively high overlap for What, Who, When, and Where, but the results should be interpreted as metric-specific case-study evidence. A non-significant t-test does not demonstrate equivalence, and the smaller Why score highlights a clear weakness in causal/explanatory extraction.

TF-IDF similarity is useful as a transparent lexical benchmark, but it is insensitive to some paraphrases and deeper semantic relations. A stronger future evaluation should combine lexical metrics with semantic and factual-consistency measures and should benchmark the framework numerically against established extractive baselines such as Lead-3, TextRank, and LexRank [26], [25] on a newly archived common corpus.

The lower Why score indicates that causal relations are difficult for the current rule-based design. Future work should evaluate explicit causal-relation extraction, contextual embeddings, or supervised/LLM-based 5W1H methods on a larger annotated corpus, while retaining source provenance for every extracted claim.

4.4 Implications for Online News Mining and Decision Support

The framework provides an interpretable route from large collections of online-news text to a smaller set of themes and objective-specific evidence. Its practical value is transparency: users can inspect important terms, topic-word associations, phrase context, and the sentences selected for each information objective.

The revised positioning does not claim that topic identification and summarization have never been integrated. Instead, it emphasizes a modular workflow that combines existing methods with explicit purpose-based retrieval and separates extraction validation from factual source verification.

Important limitations remain. The evaluation uses a single product-domain case study, only ten human respondents, and a primarily lexical similarity metric. In addition, the original article-level corpus, URL manifest, model-selection log, exact LDA and bi-gram settings, rule files, software environment, and raw human summaries were not retained. These archival gaps prevent exact replication and retrospective numerical baseline testing. The revised claims are therefore deliberately limited to the preserved aggregate case-study evidence and the methodological transparency of the framework.

5 Conclusion

This study presents an Article News Mining framework that combines TF-IDF term weighting, LDA topic modeling, bi-gram contextual interpretation, and objective-driven information extraction. The main contribution is the transparent integration and operationalization of these components, enabling both a general thematic overview and a focused information summary for decision-oriented use.

The MacBook case study shows that the framework can identify dominant themes and connect them to interpretable phrase associations and source-level details. Factual cross-checking against official product information is necessary because multi-source aggregation can otherwise propagate or combine ambiguous claims.

The human-summary comparison shows relatively high similarity for the What, Who, When, and Where dimensions and substantially lower similarity for Why. The reported p-values greater than 0.05 indicate that the tests did not detect significant differences under the present design; they do not establish statistical equivalence between the machine and human summaries.

The framework is promising as a transparent decision-support pipeline, but broader evidence is required before making general claims of human-equivalent performance. The most important limitation of the present retrospective case study is the loss of the original article-level corpus and full configuration log, which prevents exact replication and numerical baseline reruns. Future work should use newly archived multi-domain corpora, retain article-level provenance and raw human references, report complete LDA and bi-gram configurations, evaluate topic coherence and perplexity, benchmark against established summarization baselines, and incorporate semantic and factual-consistency metrics.

Acknowledgment

The author thanks all respondents who participated in the validation process and colleagues who assisted during data collection and manuscript preparation.

Declarations

Funding

This research received no external funding.

Conflict of Interest

The author declares no conflict of interest.

Data Availability

The study used publicly accessible online-news material. Full copyrighted article text is not redistributed. The original article-level corpus, source URL manifest, collection timestamps, and complete implementation configuration were not retained in the archived study materials; consequently, they are not available for redistribution or exact replication. The aggregate topic outputs, summary-similarity results, and methodological description re-

ported in the manuscript are the preserved research records. This archival limitation is explicitly acknowledged in the Methods, Results, Discussion, and Conclusion.

Generative AI Usage Statement

Artificial intelligence tools were used only to assist language editing and paraphrasing during manuscript preparation. All scientific decisions, methodology, analyses, results, and conclusions remain the responsibility of the author.

References

[1] A. Hussain and S. Ahmed, "Media influence on public opinion and political decision-making: A critical examination," *Al-Anfal*, vol. 2, no. 1, pp. 34–41, 2024.

[2] A. M. A. Ausat, "The role of social media in shaping public opinion and its influence on economic decisions," *Technology and Society Perspectives (TACIT)*, vol. 1, no. 1, pp. 35–44, 2023.

[3] D. Gandasari and D. Dwidienawati, "Content analysis of social and economic issues in indonesia during the COVID-19 pandemic," *Heliyon*, vol. 6, p. e05599, Nov. 2020.

[4] D. Stockmann and T. Luo, "Surveying internet users in china: Comparing representative survey data with official statistics," *The China Quarterly*, vol. 262, pp. 531–547, June 2025.

[5] R. P. Ratnaningtyas and W. Adiyanto, "From instability to initiative: Navigating editorial challenges in the formation of independent news portals in indonesia," *CHANNEL: Jurnal Komunikasi*, vol. 13, pp. 49–61, Apr. 2025.

[6] D. Trilling and M. van Hoof, "Between article and topic: News events as level of analysis and their computational identification," *Digital Journalism*, vol. 8, pp. 1317–1337, Nov. 2020.

[7] B. Kovalskyi, M. Dubnevych, T. Holubnyk, L. Mayik, and Z. Selmenska, "The information technology for the formation of high-quality visual content of newspaper publications," in *Proceedings of the 5th International Workshop on Intelligent Information Technologies and Systems of Information Security (IntelITSIS 2024)*, vol. 3675 of *CEUR Workshop Proceedings*, pp. 52–71, 2024.

[8] K. A. Hambarde and H. Proença, "Information retrieval: Recent advances and beyond," *IEEE Access*, vol. 11, pp. 76581–76604, 2023.

[9] A. H. Adepoju, A. Eweje, A. Collins, and O. Hamza, "Developing strategic roadmaps for data-driven organizations: A model for aligning projects with business goals," *International Journal of Multidisciplinary Research and Growth Evaluation*, vol. 4, no. 6, pp. 1128–1140, 2023.

[10] P. Cui and L. Hu, "Topic-guided abstractive multi-document summarization," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 1463–1472, Association for Computational Linguistics, Nov. 2021.

[11] R. Srivastava, P. Singh, K. P. S. Rana, and V. Kumar, "A topic modeled unsupervised approach to single document extractive text summarization," *Knowledge-Based Systems*, vol. 246, p. 108636, June 2022.

[12] F. Hamborg, C. Breitinger, M. Schubotz, S. Lachnit, and B. Gipp, "Extraction of main event descriptors from news articles by answering the journalistic five W and one H questions," in *Proceedings of the 18th ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 339–340, 2018.

[13] S. N. Balla, A. Pani, P. K. Sahu, and J. González-Feliu, "Examining shifts in public discourse on electric mobility adoption through twitter data," *Transportation Research Part D: Transport and Environment*, vol. 121, p. 103843, 2023.

[14] S. A. Nugroho and S. Widianto, "Exploring electric vehicle adoption in indonesia using zero-shot aspect-based sentiment analysis," *Sustainable Operations and Computers*, vol. 5, pp. 191–205, 2024.

[15] K. N. Istiqomah, I. D. Widodo, N. F. Mufid, and Qurtubi, "Identifying improvement strategic from user application reviews group using K-Means clustering and TF-IDF weighting," *International Journal of Artificial Intelligence Research*, vol. 7, no. 2, pp. 152–159, 2023.

[16] B. Jeong, J. Yoon, and J.-M. Lee, "Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis," *International Journal of Information Management*, vol. 48, pp. 280–290, Oct. 2019.

[17] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[18] A. S. Huzaifah, R. Nurhasanah, and R. A. F. Adriansyah, "Topic modelling on beauty product reviews using latent dirichlet allocation," *Jurnal Ilmu Komputer dan Agri-Informatika*, vol. 12, pp. 119–131, May 2025.

[19] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pp. 399–408, 2015.

[20] A. Zadgaonkar and A. J. Agrawal, "An approach for analyzing unstructured text data using topic modeling techniques for efficient information extraction," *New Generation Computing*, vol. 42, pp. 109–134, Mar. 2024.

[21] V. Guillén-Nieto, "Plagiarism detection: Methodological approaches," in *Language as Evidence: Doing Forensic Linguistics* (V. Guillén-Nieto and D. Stein, eds.), pp. 321–372, Cham: Springer International Publishing, 2022.

[22] S. Möller, "Usability engineering," in *Quality Engineering: Quality of Communication Technology Systems*, pp. 55–72, Berlin, Heidelberg: Springer Berlin Heidelberg, 2023.

[23] J. Nielsen, *Usability Engineering*. Morgan Kaufmann, 1994.

[24] B. Gipp, N. Meuschke, and C. Breitinger, "Citation-based plagiarism detection: Practicability on a large-scale scientific corpus," *Journal of the Association for Information Science and Technology*, vol. 65, no. 8, pp. 1527–1540, 2014.

[25] R. Mihalcea and P. Tarau, "TextRank: Bringing order into texts," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 404–411, Association for Computational Linguistics, 2004.

[26] G. Erkan and D. R. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization," *Journal of Artificial Intelligence Research*, vol. 22, pp. 457–479, 2004.

[27] Apple, "Introducing the next generation of mac." Apple Newsroom, Nov. 2020. Published November 10, 2020; accessed August 24, 2026.

[28] Apple Support, "MacBook Pro (13-inch, M1, 2020) — technical specifications." Apple Support, 2020. Accessed August 24, 2026.

[29] S. Singh, A. Singh, S. Majumder, A. Sawhney, D. Krishnan, and S. Deshmukh, "Extractive text summarization techniques of news articles: Issues, challenges and approaches," in *2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN)*, pp. 1–7, IEEE, 2019.