Science in Information Technology Letters
Vol. 3., No. 2, November 2022, pp. 116-132
ISSN 2722-4139
http://pubs2.ascee.org/index.php/sitech

Advanced product review summarization in e-commerce marketplaces: elevating beyond tf-idf and lexrank method

Rita Melina Anggraeni [a,1], Dewi Pramudi Ismi [a,2,*]

[a] Informatics Department, Universitas Ahmad Dahlan, Indonesia
[1] rita1600018226@webmail.uad.ac.id; [2] dewi.ismi@tif.uad.ac.id
* Corresponding Author

ARTICLE INFO

Article history
Received 12 October, 2022
Revised 28 October, 2022
Accepted 14 November, 2022

Keywords
Text summarization
Review summarization
LexRank method
E-commerce marketplace
ROUGE evaluation

ABSTRACT

In the fiercely competitive domain of online product sales, wherein engendering trust among prospective buyers assumes paramount significance, the role of product reviews cannot be understated. However, a prevailing issue in online marketplaces resides in the presence of product reviews that do not consistently align with the overall product rating. Furthermore, the sheer abundance of comments often leads potential consumers to confine their scrutiny to the initial comments, thus leaving a substantial volume of reviews unexplored. To rectify this challenge, this study introduces an automated text summarization system for product reviews, leveraging the LexRank methodology. This system underwent rigorous evaluation using the Rouge metric, with results manifesting substantial promise. At a threshold of 0.1, Rouge-1 exhibited an accuracy of 16.67%, while Rouge-2 scored 3.01%, and Rouge-L reached 16.50%. At a threshold of 0.2, Rouge-1 yielded a score of 16.08%, Rouge-2 registered 2.64%, and Rouge-L scored 16.57%. The second evaluation, performed with a distinct test dataset, notably excelled, emphasizing the system's competence. Specifically, at the 0.2 threshold, the system displayed superior performance, underscoring its efficacy in refining product review summarization within online marketplaces.

1. Introduction

The contemporary technological landscape has precipitated a discernible paradigm shift in consumer behavior [1], notably transitioning from conventional brick-and-mortar commerce to the realm of digital transactions, chiefly channeled through online marketplaces [2], [3]. These marketplaces, prominently anchored in web-based Internet platforms, serve as multifaceted arenas for diverse commercial endeavors [4], empowering consumers to meticulously discern and select suppliers in alignment with their precise criteria, thereby affording them access to products at prices reflective of the prevailing market dynamics [5], [6]. In parallel, these platforms furnish purveyors and suppliers with the valuable capability to ascertain enterprises actively seeking their products and services [7], [8].

In the backdrop of a burgeoning and intensely competitive online product sales arena [9], the cultivation of trust among prospective buyers stands as an imperative objective [10]. The critical role of product reviews as a determinant in purchase decisions has concomitantly surged [11], [12]. A profusion of positive reviews augments the likelihood of product acquisition, while conversely, an influx of adverse reviews can deter potential buyers [13]. Nonetheless, within the expansive expanse of product reviews on online marketplaces, a pervasive quandary surfaces - the incongruence between reviews and the aggregate product rating [14]. This predicament is compounded by the sheer profusion of comments that, often, inundates consumers, prompting them to restrict their perusal to the preliminary comments on a product's review page. This myopic approach, however, inadvertently relegates a substantial corpus of valuable insights to obscurity, beyond the purview of prospective purchasers [15], [16].

In response to this predicament, this research aspires to introduce a methodical resolution in the form of an automated text summarization system. Specifically, it harnesses the formidable LexRank methodology [17]–[20], offering the potential to revolutionize the presentation and consumption of product reviews in online marketplaces. The method encompasses an array of rudimentary processes, encompassing the standardization of terminological conventions, the meticulous preprocessing of text data, the judicious employment of term frequency-inverse document frequency (TF-IDF) weighting [21], [22] to evaluate word salience, and the adept application of the LexRank algorithm [23], [24]. Subsequently, the system's efficacy undergoes rigorous evaluation utilizing the ROUGE metric [25], [26], a robust analytical instrument for assessing the quality and cohesiveness of the resultant summaries.

Throughout the expanse of this article, we shall endeavor to traverse the intricacies of our research, presenting a meticulous exploration of the system's inception, implementation, and appraisal in Section 2. Subsequent sections shall elucidate the methodology, expound upon empirical findings and engage in detailed discussions in Section 3. and culminate with comprehensive conclusions in Section 4. The overarching aim is to cast illumination upon LexRank's profound potential to elevate the art of product review summarization within the dynamically expanding ambit of e-commerce marketplaces.

2. Method

The research methodology adopted in this study encompasses several intricate stages, as delineated in Fig. 1, which offers a visual representation of the sequential phases involved in this investigation.

2.1. Data Mining, Text Mining and Natural Language Processing

The research methodology in this study encompasses a comprehensive approach, integrating data mining, text mining, and natural language processing (NLP) to extract valuable insights from a vast corpus of product reviews within the e-commerce domain [27]–[29]. These interconnected phases are depicted in Fig. 1, presenting a holistic view of the research framework.

Data mining, often referred to as knowledge discovery, serves as the foundational phase for knowledge extraction [30]. It involves the meticulous process of unearthing meaningful patterns, associations, and predictive trends from extensive datasets. In the context of this research, data mining is instrumental in sourcing and structuring unstructured text data from product reviews, paving the way for subsequent text mining and NLP processes. It encompasses several critical components :

- Knowledge Discovery: Data mining aims to extract actionable knowledge from data, delving deeper than conventional data analysis to uncover hidden patterns.

- Unstructured Data Processing: It deals with diverse data types, including unstructured text data, demanding advanced techniques for transformation and structuring.
- Pattern Identification: Data mining employs advanced algorithms to identify patterns within the data, enabling the detection of associations and classifications.
- Predictive Modeling: Forecasting trends and consumer behaviors is a key objective, achieved through predictive modeling techniques.
- Feature Selection: The selection of pertinent features from the dataset reduces dimensionality, enhancing downstream analysis.
- Ethical Considerations: Ensuring ethical data handling and privacy protection are paramount in data mining.

Text mining is the subsequent stage, focusing on the automated analysis of textual content, particularly product reviews [31], [32]. Its role is critical in extracting information and sentiments from unstructured text. Key facets of text mining encompass :

- Automated Text Analysis: Leveraging algorithms for automated information extraction from text, especially in the context of product reviews.
- Sentiment Analysis: Determining sentiment expressed in reviews, whether positive, negative, or neutral, is fundamental to understanding consumer opinions.
- Keyword Identification: Identifying and ranking keywords in reviews helps in discerning the most critical product attributes.
- Topic Modeling: Analyzing themes and topics prevalent in reviews reveals the most-discussed aspects.
- Scalability and Efficiency: Handling large volumes of text data efficiently is essential in text mining.
- Ethical Considerations: Ensuring ethical handling of textual data, including privacy concerns, is imperative.

Natural Language Processing (NLP) is an integral part of this research, extending to the analysis of textual content [33]. NLP focuses on understanding and manipulating human language, enabling further in-depth analysis of product reviews :

- Language Understanding: NLP parses sentences, identifies parts of speech, and handles linguistic nuances, enhancing text mining.
- Multilingual Analysis: NLP should be adaptable to multiple languages, crucial in the context of multilingual product reviews.
- Cross-Document Analysis: NLP enables the examination of relationships and patterns across multiple documents.
- Ethical Considerations: Ethical data handling and privacy protection are equally vital within the NLP phase.

2.2. TF-IDF Weighting

Term Frequency-Inverse Document Frequency (TF-IDF) [34] weighting is a fundamental method employed in information retrieval, text mining, and natural language processing (NLP). This section delves into the intricacies of TF-IDF, its components, and its application within the context of this research :

- Term Frequency (TF): TF is the first component of TF-IDF and represents the frequency of a term within a document. It measures how often a term occurs within a specific document. In the context of this research, TF quantifies the frequency of words and phrases within product reviews. It's worth noting that TF values may vary significantly depending on the document's length, as longer documents are more likely to have higher TF values.
- Inverse Document Frequency (IDF): IDF is the second component of TF-IDF and assesses the significance of a term. It measures how unique or rare a term is across a corpus of documents. Terms that appear in many documents have lower IDF values, indicating they are less significant, while terms that appear in only a few documents have higher IDF values, indicating their importance. For this research, the IDF component helps in identifying terms that are unique to certain products or product features, contributing to their significance in summarization.
- TF-IDF Calculation: The TF-IDF score for a term within a specific document is determined by multiplying its TF and IDF values. This calculation helps to identify terms that are both frequent within a document and unique across the corpus, making them prime candidates for summarization. The TF-IDF score emphasizes terms that are specific to individual documents while diminishing the importance of common terms.
- The inverse document frequency (IDF) is often calculated using the formula (1), where D represents the total number of documents in the corpus, and df denotes the number of documents in which the term appears. This formula assesses how widespread or rare a term is across the entire dataset.

$$IDF = log\frac{D}{df} \tag{1}$$

$$TF - IDF = TF * IDF \tag{2}$$

- Equation (2): The final TF-IDF score for a term is obtained by multiplying the term's TF by its IDF, as per Equation (2). This calculation leads to a score that reflects the term's significance within the context of the document and the entire corpus.
- Importance in Summarization: In the context of product review summarization, TF-IDF plays a pivotal role in identifying salient terms and phrases that encapsulate the essence of reviews. The method is instrumental in selecting terms that are frequent within a specific product review (high TF) and rare across the entire collection of reviews (high IDF).
- Variants and Adaptations: While the core TF-IDF method is widely used, numerous adaptations and variants exist, each tailored to specific applications. Some adaptations consider factors like document length normalization to mitigate bias towards longer documents. The choice of a specific TF-IDF variant should align with the research objectives and characteristics of the dataset.
- Application Challenges: Applying TF-IDF in the e-commerce domain can present challenges related to handling diverse languages, interpreting the significance of multi-word terms, and accommodating large datasets. Researchers may need to address these challenges to ensure effective summarization.

- Ethical Considerations: The application of TF-IDF should adhere to ethical data handling practices, particularly concerning privacy and data protection, as personal information may be present in product reviews.

Understanding TF-IDF weighting is vital for the subsequent stages of text summarization [35], [36]. This method allows the research to identify key terms and phrases that represent the most pertinent information within product reviews, ultimately contributing to the creation of concise and meaningful summaries.

2.3. Data Collection and Preparation

Text summarization is a critical phase within this research, focusing on the extraction of succinct and informative summaries from extensive textual data, particularly product reviews [37]. In this section, we delve into the intricacies of text summarization, its techniques, and its importance within the e-commerce domain :

- Objective of Summarization: The primary goal of text summarization is to distill the most vital and relevant information from a document or set of documents. In the context of e-commerce, this process aims to condense the numerous product reviews into shorter, more manageable summaries that capture the essence of consumer feedback.
- Two Techniques: Extractive and Abstractive: Text summarization techniques can be broadly categorized into two methods: extractive and abstractive. Extractive summarization involves selecting sentences or phrases directly from the source text, while abstractive summarization generates new sentences that capture the main ideas. Researchers often choose the technique that aligns with the research objectives and dataset characteristics.
- Role of TF-IDF: As discussed in Section 2.2, TF-IDF plays a crucial role in text summarization. The method assists in identifying the most salient terms within product reviews, which can then be integrated into summaries. Terms with high TF-IDF scores are likely candidates for inclusion in summaries, as they are both frequent within specific reviews and unique across the corpus.
- Challenge of Multi-Document Summarization: In scenarios where numerous product reviews are available for a single product, multi-document summarization is essential. This process involves aggregating feedback from multiple sources into a coherent summary. Handling redundancy and diversity in opinions within this context can be challenging.
- Use of LexRank: The research leverages LexRank, an extractive summarization method based on graph theory. LexRank determines the importance of sentences in product reviews based on their similarity and centrality within the network of reviews. The LexRank approach provides a systematic and data-driven method for selecting the most representative sentences.
- LexRank Algorithm: The LexRank algorithm employs the idf-modified-cosine (3) similarity measure to assess the similarity between sentences. Sentences that are highly similar are deemed to be of greater significance in the summary. The algorithm also incorporates degree centrality, which assesses the relationships between sentences in the network, allowing for the identification of central sentences (4).

$$idf-modified-cosine(x,y) = \frac{\sum w \in x,y \; tf_{w,x} tf_{w,y} (idf_w)^2}{\sqrt{\sum x_i \in x \left(tf_{x_i,x} idf_{x_i}\right)^2} \text{ x } \sqrt{\sum y_i \in y \left(tf_{y_i,y} idf_{y_i}\right)^2}} \tag{3}$$

$$cosine\ matrix = \frac{cosine\ matrix}{degree\ centrality} \tag{4}$$

- Setting Thresholds: The research sets specific thresholds for the idf-modified-cosine similarity measure to determine which sentences are included in the summary. These thresholds can be adjusted to control the level of detail and length of the final summary, depending on the desired level of conciseness.
- Evaluation Metrics: Assessing the quality of the generated summaries is pivotal. The research employs the ROUGE metric, which focuses on the overlap and similarity between the generated summary and a reference summary. ROUGE provides a quantitative measure of the quality of the summarization, assessing aspects like recall and precision.
- Scalability and Automation: In the context of e-commerce, where vast numbers of product reviews are available, scalability and automation are essential. Summarization techniques need to be efficient and capable of handling large datasets.
- Ethical Considerations: As with other phases of data analysis, ethical considerations remain paramount in text summarization. Ensuring the privacy and data protection of consumers who have provided product reviews is a critical ethical concern.

Text summarization is the linchpin of the research, allowing for the extraction of the most pertinent information from product reviews. The method, incorporating TF-IDF and LexRank, enables the generation of concise and informative summaries that can assist both consumers and businesses in making informed decisions within the e-commerce domain.

2.4. LexRank: An Extractive Summarization Method

LexRank [20], a method for extractive summarization, stands as a pivotal element of this research, facilitating the systematic selection of sentences to create coherent product review summaries. In this section, we delve deeper into the mechanics of LexRank and its relevance in the e-commerce domain :

- Graph-Based Summarization: LexRank employs a graph-based approach to summarization, utilizing the principles of centrality and similarity to identify the most significant sentences within the corpus of product reviews. Each sentence is represented as a node in the graph, and edges are established based on the similarity of sentences.
- Eigenvalue Centrality (4): The core principle of LexRank is the concept of Eigenvector Centrality. Eigenvector centrality (5) assigns importance scores to each sentence, considering not only the similarity of a sentence to others but also the importance of the sentences it is connected to. This ensures that important sentences are both similar to other sentences and well-connected within the network.

$$Ax = \lambda x \tag{4}$$

$$det(\lambda I - A) = 0 \tag{5}$$

- Fully Connected Graph: LexRank assumes a fully connected, undirected graph in which each sentence is a vertex and sentence similarity serves as the edges. The edges are weighted based on a similarity metric, typically idf-modified-cosine similarity, which quantifies the resemblance between sentences.

- Similarity Measure: The similarity measure between sentences is crucial to LexRank's functioning. The idf-modified-cosine similarity is a common choice, comparing sentence vectors and incorporating inverse document frequency (IDF) to emphasize important terms that are rare in the corpus. The formula for idf-modified-cosine measures the overlap of terms between sentences and is central to the process (3).
- Threshold-Based Sentence Selection: LexRank employs threshold values to determine which sentences are included in the final summary. By setting specific similarity thresholds, the method controls the length and detail of the generated summary. Higher thresholds result in more concise summaries, whereas lower thresholds yield more extensive summaries.
- Degree Centrality: In addition to Eigenvector Centrality, LexRank uses degree centrality as another dimension for assessing the significance of sentences. Degree centrality quantifies the relationships between sentences and can help identify central sentences that may be pivotal in the summary.
- Scalability and Adaptability: LexRank has the advantage of scalability and adaptability to a wide range of textual datasets. Its effectiveness in generating concise summaries is well-suited for the large volume of product reviews that are typical in the e-commerce domain.
- Competitive Accuracy: LexRank is competitive with other summarization techniques, particularly in extractive multi-document summarization. Its ability to identify central sentences and its data-driven approach to threshold selection make it a powerful tool for producing meaningful summaries.
- Integration with Text Mining: LexRank is part of a broader process that encompasses text mining and natural language processing. The method complements text mining efforts by identifying and selecting the most critical sentences from a vast collection of reviews.
- Ethical Considerations: Ethical considerations in the use of LexRank are crucial, particularly in ensuring the privacy and data protection of consumers who have contributed product reviews.

LexRank plays a fundamental role in the research by providing a systematic and data-driven approach to extract the most salient sentences from a pool of product reviews. This method aligns with the overarching objective of creating concise, informative summaries that assist both consumers and businesses in the e-commerce sector in making informed decisions.

2.5. ROUGE: Evaluating the Quality of Summaries

ROUGE, an acronym for Recall-Oriented Understudy for Gisting Evaluation, stands as a crucial metric in this research, serving to assess the quality of the generated product review summaries [38], [39]. In this section, we delve deeper into the mechanics of ROUGE and its significance in the context of e-commerce product review summarization :

- Purpose of ROUGE: ROUGE is a suite of metrics designed to evaluate the quality of machine-generated text summaries by comparing them to reference summaries created manually. In the e-commerce domain, where the quality of product review summaries is paramount, ROUGE serves as a benchmark for measuring the effectiveness of the summarization process.

- Recall and Precision: ROUGE metrics primarily focus on two key dimensions: recall and precision. Recall assesses the extent to which the machine-generated summary captures information present in the reference summary. Precision measures how much of the machine-generated summary is relevant and not extraneous. Striking a balance between high recall and precision is vital to create informative yet concise summaries.
- N-Gram Overlap Metrics: ROUGE utilizes n-gram overlap to calculate recall and precision. It assesses how well the generated summary aligns with the reference summary in terms of shared n-grams (contiguous sequences of words or characters). Common n-gram measures used in ROUGE include ROUGE-1 (unigrams), ROUGE-2 (bigrams), and ROUGE-L (the longest common subsequence). These measures provide a nuanced evaluation of the summaries, taking into account the continuity and structure of the text.
- F1 Score: The F1 score, a harmonic mean of recall and precision, provides a comprehensive evaluation of the summary's quality. A high F1 score suggests that the summary successfully balances capturing relevant information and maintaining conciseness.
- Threshold-Based Evaluation: ROUGE metrics allow for threshold-based evaluation, enabling researchers to customize the assessment based on the desired level of detail and conciseness in the summary. By adjusting the threshold, it is possible to evaluate the summary at different levels of granularity.
- Comparing to Human-Generated Summaries: ROUGE metrics are invaluable for comparing machine-generated summaries to human-generated reference summaries. The metrics provide quantitative data to assess how well the machine-generated summary aligns with the human-generated gold standard.
- Multi-Document Summarization: In cases where multiple product reviews are aggregated into a single summary, ROUGE is particularly useful. It evaluates the ability of the machine-generated summary to capture the sentiments and key information from a variety of reviews. This is particularly important in e-commerce, where products often receive multiple reviews that express diverse opinions.
- Ethical Considerations: ROUGE metrics play a role in the ethical assessment of summarization processes. Ensuring that the machine-generated summaries respect the privacy and data protection of consumers who have provided product reviews is an ethical imperative.

ROUGE metrics are integral to the research, offering a quantitative and data-driven means to evaluate the quality of the generated product review summaries. By providing a structured and objective approach to assessing the effectiveness of the summarization process, ROUGE aids in the creation of summaries that are both informative and concise, aligning with the needs of consumers and businesses in the e-commerce sector.

3. Results and Discussion

In this section, we will discuss the results and findings obtained from the various stages of data processing, including data loading, standardization, tokenization, filtering, stemming, TF-IDF weighting, LexRank calculation, and summarization. We will also present the evaluation results using Rouge metrics and discuss the implications.

3.1. Data Preparation

In this section, we outline the preparatory steps taken to transform raw product comments into a format suitable for analysis. The key stages of data preparation include data loading, standardizing text through word normalization, tokenization, filtering, stemming, and applying the TF-IDF (Term Frequency-Inverse Document Frequency) weighting scheme.

The initial step involved the collection of product comments, each of which was assigned a unique ID per document (K1 to K10). These comments were then loaded into the analysis program, allowing for further data processing. A summary of the loaded data can be seen in Table 1.

Table 1. Load Data

ID	Comments
K1	Produk original,kualitas produk,pengiriman baik sooo reallllll pict,
K2	Barang baguuus, sesuai deskripsi, sesuai gambar, pengiriman agak lama karena dr luar negeri
K3	Case nya bagus, tebel. ga nyesel deh beli dsni
K4	barang sudah datang dengan selamat, tanpa ada yang kurang dan salah, pengiriman lumayan cepat, respon penjual juga ramah, dan paling penting harga paling murah tapi kualitas bagus banget, suka deh belanja disini
K5	Bagus rekomended banget
K6	Cantiikkk yang datang sesuai foto dan lengkao tanoa kurang 1 pun
K7	Karena memang barangnya sangat baguss, aluss dan soft banget, harga murmer jdi tunggu apalagi buruan diorder yaak softcasenya
K8	Barang sesuai dgn gambar dan permintaan. Sangat bagus dan lucu-lucu gambarnya, gak nyesel beli 3 untuk 1 hp doang. Biar buat ganti-ganti. Bahannya lumayan tebel utk harga segini. Makasih seller.
K9	Udah belanjaaa kesekiaan kalii nya disiniiii dan ga pernahh ngecewaiinnn, haraganya jugaa murahh dan pengirimannnya juga ga terlalu lama walaupun dari luarrrr
K10	Barang sesuai pesanan, dan aman sampai tujuan, recommended seller deh makasih ya kaak

This sample dataset consists of 10 product comments, each represented by a unique ID. Next phase, word normalization was executed, which involved converting slang or non-standard terms into standard words using a predefined dictionary. The "colloquial-indonesian-lexicon.csv" file contained a list of standard words used for this purpose. Tokenization was carried out to segment the text into individual words, allowing for further analysis on a word level. Each word was labeled with a header for identification. The filtering process was employed to eliminate less informative words, such as stop words, conjunctions, and other irrelevant terms, which may not contribute significantly to the analysis. Stemming was used to reduce words to their base or root form by removing affixes, making it possible to group words with the same root together. The TF-IDF weighting process was applied to calculate the importance of each term within the comments. It involved two primary components: TF (Term Frequency), which measures the frequency of each term within a document, and IDF (Inverse Document Frequency), which quantifies the importance of a term across the entire corpus of documents. The final step involved calculating the TF-IDF weights for each term within a document, providing a numerical representation of term importance.

These data preparation steps are crucial in ensuring that the subsequent analysis is conducted on standardized, structured, and relevant textual data, facilitating effective information retrieval and extraction. The processed data is then ready for more advanced analyses, as detailed in the following sections.

3.2. Equations

In this section, we delve into the text analysis process, which encompasses word tokenization, filtering, stemming, TF-IDF weighting, and a critical analysis of comment similarity through the LexRank algorithm. Furthermore, we present how these techniques were employed to generate product comment summaries, which are evaluated using Rouge (Recall-Oriented Understudy for Gisting Evaluation) metrics.

Tokenization is a critical component of text analysis that divides the text into individual tokens, typically words or phrases. In our analysis, tokenization was used to break down the product comments into their constituent words, which allowed us to explore the structure of the text and determine the frequency of terms within each comment. The tokenization process also assigned a header to each token for identification purposes. The filtering process followed tokenization. Its primary function was to remove words that were deemed less informative for the analysis. This involved eliminating stop words (e.g., conjunctions, prepositions) and other terms that do not significantly contribute to the overall meaning of the comments.

Stemming was employed to reduce inflected words to their root forms, providing a common base for words with similar meanings. By applying stemming, the analysis was able to treat variations of a word as the same term, thus enhancing the quality of the results. The TF-IDF (Term Frequency-Inverse Document Frequency) weighting technique was used to evaluate the importance of each term within the product comments. TF measured the frequency of a term within a comment, while IDF quantified the importance of that term across the entire corpus of comments. By multiplying these two components, we derived a weight for each term, signifying its significance within a particular comment.

The LexRank algorithm was applied to assess the similarity of product comments based on the computed TF-IDF weights. LexRank employs the Idf-Modified-Cosine approach, which computes the weights of each term within sentence pairs using TF-IDF. The results were then normalized, with a threshold of 0.1, 0.2, and 0.3 applied. As shown in Table 2.

Table 2. Load Data

	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
K1	1	0,02195	0	0,05053	0	0	0	0	0,01661	0
K2	0,02195	1	0,01160	0,03114	0,01842	0,06340	0,01872	0,14931	0,15922	0,06633
K3	0	0,01160	1	0,07599	0,02343	0	0,00838	0,23693	0,14631	0,05663
K4	0,05053	0,03114	0,07599	1	0,06805	0,11023	0,06377	0,04618	0,10685	0,03341
K5	0	0,01842	0,02343	0,06805	1	0	0,08727	0,01035	0	0
K6	0	0,06340	0	0,11023	0	1	0	0,07279	0	0,02786
K7	0	0,01872	0,00838	0,06377	0,08727	0	1	0,05922	0,04375	0,01066
K8	0	0,14931	0,23693	0,04618	0,01035	0,07279	0,05922	1	0,05558	0,15694
K9	0,01661	0,15922	0,14631	0,10685	0	0	0,04375	0,05558	1	0
K10	0	0,06633	0,05663	0,03341	0	0,02786	0,01066	0,15694	0	1

The similarity between each pair of comments was computed using the Idf-Modified-Cosine method. Pairs of comments with similarity values greater than 0 were considered relevant and formed edges in the graph, as illustrated in Fig. 1.

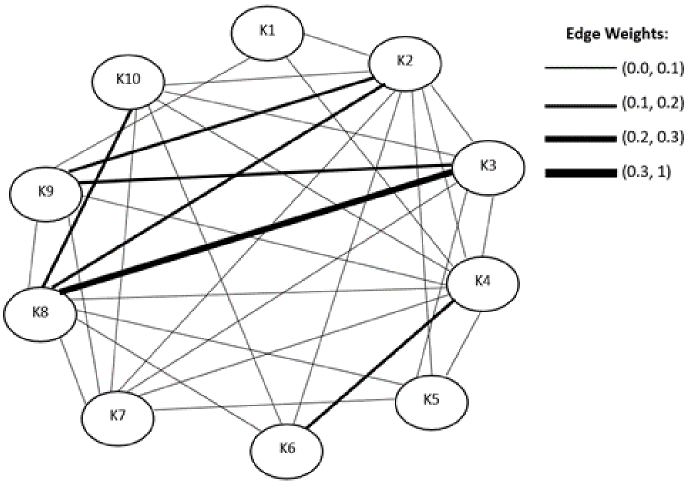

Fig. 1. Idf-Modified-Cosine Graph

After determining the similarity values, the next step was to rank the documents based on the degree centrality of each vertex. Degree centrality quantifies the number of edges connected to a vertex. This was done with different thresholds (0.1, 0.2, and 0.3), as depicted in Fig. 2.

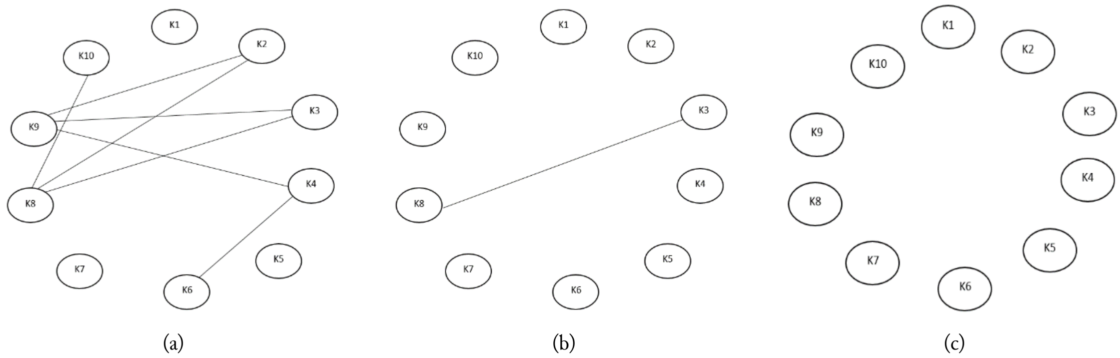

Fig. 2. Graph at Threshold of (a) 0.1, (b) 0.2 and (c) 0.3

The next step involves updating the similarity values by dividing the old similarity by degree centrality (Table 3).

Table 3. Degree Centrality Results

ID	degree(0.1)	degree (0.2)	degree (0.3)
K1	1	1	1
K2	3	1	1
K3	3	2	1
K4	3	1	1
K5	1	1	1
K6	2	1	1
K7	1	1	1
K8	4	2	1
K9	4	1	1
K10	2	1	1

Three updated similarity tables are generated for different thresholds (0.1, 0.2, and 0.3), as shown in Table 4. This step refines the edge weights between sentences.

Table 4. Cosine Matrix with Threshold 0.1, 0.2, and 0.3

	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
					Threshold 0.1					
K1	1	0	0	0	0	0	0	0	0	0
K2	0	0,33	0	0	0	0	0	0,25	0,25	0
K3	0	0	0,33	0	0	0	0	0,25	0,25	0
K4	0	0	0	0,33	0	0,5	0	0	0,25	0
K5	0	0	0	0	1	0	0	0	0	0
K6	0	0	0	0,5	0	0,5	0	0	0	0
K7	0	0	0	0	0	0	1	0	0	0
K8	0	0,25	0,25	0	0	0	0	0,25	0	0,5
K9	0	0,25	0,25	0,25	0	0	0	0	0,25	0
K10	0	0	0	0	0	0	0	0,5	0	0,5
					Threshold 0.2					
K1	1	0	0	0	0	0	0	0	0	0
K2	0	1	0	0	0	0	0	0	0	0
K3	0	0	0,5	0	0	0	0	0,5	0	0
K4	0	0	0	1	0	0	0	0	0	0
K5	0	0	0	0	1	0	0	0	0	0
K6	0	0	0	0	0	1	0	0	0	0
K7	0	0	0	0	0	0	1	0	0	0
K8	0	0	0,5	0	0	0	0	0,5	0	0
K9	0	0	0	0	0	0	0	0	1	0
K10	0	0	0	0	0	0	0	0	0	1
					Threshold 0.3					
K1	1	0	0	0	0	0	0	0	0	0
K2	0	1	0	0	0	0	0	0	0	0
K3	0	0	1	0	0	0	0	0	0	0
K4	0	0	0	1	0	0	0	0	0	0
K5	0	0	0	0	1	0	0	0	0	0
K6	0	0	0	0	0	1	0	0	0	0
K7	0	0	0	0	0	0	1	0	0	0
K8	0	0	0	0	0	0	0	1	0	0
K9	0	0	0	0	0	0	0	0	1	0
K10	0	0	0	0	0	0	0	0	0	1

To find the most significant sentences, the Power Method is applied to calculate the leading eigenvalues and eigenvectors of the cosine matrix at different thresholds. The eigenvectors represent the final weights assigned to each sentence, allowing for the selection of the most important sentences for summarization.

The eigenvectors for different thresholds are shown in Table 5. These eigenvectors represent the sentence weights for summarization.

3.2.1. Sentence Ranking

Based on the eigenvector values, sentences are ranked from highest to lowest. The top-ranked sentences form the summary. The number of sentences in the summary can be adjusted based on the desired length. Finally, the top-ranked sentences are selected to generate the summarized text from the original product comments. The summary offers a concise representation of the most significant information contained in the comments.

Table 5. Power Method Results

Threshold 0.1	Threshold 0.2	Threshold 0.3
1,0	1.41421	1,0
1.0	1,0	1,0
1.0	1,0	1,0
0.82858	1,0	1,0
0.50935	1,0	1,0
0.48197	1,0	1,0
-0.40156	1,0	1,0
-0.63620	1,0	1,0
-0.81912	1,0	1,0
-2.05041	2.2204e-16	1,0

3.3. Some Common Mistakes

In this section, we will discuss the findings and implications of our text summarization research, focusing on the results obtained from the two testing phases: the first testing involving 100 datasets and the second testing with 10 datasets. Our analysis will center on the performance of our system in generating extractive summaries and its sensitivity to different threshold values, as well as the implications of using extractive summarization as opposed to abstractive summarization. We compared the system-generated summaries with human-generated summaries. The results, as indicated in Table 6, revealed relatively low F1-scores for Rouge-1, Rouge-2, and Rouge-L. This outcome can be attributed to our system's use of extractive summarization techniques. Extractive summarization entails selecting portions of text directly from the source document to construct the summary. In contrast, human-generated summaries often exhibit abstractive summarization characteristics, where the text is rephrased and restructured to form a more concise and coherent summary. This disparity in summarization approaches can lead to discrepancies in evaluation scores.

Table 6. Results of the First Rouge Testing Phase

Testing Average	Threshold 01			Threshold 02			Threshold 03		
	F_1-Score			*F_1-Score*			*F_1-Score*		
	Rouge			*Rouge*			*Rouge*		
	1	*2*	*L*	*1*	*2*	*L*	*1*	*2*	*L*
100 Testing	0.1667	0.0301	0.1650	0.1668	0.0264	0.1657	0.1849	0.0299	0.1843
X100%	16.67	3.01	16.50	16.08	2.64	16.57	18.49	2.99	18.43

An interesting observation from the second testing phase, presented in Table 7, is the sensitivity of our system's performance to threshold values. The F1-scores varied with the choice of threshold (threshold 01, threshold 02, and threshold 03). Notably, the system performed relatively better at a threshold value of 0.2. This indicates that the threshold value plays a crucial role in the summarization process. Further investigation is needed to determine the optimal threshold that aligns with the desired summarization quality.

Table 7. Sensitivity to Threshold Values

Testing	Threshold 01			Threshold 02			Threshold 03		
	F_1-Score			*F_1-Score*			*F_1-Score*		
	Rouge			*Rouge*			*Rouge*		
	1	*2*	*L*	*1*	*2*	*L*	*1*	*2*	*L*
1	0.37	0.24	0.39	0.36	0.25	0.39	0.35	0.21	0.36
2	0.42	0.31	0.41	0.35	0.21	0.38	0.37	0.21	0.39
3	0.30	0.19	0.31	0.40	0.29	0.41	0.37	0.23	0.37
4	0.28	0.16	0.29	0.36	0.23	0.38	0.42	0.26	0.41
5	0.28	0.17	0.30	0.37	0.24	0.39	0.33	0.18	0.34
6	0.32	0.20	0.32	0.43	0.30	0.45	0.30	0.14	0.30
7	0.31	0.18	0.33	0.37	0.21	0.38	0.38	0.24	0.38
8	0.30	0.18	0.31	0.30	0.19	0.31	0.32	0.16	0.32
9	0.31	0.17	0.32	0.28	0.15	0.29	0.36	0.21	0.36
10	0.30	0.18	0.29	0.36	0.20	0.38	0.36	0.23	0.36
Average	0.33	0.21	0.34	0.36	0.23	0.38	0.36	0.21	0.36
X 100%	33	21	34	36	23	38	36	21	36

While the results show room for improvement, it's essential to recognize the system's strengths. Our system demonstrated its capability to handle large-scale summarization tasks, as evidenced by the use of 100 datasets. Additionally, it showed sensitivity to the choice of threshold values, offering opportunities for fine-tuning to improve performance.

Our system's performance was limited by its reliance on extractive summarization. To bridge the performance gap with human summaries, we should consider incorporating abstractive summarization techniques. Future research should explore methods for achieving a balance between extraction and abstraction, aiming for more contextually coherent summaries.

Furthermore, the evaluation process can benefit from a more extensive dataset for assessing system performance comprehensively. It is also essential to explore other evaluation metrics beyond Rouge to gain a more comprehensive understanding of summarization quality.

4. Conclusion

The research has resulted in the successful development of a Python-based system for text summarization, which greatly aids readers in extracting the essential information from product reviews with ease. This system effectively employs the LexRank method to produce succinct summaries of product reviews. The evaluation of the system's performance using Rouge testing, involving the comparison of system-generated summaries with human-crafted ones using 10 datasets, revealed varying levels of accuracy with different threshold values. With a threshold value of 0.1, the system achieved Rouge-1 accuracy of 33%, Rouge-2 accuracy of 21%, and Rouge-L accuracy of 34%. Notably, when the threshold was increased to 0.2, the system displayed substantial improvement, with Rouge-1 accuracy at 36%, Rouge-2 accuracy at 23%, and Rouge-L accuracy at 38%. Conversely, a threshold of 0.3 yielded Rouge-1 accuracy of 36%, Rouge-2 accuracy of 21%, and Rouge-L accuracy of 36%. In conclusion, the system's performance was notably superior at the threshold of 0.2 in comparison to thresholds of 0.1 and 0.3, signifying its proficiency in automatically summarizing product reviews within marketplace contexts. This research offers the potential for more effective and accurate product review summaries, which, in turn, can enhance consumer decision-making processes.

Acknowledgment

The authors would like to acknowledge the support and resources provided by the Study Program in Informatics Engineering at the Faculty of Industrial Engineering, Universitas Ahmad Dahlan, Yogyakarta, Indonesia. Their assistance and contributions have been instrumental in the successful execution of this research project.

References

[1] B. De Quero Navarro and A. Joy, “Confronting the Postmodern Intersection of Markets, Development, Globalization, and Technology: The Necessary Radical Vision of Nikhilesh Dholakia,” *Mark. Glob. Dev. Rev.*, vol. 7, no. 2, pp. 1-13, 2022, doi: 10.23860/MGDR-2022-07-02-02.

[2] B. Barann, J. H. Betzing, M. Niemann, B. Hoffmeister, and J. Becker, “Exploring customers’ likeliness to use e-service touchpoints in brick and mortar retail,” *Electron. Mark.*, vol. 32, no. 2, pp. 523–545, Jun. 2022, doi: 10.1007/s12525-020-00445-0.

[3] I. Bubanja and M. Vidas-Bubanja, “Managing trade transactions in the covid era: The rise of e-commerce,” *J. Eng. Manag. Compet.*, vol. 12, no. 1, pp. 20–34, 2022, doi: 10.5937/jemc2201020B.

[4] R. Chai, “Leveraging Corporate Social Innovation for Value Co-Creation: The Case of Superapps, an Integrative Literature Review,” *SSRN Electron. J.*, p. 20, 2021, doi: 10.2139/ssrn.4004373.

[5] Y. K. Dwivedi *et al.*, “Setting the future of digital and social media marketing research: Perspectives and research propositions,” *Int. J. Inf. Manage.*, vol. 59, p. 102168, Aug. 2021, doi: 10.1016/j.ijinfomgt.2020.102168.

[6] Y. Kim and S. Suh, “The Core Value of Sustainable Fashion: A Case Study on ‘Market Gredit,’” *Sustainability*, vol. 14, no. 21, p. 14423, Nov. 2022, doi: 10.3390/su142114423.

[7] Y. Wang, F. Jia, T. Schoenherr, and Y. Gong, “Supply Chain-Based Business Model Innovation: The Case of a Cross-Border E-Commerce Company,” *Sustainability*, vol. 10, no. 12, p. 4362, Nov. 2018, doi: 10.3390/su10124362.

[8] M. Matarazzo, L. Penco, G. Profumo, and R. Quaglia, “Digital transformation and customer value creation in Made in Italy SMEs: A dynamic capabilities perspective,” *J. Bus. Res.*, vol. 123, pp. 642–656, Feb. 2021, doi: 10.1016/j.jbusres.2020.10.033.

[9] N. . Saravana Kumar, K. Hariprasath, N. Kaviyavarshini, and A. Kavinya, “A study on forecasting bigmart sales using optimized machine learning techniques,” *Sci. Inf. Technol. Lett.*, vol. 1, no. 2, pp. 52–59, Nov. 2020, doi: 10.31763/sitech.v1i2.167.

[10] P.-S. Lo, Y. K. Dwivedi, G. Wei-Han Tan, K.-B. Ooi, E. Cheng-Xi Aw, and B. Metri, “Why do consumers buy impulsively during live streaming? A deep learning-based dual-stage SEM-ANN analysis,” *J. Bus. Res.*, vol. 147, pp. 325–337, Aug. 2022, doi: 10.1016/j.jbusres.2022.04.013.

[11] R. Bandara, M. Fernando, and S. Akter, “Privacy concerns in E-commerce: A taxonomy and a future research agenda,” *Electron. Mark.*, vol. 30, no. 3, pp. 629–647, Sep. 2020, doi: 10.1007/s12525-019-00375-6.

[12] B. Lu and S. Zhang, “A Conjoint Approach to Understanding Online Buyers’ Decisions towards Online Marketplaces,” *J. Theor. Appl. Electron. Commer. Res.*, vol. 15, no. 3, pp. 69–83, Sep. 2020, doi: 10.4067/S0718-18762020000300106.

[13] M. Nachar, “Factors that predict the adoption of online shopping in Saudi Arabia,” *Walden University*, 2019. [Online]. Available at: https://scholarworks.waldenu.edu/dissertations/6630/#:~:text.

[14] J. F. Gaski, “A history of brand misdefinition – with corresponding implications for mismeasurement and incoherent brand theory,” *J. Prod. Brand Manag.*, vol. 29, no. 4, pp. 517–530, Mar. 2020, doi: 10.1108/JPBM-11-2018-2124.

[15] L. Yang, Y. Li, J. Wang, and R. S. Sherratt, “Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning,” *IEEE Access*, vol. 8, pp. 23522–23530, 2020, doi: 10.1109/ACCESS.2020.2969854.

[16] L. Zheng, P. Jin, J. Zhao, and L. Yue, “Multi-dimensional Sentiment Analysis for Large-Scale E-commerce Reviews,” *Decker, H., Lhotská, L., Link, S., Spies, M., Wagner, R.R. Database Expert Syst. Appl. DEXA 2014. Lect. Notes Comput. Sci.*, vol. 8645, pp. 449–463, 2014, doi: 10.1007/978-3-319-10085-2_41.

[17] M. R. Ramadhan, S. N. Endah, and A. B. J. Mantau, “Implementation of Textrank Algorithm in Product Review Summarization,” in *2020 4th International Conference on Informatics and Computational Sciences (ICICoS)*, Nov. 2020, pp. 1–5, doi: 10.1109/ICICoS51170.2020.9299005.

[18] S. Novgorodov, I. Guy, G. Elad, and K. Radinsky, “Generating Product Descriptions from User Reviews,” in *The World Wide Web Conference*, May 2019, pp. 1354–1364, doi: 10.1145/3308558.3313532.

[19] S. Novgorodov, I. Guy, G. Elad, and K. Radinsky, “Descriptions from the Customers: Comparative Analysis of Review-based Product Description Generation Methods,” *ACM Trans. Internet Technol.*, vol. 20, no. 4, pp. 1–31, Nov. 2020, doi: 10.1145/3418202.

[20] S. Deo and D. Banik, “Text Summarization using Textrank and Lexrank through Latent Semantic analysis,” in *2022 OITS International Conference on Information Technology (OCIT)*, Dec. 2022, pp. 113–118, doi: 10.1109/OCIT56763.2022.00031.

[21] N. H. Cahyana, S. Saifullah, Y. Fauziah, A. S. Aribowo, and R. Drezewski, “Semi-supervised Text Annotation for Hate Speech Detection using K-Nearest Neighbors and Term Frequency-Inverse Document Frequency,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 10, pp. 147-151, 2022, doi: 10.14569/IJACSA.2022.0131020.

[22] S. Saifullah, Y. Fauziah, and A. S. Aribowo, “Comparison of Machine Learning for Sentiment Analysis in Detecting Anxiety Based on Social Media Data,” Jurnal Informatika, vol 15, no.1, pp. 45-55, Jan. 2021, doi: 10.26555/jifo.v15i1.a20111.

[23] D. Yadav *et al.*, “Qualitative Analysis of Text Summarization Techniques and Its Applications in Health Domain,” *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–14, Feb. 2022, doi: 10.1155/2022/3411881.

[24] M. S M, R. M P, A. R E, and E. S. G SR, “Text Summarization Using Text Frequency Ranking Sentence Prediction,” in *2020 4th International Conference on Computer, Communication and Signal Processing (ICCCSP)*, Sep. 2020, pp. 1–5, doi: 10.1109/ICCCSP49186.2020.9315203.

[25] M. Al-Dhelaan and A. Al-Suhaim, “Sentiment diversification for short review summarization,” in *Proceedings of the International Conference on Web Intelligence*, Aug. 2017, pp. 723–729, doi: 10.1145/3106426.3106525.

[26] N. Modani, E. Khabiri, H. Srinivasan, and J. Caverlee, “Creating Diverse Product Review Summaries: A Graph Approach,” *n Wang, J., al. Web Inf. Syst. Eng. – WISE 2015. WISE 2015. Lect. Notes Comput. Sci.*, vol. 9418, pp. 169–184, 2015, doi: 10.1007/978-3-319-26190-4_12.

[27] T. Shaik *et al.*, “A Review of the Trends and Challenges in Adopting Natural Language Processing Methods for Education Feedback Analysis,” *IEEE Access*, vol. 10, pp. 56720–56739, 2022, doi: 10.1109/ACCESS.2022.3177752.

[28] U. Singh, A. Saraswat, H. K. Azad, K. Abhishek, and S. Shitharth, “Towards improving e-commerce customer review analysis for sentiment detection,” *Sci. Rep.*, vol. 12, no. 1, p. 21983, Dec. 2022, doi: 10.1038/s41598-022-26432-3.

[29] S. Kumar, A. K. Kar, and P. V. Ilavarasan, “Applications of text mining in services management: A systematic literature review,” *Int. J. Inf. Manag. Data Insights*, vol. 1, no. 1, p. 100008, Apr. 2021, doi: 10.1016/j.jjimei.2021.100008.

[30] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data," *Front. Energy Res.*, vol. 9, Mar. pp. 1-17, 2021, doi: 10.3389/fenrg.2021.652801.

[31] D. Antons, E. Grünwald, P. Cichy, and T. O. Salge, "The application of text mining methods in innovation research: current state, evolution patterns, and development priorities," *R&D Manag.*, vol. 50, no. 3, pp. 329–351, Jun. 2020, doi: 10.1111/radm.12408.

[32] Kauffmann, Peral, Gil, Ferrández, Sellers, and Mora, "Managing Marketing Decision-Making with Sentiment Analysis: An Evaluation of the Main Product Features Using Text Data Mining," *Sustainability*, vol. 11, no. 15, p. 4235, Aug. 2019, doi: 10.3390/su11154235.

[33] Y. Wang and C. Zhang, "Using the full-text content of academic articles to identify and evaluate algorithm entities in the domain of natural language processing," *J. Informetr.*, vol. 14, no. 4, p. 101091, Nov. 2020, doi: 10.1016/j.joi.2020.101091.

[34] Y. Fauziah, S. Saifullah, and A. S. Aribowo, "Design Text Mining for Anxiety Detection using Machine Learning based-on Social Media Data during COVID-19 pandemic," in *Proceeding of LPPM UPN "Veteran" Yogyakarta Conference Series 2020–Engineering and Science Series*, 2020, vol. 1, no. 1, pp. 253–261. [Online]. Available at: https://proceeding.researchsynergypress.com/index.php/ess/article/view/117.

[35] S. Zaware, D. Patadiya, A. Gaikwad, S. Gulhane, and A. Thakare, "Text Summarization using TF-IDF and Textrank algorithm," in *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, Jun. 2021, pp. 1399–1407, doi: 10.1109/ICOEI51242.2021.9453071.

[36] M. Y. Saeed, M. Awais, R. Talib, and M. Younas, "Unstructured Text Documents Summarization With Multi-Stage Clustering," *IEEE Access*, vol. 8, pp. 212838–212854, 2020, doi: 10.1109/ACCESS.2020.3040506.

[37] Z. Nasar, S. W. Jaffry, and M. K. Malik, "Textual keyword extraction and summarization: State-of-the-art," *Inf. Process. Manag.*, vol. 56, no. 6, p. 102088, Nov. 2019, doi: 10.1016/j.ipm.2019.102088.

[38] G. N. H., R. Siautama, A. C. I. A., and D. Suhartono, "Extractive Hotel Review Summarization based on TF/IDF and Adjective-Noun Pairing by Considering Annual Sentiment Trends," *Procedia Comput. Sci.*, vol. 179, pp. 558–565, 2021, doi: 10.1016/j.procs.2021.01.040.

[39] C. Ma, W. E. Zhang, M. Guo, H. Wang, and Q. Z. Sheng, "Multi-document Summarization via Deep Learning Techniques: A Survey," *ACM Comput. Surv.*, vol. 55, no. 5, pp. 1–37, May 2023, doi: 10.1145/3529754.