International Journal on Digital Libraries
https://doi.org/10.1007/s00799-024-00402-8

Structured abstract generator (SAG) model: analysis of IMRAD structure of articles and its effect on extractive summarization

Ayşe Esra Özkan Çelik[1] · **Umut Al**[2]

Received: 11 July 2023 / Revised: 25 March 2024 / Accepted: 1 April 2024

Abstract

An abstract is the most crucial element that may convince readers to read the complete text of a scientific publication. However, studies show that in terms of organization, readability, and style, abstracts are also among the most troublesome parts of the pertinent manuscript. The ultimate goal of this article is to produce better understandable abstracts with automatic methods that will contribute to scientific communication in Turkish. We propose a summarization system based on extractive techniques combining general features that have been shown to be beneficial for Turkish. To construct the data set for this aim, a sample of 421 peer-reviewed Turkish articles in the field of librarianship and information science was developed. First, the structure of the full-texts, and their readability in comparison with author abstracts, were examined for text quality evaluation. A content-based evaluation of the system outputs was then carried out. System outputs, in cases of using and ignoring structural features of full-texts, were compared. Structured outputs outperformed classical outputs in terms of content and text quality. Each output group has better readability levels than their original abstracts. Additionally, it was discovered that higher-quality outputs are correlated with more structured full-texts, highlighting the importance of structural writing. Finally, it was determined that our system can facilitate the scholarly communication process as an auxiliary tool for authors and editors. Findings also indicate the significance of structural writing for better scholarly communication.

Keywords Abstracts · Readability · Scholarly communication · Automatic text summarization

1 Introduction

Abstracts are the most important textual tools in enabling potential readers to read the relevant full-texts from the huge stack of electronic information retrieved through the Internet. It is reported that there is a correlation between a scientific article's readability and impact determined by its subsequent citations or the possibility of being published in a top 5 journal in a relevant subject [1, 2]. However, compared to the relevant full-texts, abstracts are even much more subject to readability issues and structural flaws in their contents [3–6].

The electronic versions of scientific publications have become more preferred than the printed ones in a short time, with their advanced functionality that accelerates the access and publishing process [7]. However, electronic formats of scientific publications are almost identical to the printed formats. Thus, the electronic forms of publications have not increased the user experience in terms of readability [8]. In contrast, online communication brings new challenges to the scientific community for analyzing retrieved documents. These challenges include the distraction caused by being online, the obligation to choose from a stack of related articles, and the difficulty of maintaining focus while navigating through linked web pages [9–11]. Research has shown that reading and comprehending a lengthy electronic text, which requires scrolling and navigating back and forth, demands more mental effort than reading a printed text [12, 13]. Screen reading has been found to be inherently distracting, mainly because of the above mentioned multitasking nature of online reading [14].

While reading lengthy electronic texts can be challenging, scientific publications are constructed and archived follow-

✉ Ayşe Esra Özkan Çelik
esra@hacettepe.edu.tr

Umut Al
umutal@hacettepe.edu.tr

[1] Library, Hacettepe University, Beytepe, 06800 Ankara, Turkey

[2] Department of Information Management, Hacettepe University, Beytepe, 06800 Ankara, Turkey

Published online: 07 May 2024

ing certain rules, making them highly structured text data [15]. The components of a scientific article, including title, abstract, keywords, article body, acknowledgments, bibliography, and appendices, each have very specific functions and are located in particular places within a manuscript. The article bodies also follow a well-defined structure over time, largely due to the introduction of the IMRAD (Introduction, Methods, Results, and Discussion) format by Pasteur in 1876 [3]. The IMRAD format is now widely adopted by the scientific community as it ensures that articles are well-organized and easy to read, regardless of whether they are published in electronic or print format. Each section has a specific role in communicating the research findings as follows:

- Introduction: What was studied and why?
- Methods: How was the study conducted?
- Results: What were the findings?
- Discussion: What do the findings mean?

Before reading the body text, readers first encounter titles and sometimes keywords that contain very limited information about the article. Abstracts, on the other hand, are the first and last stop for the reader to learn the content before proceeding to review the full-text. Therefore, for most readers, an article is as interesting as its abstract. Studies have shown that nearly half of the readers of scientific articles who read the abstracts also read the full-texts [16]. In a study, users' transaction records of more than 1000 scientists, and 17,000 sessions on ScienceDirect were examined [6, 17]. It was found that at least 20% of the users only read abstracts and that they trust the abstracts to select the relevant articles and to provide the necessary preliminary information for their research.

The language used in the abstract should be clear enough so that everyone can understand it, even if they don't know much about the topic or English isn't their first language. However, it's often the case that abstracts are more difficult to read than the main body of an article [3–5, 18, 19]. Moreover, the abstract section should also cover the major information given in the full-text. Studies have found that skipping necessary information in abstracts is a frequently observed problem [6, 20–22].

How can abstracts be written to persuade readers to read the full text, especially if the reader has difficulty understanding the abstract? Structured abstract writing may be a solution, as it can improve readability and comprehension by dividing the text into subheadings [23]. In this way the informativeness of the abstract increases. When compared to unstructured abstracts, structured abstracts have significantly higher information quality [24]. Further, the indexing performance of the publication increases. It provides ease of access to the user and increased relevance in search results. This facilitates access to the article for all users with varying degrees of familiarity with the subject of the publication. The structural headings can help readers to find and understand the information they need more easily. It is easier for the author to write an abstract using a structured format than a classical one. The author cannot forget to mention all parts of the publication in the abstract. In that manner, abstract full-text consistency increases. It is preferred more by the readers and authors than the classical versions [23].

Given the critical role of abstracts in scholarly communication, this study is conducted to enhance the informativeness of abstracts by utilizing the high readability of full-text sentences and the structured ordering inherited from the full-text articles.

2 Literature review

The main research topics related to abstracts in the literature deal with organizational issues, readability issues and presentation issues in general. Many researchers have found that abstracts do not follow the structural order followed in the full-text, if the journal does not have a specific policy on this issue.

In the process of deciding whether to read the full text of an academic article, readers are most interested in descriptive information about the research problem, method, or results. Skipping information about these parts in abstracts is a frequently observed problem [6, 20–22]. The abstract of a scientific paper often contains long, inverted sentences with conjunctions and intensive use of specific technical terms or jargon related to the field. The conscious preference for such sophisticated language features has resulted in abstracts becoming progressively more difficult to read over time. The readability of an abstract is usually found more difficult than the other parts of the article [3–5, 18, 19]. Although the subject of the presentation is an element that should be considered separately from the readability context [25], it is difficult to read an abstract written in a single block without paragraphs and subtitles, in fonts smaller than the full-text, and sometimes in italics [26, 27]. The abstract formats required by journals vary. The two most dominant formats are classical (or traditional) abstracts and structured abstracts. Classical abstracts which are preferred by most journals, are not produced in a format that will attract the attention of the reader within the scope of the presentation. Abstracts that are written in a single block in an unstructured format, without paragraphs and subheadings, are generally called classical. Structured abstracts must be produced by filling in all the structural titles specified by the journal.

Luhn [28] carried out his pioneering work in the field of automatic text summarization in order to save the reader time and effort in finding useful information in an article or

report when the widespread use of the Internet and information technologies were not yet on the agenda. Since then, the summarization of scientific textual data has become a necessary and crucial task in Natural Language Processing (NLP) [29, 30]. However, there are certain difficulties such as the abstract generation, having labeled training and test corpora, and the scaling of collections of large documents.

Research in automatic text summarization has witnessed a proliferation of techniques since the beginning. The process generally involves several stages, including pre-processing the source document, extracting relevant features, and applying a summary generation method or algorithms. In the pre-processing stage, text documents are prepared for the next stages using linguistic techniques such as sentence segmentation, punctuation removal, stop word filtering, stemming, etc. Then, words are converted to numbers for computers to decode language patterns. Common methods include bag-of-words, n-grams, tf-idf, and word embeddings. For feature extraction, some of the commonly used features [31] that are used at both the word and sentence level to identify and extract salient sentences from documents are listed below:

Word level features

- Keywords (content words): Nouns, verbs, adjectives, and adverbs with high TF-IDF scores suggesting sentence importance.
- Title words: Sentences containing words from the title are likely to be relevant to the topic of the document.
- Cue Phrases: Phrases such as "conclusion", "because", "this information", etc. that indicate structure or importance.
- Biased words: Domain-specific words that reflect the topic of the document are considered important.
- Capitalized words: Names or acronyms such as "UNICEF" that indicate important entities.

Sentence level features

- Sentence Location: Sentences in the document are prioritized due to information hierarchy. For instance, beginning and ending sentences are likely to hold more weight.
- Length: Optimal length of sentences plays an important role in identifying excessive detail or lack of information.
- Paragraph Location: Similar to sentence location, beginning and ending paragraphs of the document carry higher weight.
- Sentence-Sentence Similarity: Sentences with higher similarity to other sentences of the document indicate their importance.

Text summarization methods are typically confined to extractive and abstractive summarization. In extractive text summarization, supervised and unsupervised learning methods are applied. Supervised learning needs a labeled dataset containing both summarized and non-summarized text, while unsupervised learning uses advanced algorithms such as fuzzy-based, graph-based, concept-based, and latent semantics to process input automatically [32].

Summarization of scientific papers is one of the applications of automatic summarization. Abstract generation-based applications and citation-based applications are two main branches of scientific article summarization. Other applications focus on specific problems such as the summarization of tables, figures, or specific sections of the related article [29]. Turkish text summarization studies primarily used extractive techniques due to a deficiency of trained corpora, a requirement that is still unmet in languages with limited resources like Turkish [33].

In addition, in scientific article summarization, single-article summarization with extractive techniques has predominantly been used with the high dominance of combinations of statistical and machine learning approaches, and intrinsic evaluation methods which are largely based on ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics [29]. The ROUGE evaluation of an automated scientific article summarization system that focused on the dataset containing academic articles shows that the extractive algorithms are better than the abstractive algorithms [34].

Our summarization model is based on a study [35] that evaluated the performance of 15 different extractive-based sentence selection methods, both individually and combined, on 20 Turkish news documents. The study aimed to select the most important sentences in a document. They analyzed the outputs of the methods based on the summaries of sentences hand-selected by 30 evaluators. The best results were obtained when the sentence position, number of common adjacencies, and inclusion of nouns were combined. While these features were combined in a linear function, their weights were kept equal.

3 Research objectives and questions

We propose a summarization model based on extractive techniques combining general sentence selection features that have been shown by human judgments to be beneficial for Turkish [35]. Our study aims to assess the suitability of the Turkish librarianship and information science (LIS) corpus for automatic summarization methods by evaluating it from a broad perspective, rather than developing our own method. We focus on the full-text structural order to improve the extractive sentence selection process. Additionally, we compare the readability levels of full texts and abstracts to emphasize the significance of readability in scholarly com-

munication. Raising awareness of this issue is also important, especially among LIS professionals.

The field of LIS is a broad and interdisciplinary field that encompasses a wide range of research topics. That is characterized by integrating research paradigms and methodologies from various disciplines [36]. This interdisciplinary nature makes LIS an ideal domain to examine the structural layouts of various approaches employed in scientific articles which can be extended to other fields. Due to this characteristic, LIS was selected as the domain in this study.

The main goal of this study is to understand the benefits of generating structured abstracts using extractive methods. We aim to identify the most feasible way to generate abstracts for scholarly communication in Turkish. It is clear that choosing the most important sentences from each structural section of a scientific article and presenting them under the structural headings will facilitate the abstract generation process. Moreover, such structural sectioning increases the semantic integrity and readability of an abstract. Our main hypothesis is "Considering the structural features of full-texts in extracting abstract sentences with automatic methods will increase the quality of the outputs". The study attempts to answer the following research questions: (1) Are the full-texts of Turkish LIS articles organized taking into consideration the basic structural features that are expected to exist in a scientific publication? (2) What is the readability of the full-texts and the abstracts of Turkish LIS articles, based on the readability scale? (3) Does using full-text structural features in extracting abstracts with automated methods improve output quality?

In our study, we examined articles published in the field of LIS with classical abstracts. The corpus was analyzed to determine whether the full-texts of the articles are more readable and better structured than the classical author abstracts. We generate a simple automatic abstract generator model that chooses the most important sentences from each structural section of each article.

4 Methodology

We utilized an extractive automatic summarization system named, Structured Abstract Generator (SAG), which depends on the extraction of the most important sentences from all structural parts of the full-texts of articles. Figure 1 demonstrates the architecture of the SAG. This section describes the methodology used in the study.

4.1 Data collection and representation

To construct a corpus for the study, *Türk Kütüphaneciliği - Turkish Librarianship* (TL) and *Bilgi Dünyası -Information World* (IW), which are major journals in the field of librarianship and information science in Turkey, have been used. Both journals asked the authors to develop classical abstracts. In addition, both journals do not set either an IMRAD or similar clear template for full-texts. However, IW draws a framework in line with the IMRAD regarding the arrangement of the content. All refereed articles written in Turkish were included in the study. Since each journal is open access, there was no problem in accessing these articles. This study is the first in Turkish to conduct a detailed full-text analysis of a large corpus of LIS literature.

In the initial stage, all articles were saved in PDF format with a unique identifier that encoded the journal name, year, volume, and issue information. For example, the identifier BD200011 indicates an article published in the year 2000, which is the 1st volume of the year and the 1st article of the volume in the IW (BD in Turkish) journal.

Once the articles were identified, they were converted into.txt format using UTF-8 character encoding to ensure the correct representation of Turkish characters. Then, article metadata was automatically extracted. This included author names, titles, abstracts, body text, and keywords, which are clear indicators of the content and are located in specific places in the document.

After processing 421 documents from two journals (172 IW, 249 TL), a relational database was created using MySQL. This database enabled the efficient processing of article full-text sentences as vectors, where each component is assigned to the corresponding structural section of the document, as well as the document's metadata. The IMRAD format, which is the most prominent organizational structure for full-text in scientific writing, was used in this study.

To facilitate further stages, web-based interfaces were developed to enable the monitoring and management of rules governing the structural layout decisions for each article. The development of a web-based system offered inherent advantages in terms of providing flexible work arrangements and enabling quick control over individuals in operator roles. The solution was designed to be compatible with both mobile and desktop devices, enabling the team to operate flexibly and remotely.

The team of operators consisted of six professionals, two undergraduate students, and four PhD students from the Department of Information Management. These individuals had prior expertise regarding the structural components of scientific articles. Two roles were identified for the expert team: operator (4 experts) and administrator (2 experts).

Operators copied and pasted the body text from these interfaces according to IMRAD headings, retaining complete control over the process. After the completion of the IMRAD marking procedure for an article, operators were unable to make any additional modifications using the interface. However, administrators retained the authorization to execute final supervision and operational functions subsequent to this stage. This control was important to ensure that

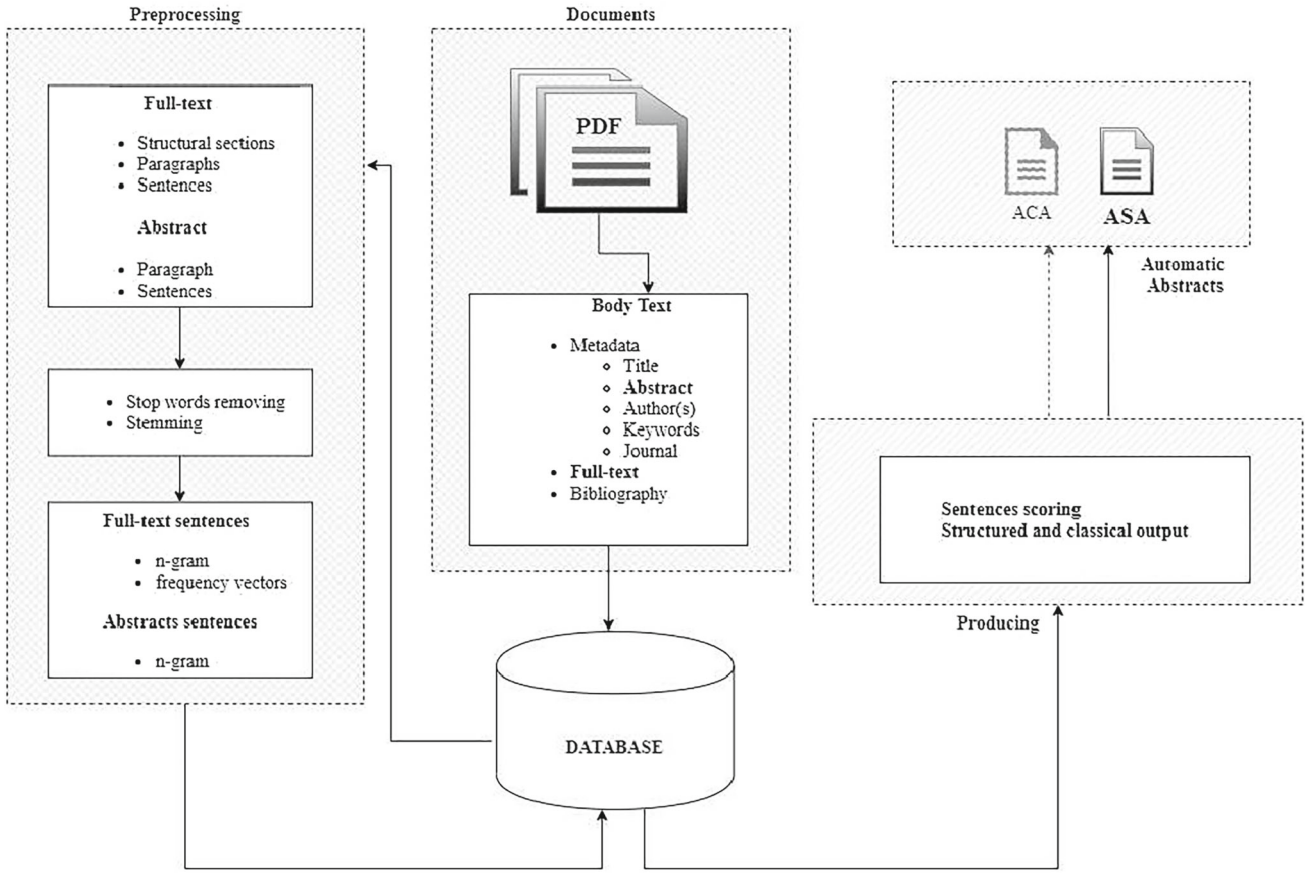

Fig. 1 SAG architecture

the IMRAD structure of the articles, which was inherited by paragraphs, was determined correctly. To ensure inter-annotator agreement of scholarship decisions, each article was tagged by at least two operators and one expert doctoral student during the manual step.

By implementing this work plan, the expert team successfully achieved the systematic and efficient classification of the boundaries and structural sections (according to the IMRAD format) of each paragraph of the body text. Consequently, the work of carefully adhering to the sequential arrangement of sentences in all articles was successfully completed within a brief timeframe. This hierarchical structure of body text was further applied to the sentence level through the utilization of a relational database. At the end of the two main steps mentioned above, 101,019 sentences were extracted from 421 articles. Next, word frequency vectors and n-gram sequences were obtained using Zemberek [37] and then stored in the database.

Table 1 shows an example of the data representation for a sentence of an article. The ID BD200011 indicates that the sentence is from the first article of the first volume of the year 2000 of the IW (BD in Turkish) journal. The remaining information refers to the 27th sentence of the 5th paragraph of the 1st IMRAD section of the relevant article. In this study, we used the following section numbers: 1 for Introduction, 2 for Method, 3 for Results, and 4 for Discussion. The title information indicates the title of the paragraph to which this sentence belongs.

4.2 Stemming

Since Turkish has an agglutinative morphology, inflectional or plural suffixes may produce multiple words from one root. Turkish words that appear in different ways in the text but have the same meaning in terms of their roots can be shown in a single way. Due to the high reduction rate provided in the size of the document-term matrix, it is strongly recommended to apply to stemming in Turkish texts [38]. For root finding, we utilize Zemberek [37], a natural language processing toolkit for Turkish for root finding. Although sentences of articles had been parsed under the supervision of the operators, we employed data-cleaning methods on the raw data.

After the stemming and data-cleaning processes, word frequency vectors are produced. Table 2 depicts the example of a vector representation of a sentence whose raw data is seen in Table 1.

Table 1 Data representation of a sentence of an article

Paper id	Sentence_no	Paragraph_no	Imrad_no	Title	Text
BD200011	27	5	1	geleceğe yönelik tartışmalar	tarım toplumundan sanayi toplumuna geçiş eğitimi nasıl etkilediyse sanayi toplumundan bilgi toplumuna geçiş de kurumların yapısında köklü değişiklikleri zorunlu kılmaktadır

Table 2 Word frequency vector example

Paper id	Sentence_no	Imrad_no	Words	Word_vector
BD200011	27	1	toplum, sanayi, geç, eğitim, yapı, bilgi, tarım, kur, kıl, değiş nasıl, etki, kök, zorunlu	4,2,2,1,1, 1,1,1,1,1 1,1,1,1

4.3 Extractive summarization and evaluation process

Extractive automatic summarization methods include the process of scoring, sorting and selecting sentences in the document. Automatic text summarization approaches and methods are employed to identify key representative sentences from the full-text. Sentences are scored based on their predetermined features, and the significance of each sentence in the document is determined by these scores. Sentence selection functions that bring together each feature by weighting are another stage of the extractive automatic summarization systems. Features used in sentence scoring are as follows.

4.3.1 Sentence position

This feature assumes that the most important information in a text is usually presented at the beginning. It assigns a higher ranking score to sentences that are closer to the beginning of the text, using the following formula

$$\mathrm{SP}\left(\mathrm{s_i}\right) = \frac{(\mathrm{n}-\mathrm{i})}{(\mathrm{n}-1)} \tag{1}$$

here i is the sequence number of the sentence in the document and n is the number of sentences in the document.

Formula 1 gives, each sentence ranking points from 1 to 0 depending on the order of appearance in the article.

4.3.2 Sentence centrality

Centrality is the most widely used feature in automatic text summarization for a variety of text types and corpora. It is based on finding the degrees of representing the basic information given in the full-text, in terms of the scoring of the sentences. It is calculated by considering how many other sentences in the document are connected to it. There are many different ways to calculate centrality. Within the scope of the study, the centrality of each sentence for a document with n sentences was obtained as in Formula (2) [39].

$$\begin{aligned} S &= \sum_{\mathrm{j}=1}^{\mathrm{n}-1} \mathrm{sim}(\mathrm{s_i}, \mathrm{s_j}) \\ F &= \sum_{\mathrm{j}=1}^{\mathrm{n}-1} \text{n-friends}(\mathrm{s_i}, \mathrm{s_j}) \\ G &= \sum_{\mathrm{j}=1}^{\mathrm{n}-1} \text{n-gram}(\mathrm{s_i}, \mathrm{s_j}) \\ SC\left(\mathrm{s_j}\right) &= \frac{S+F+G}{\mathrm{n}-1} \end{aligned} \tag{2}$$

here $\mathrm{i} \neq \mathrm{j}$ and $\cos\left(\mathrm{s_i}, \mathrm{s_j}\right) \geq 0.16$.

Sentence centrality based on three factors: the similarity between a sentence $\mathrm{s_i}$ and other sentences $\mathrm{s_j}$ in the document, the number of shared words (n-friends) between $\mathrm{s_i}$ and $\mathrm{s_j}$, and the presence of common n-grams between them. The resulting sum is then normalized by dividing it by n-1, where n is the number of sentences in the document. An experimentally determined threshold value of $\cos\left(\mathrm{s_i}, \mathrm{s_j}\right) \geq 0.16$ was found

to be appropriate. Accordingly,

$$\text{n-friends}\left(s_i, s_j\right) = \frac{\left|s_i\ (\text{friends}) \cap s_j\ (\text{friends})\right|}{\left|s_i\ (\text{friends}) \cup s_j\ (\text{friends})\right|} \tag{3}$$

$$\text{n-grams}\left(s_i, s_j\right) = \frac{\left|s_i\ (\text{n-grams}) \cap s_j\ (\text{n-grams})\right|}{\left|s_i\ (\text{n-grams}) \cup s_j\ (\text{n-grams})\right|} \tag{4}$$

where $i \neq j$. Here, the number of shared affinities are calculated as in Formula 3 over sets of sentences similar to both s_i and s_j. 2-grams were used for shared n-grams in Formula 4. |X| gives the number of elements of the set X.

The sim value of each sentence is calculated using the cosine similarity measure [40]. Cosine similarity is one of the most preferred methods to compare two texts and to make decisions over the similarity between them.

Let X and Y be vector representations of the two sentences to be compared. Given the Euclidean norm of X, $||X|| = \sqrt{x_1^2 + x_2^2 + \cdots + x_p^2}$ and the vector product of X and Y, to be defined by $XY = x_1y_1 + x_2y_2 + \cdots + x_py_p$, the cosine value of the angle θ between the two vectors gives the similarity value of the two sentences represented by these two vectors as in Formula (5) [41].

$$\text{sim}\,(X,Y) = cos(\theta) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \tag{5}$$

4.3.3 Noun score

Another feature discussed in this study is whether the sentences contain nouns. The nouns in the texts transmit the information about the content of the text. Therefore, the text summarization system gives points to the sentences containing nouns according to the number of nouns they contain. Zemberek [37] was used to calculate the score. That score (NS) of each sentence was added to the formula after normalizing by a count of all words of the related sentence.

4.3.4 Ranking score

By combining the linear Formula (6), which accepts the weights of all three mentioned features as equal, the ranking scores $RS(s_i)$ are calculated as follows.

$$RS(s_i) = SP(s_i) + SC(s_i) + NS(s_i) \tag{6}$$

here i is the sequence number of the sentence s_i in the document. The word frequency vectors and n-gram sequences stored in the database were used in sentence score calculation.

4.3.5 Generating automatic abstracts

The intended outputs of our system are automatic structured abstracts (ASA). In addition to these outputs, we evaluated the impact of considering structural features on the performance of an extractive-based text summarization system with automatic classical abstracts (ACA) without using structural features, with the same ranking function. The structural section marking of the corpus full-texts is compatible with the widely accepted and well-known IMRAD headings, so the layout of the ASA output of our system is also compatible with IMRAD.

The word limit for our system's output was determined by reviewing the TL and IW journal guides. The journal TL does not have a word limit for abstracts, while the journal IW has a 250-word limit, which we considered reasonable. Usually, journal guides indicate a word limit for abstracts, with the range being from 150 to 300 words (APA, 2010). As such, we set a 250-word limit for the output of our automated structural abstract system.

For ASAs, the 250-word limit is divided equally among the structural sections of the article. The highest-scoring sentences are selected from each section until the word limit for that section is reached. In this step, sentences are first sorted according to their structural section and then according to their score. For ACAs, the highest-scoring sentences are selected from the entire article until the 250-word limit is reached. In this step, we only sort sentences according to their score.

4.3.6 Evaluation process

In this study, the effect of selecting sentences by considering the structural features of the full-text while generating abstracts was measured using automatic methods. The evaluation is conducted in three stages. Firstly, the distribution of selected sentences for ASA and ACAs within the full text is compared to ensure that the automatic summaries are representative. Next, the full text, original abstract, ASA, and ACA are evaluated for readability to determine whether the automatic summaries are easier to understand than the author summaries. Finally, structural (ASA) and non-structural (ACA) automatic summaries are compared using n-gram co-occurrence between the original abstracts to measure quality and effectiveness. ROUGE scores [42] are used to compare n-grams in the reference summaries and the extracted summaries as a standard of automatic evaluation of document summarization.

ROUGE evaluation

The ROUGE evaluation approach is based on n-gram co-occurrence, longest common subsequence, and weighted longest common subsequence between the ideal summary

and the extracted summary [42]. The n-grams are ordered terms of length n derived from a given sequence of text used to find the association statistic between reference summary and candidate summary. Formula (7) calculates the nominal value for each ROUGE-N between the candidate abstract and the reference abstract(s).

$$\text{ROUGE-N} = \frac{\sum_{S\epsilon\{RefS\}} \sum_{gram_n} Count_{match}\left(gram_n\right)}{\sum_{S\epsilon\{RefS\}} \sum_{gram_n} Count\left(gram_n\right)} \quad (7)$$

where n is the length of n-grams and $Count_{match}$ is the maximum number of n-gram overlaps seen in the reference and candidate abstracts [42]. When X and Y represent two different pieces of text, the overlap between them is calculated as in Formula (8) [43]. ||X|| represents the size of the relevant text.

$$\text{overlap}(X, Y) = \frac{||X \cap Y||}{(||X|| + ||Y|| - ||X \cap Y||)} \quad (8)$$

It is a common approach to use abstracts written by the author as reference abstracts in the evaluation process when performing automatic summarization studies for academic articles. Within the scope of this study, author abstracts were used as reference abstracts to calculate the n-gram overlaps of the system outputs with the recall, precision, and F-score scores obtained based on the ROUGE measurements. The ROUGE 2.0 [44] package was employed in this stage. This comparison is obtained using the mean recall, mean precision, and mean F-score values relative to the ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-SU4 measurements. The precision value is obtained by dividing the total number of instances included in the ideal and system-generated summaries by the number of instances in the system summary. The recall value is calculated by dividing the number of instances in the ideal and system summaries together by the total number of instances in the ideal summary. The F-score value is obtained by combining the precision and recall values. The simplest way to obtain F-score is to calculate the harmonic mean of these two values [45]. For more reliable results on a sentence basis, two ROUGE values are used during the evaluation phase. These are ROUGE-L and ROUGE-SU4. ROUGE-L gives the longest common subsequence (LCS) measurement and is calculated by Formula (9) [43].

$$\text{LCS}(X, Y) = \frac{\text{length}(X) + \text{length}(Y) - \text{edit}_{di}(X, Y)}{2} \quad (9)$$

here LCS(X, Y) is the length of the longest common subsequence of X and Y. Length values are the length value of the relevant texts whereas edit_{di} (X, Y) is the minimum number of deletion and addition operations which are required to transform X into Y [46]. The LCS is sensitive to how information is ordered in the text. The disadvantage of ROUGE-L is that it may catch the main word sequence in the text and skip the side subjects that create shorter sequences [42]. ROUGE-SU4, which evaluates any word pair by allowing arbitrary spaces in the sentence order, measures the 2-gram association created by skipping four 1-grams at most [42].

Table 3 Atesman's readability scale

Readability value	Readability scale of text
90–100	Very easy
70–89	Easy
50–69	Fairly difficult
30–49	Difficult
0–29	Very difficult

Readability of texts

Reading is a complex process that requires readers to make sense of the given message, comprehend it, and finally interpret it [47]. The suitability of the text for the target audience can be determined through readability calculations.

Although a language-specific formula has not been produced to measure the readability of Turkish texts, an adaptation of the well-known formula called "Flesch Reading Ease" (FRE) [48] has been widely used since 1997. This adaptation is known as Atesman's Readability Formula [49], which calculates the readability of a text based on the average syllable length of the words in the text and the average number of words per sentence.

The Atesman's Readability Values (ARVs) are calculated with the formulas given in (10) and (11) below:

$$\text{readability} = 198.825 - 40.175a - 2.610b \quad (10)$$

$$a = \frac{\text{count of syllable}}{\text{count of word}}, b = \frac{\text{count of word}}{\text{count of sentence}} \quad (11)$$

The readability scale for Turkish texts using ARVs is given in Table 3.

Academic texts are typically challenging since they contain a lot of jargon specific to the study domain and lengthy sentences with conjunctions. In our study, we have a domain-specific corpus of articles with similar linguistic characteristics. Thus, it is believed that assessing the text's readability based on the length of sentences and words will be distinctive. While examining the characteristics of the corpus, we calculated the readability values of the body text and traditional abstracts of each article using ARVs. Finally, we compared these calculations with the ARVs of system outputs.

Table 4 Count of IMRAD patterns used in the articles

IMRAD (#)	Article (%)	Pattern	Article (#)
1	4.7	I	19
		R	1
2	45.8	I,D	193
3	6.8	I,M,R	3
		I,M,D	1
		I,R,D	25
4	42.5	I,M,R,D	179

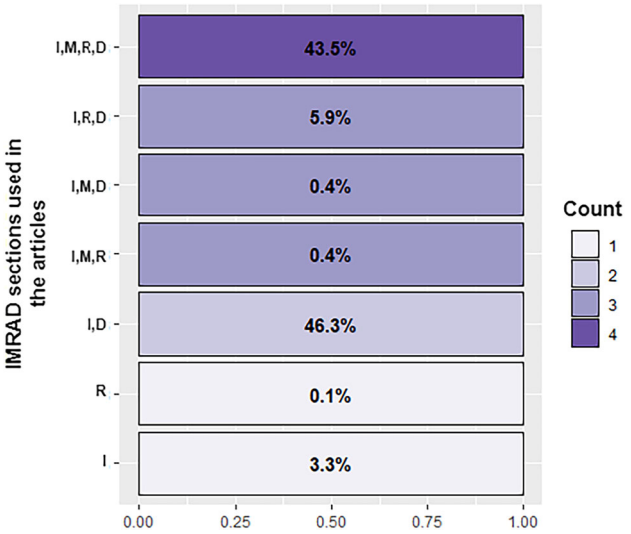

Fig. 2 Percentages of IMRAD patterns of the corpus. The color code darkens as the count of sections that are compatible with IMRAD increases

5 Results

It's crucial to ensure the corpus texts are structured in a way that supports our analysis. All IMRAD patterns used in the articles are represented in Table 4. The number of articles in which each pattern is used and the percentages of these articles in the corpus can also be seen in Table 4.

Figure 2 depicts how the structural order of articles influences the weight distribution of sentences. In Fig. 2, we see that at least half of the sentences (43.5%) come from articles that use a proper IMRAD format (I, M, R, D). With the addition of sentences coming from articles with an introduction and discussion (I, D) (46.3%), we can say that 89.8% of the sentences come from articles with an acceptable IMRAD structure, as there is a consensus that these types of articles are also suitable for non-experimental social science topics.

However, it is important to note that every scientific article must contain research question(s) and a method adopted to investigate the question(s). Therefore, the findings about the research question(s) should also be included in the articles. Articles with methods without results (I,M,D), results without methods (I,R,D), or methods and results without discussion (I,M,R) are incompatible with academic writing, as they do not provide a complete account of the research. However, these sentences are the minority of our corpus, constituting only 6.7% (5.9% + 0.4% + 0.4%) of the total. Also, articles consisting of a single IMRAD section including introduction (I), or result (R) remain a minority (3.3% + 0.1% = 3.4%). If such incompatible structural patterns were prevalent, using the SAG system on Turkish LIS articles would be considered inappropriate.

The implications of incompatible structural orders in Turkish LIS articles, particularly those without a method section (I,R,D) (5.9%) or with only an introduction section (I) (3.3%), are worth examining to determine whether they are a domain-specific format or a sign of incomplete content. Having only two IMRAD sections is also worth examining. We defer discussion of these implications to future work, as they are beyond the scope of the present study.

As a result, our corpus reflects the implications of this on the feasibility of extracting automatic structural abstracts.

Figure 3 presents boxplots comparing the readability scores of different groups (original abstracts, full-text articles, ASAs, and ACAs) within the corpus. The area between the red horizontal line (y = 29) and the black horizontal line (y = 49) limits the "difficult" area in the graphic depending on the readability scale. The area below the red line indicates "very difficult" readability and the area above the black line indicates "medium difficulty" readability levels. The collection of original abstracts produced by the author is located at the bottom of Fig. 3 which is almost entirely classified as "very difficult". The full-texts are clearly limited within the "difficult" readability range. The majority of ACA, ASA, and average of the readability values of these texts appear in the "difficult" readability area.

As can be seen in Fig. 3, an important finding is a divergence between abstracts and full-texts depending on their readability levels. Author abstracts have a "very difficult" readability level on a corpus basis, while the respective full-texts have a level of "difficult" readability. The readability level of the automatic abstracts produced by the SAG is found between the original abstract and the full-text, and they have almost the same readability level as the full-texts. In addition, there is no statistically significant correlation between the ARV values of abstracts and those of the full text (r = 0.18) (Fig. 4). Therefore, it can be stated that the authors did not show a similar approach in terms of factors that will affect the readability of the full-text and abstracts of their articles. This finding supports the view that the authors deliberately choose difficult-to-read language features when writing abstracts.

It has been seen that the full-text and ASAs distributions based on all IMRAD schemes are proportionally quite similar to each other in Figs. 2 and 5b. Since ASAs take into

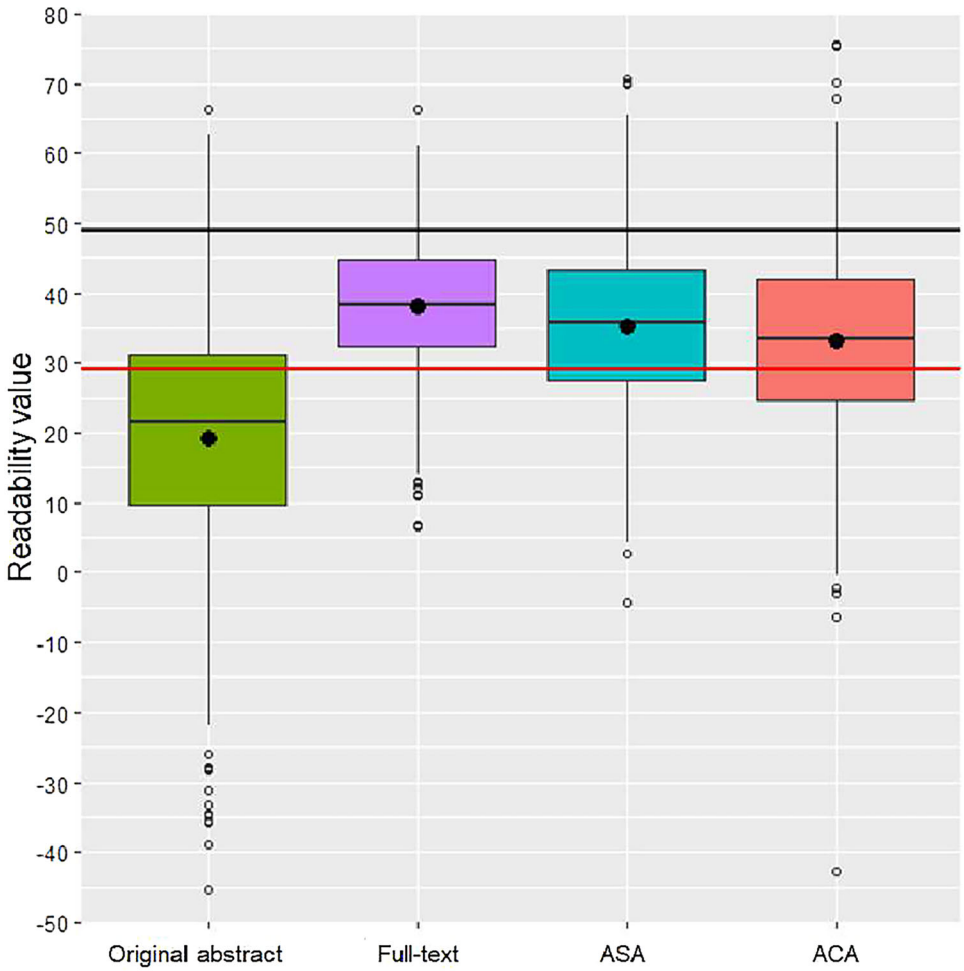

Fig. 3 Readability boxplots of abstract, full-text, ASA, ACA

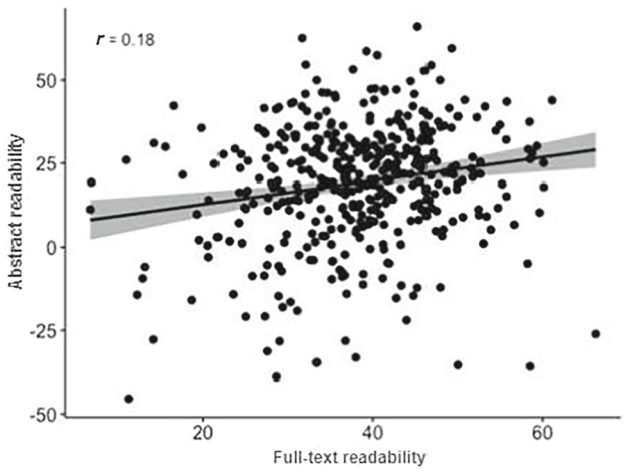

Fig. 4 Correlation between the ARV values of abstracts and those of the full-text

account the structural sections of the full-text when selecting sentences, it is not surprising that the system's structural outputs also reflect the well-structured order of the full-text. However, this graph reveals that, in terms of the amount of structural content in the corpus, the sentence weights of articles with all four IMRAD sections should be represented equally in abstracts. It also suggests that the corpus, which consists of articles selected from the field of LIS and produced with classical abstracts, is actually suitable for structural abstracting.

On the other hand, Fig. 5a, which gives the distribution of ACA sentences based on the structural format, differs clearly both from Figs. 3 and 5b. The weight of the output sentences taken from the articles that have the pattern of four IMRAD sections for ACA is found to be 41.6% (= 1.3% + 17.5% + 12.6% + 10.2%). Only 1.3% of all ACA sentences in the articles with four IMRAD sections consist of four IMRAD sections themselves. Articles with four IMRAD sections account for 17.5% of the ACA sentences in this group, 12.6% for two IMRAD sections, and 10.2% for a single IMRAD section.

When the IMRAD section numbers of ACA's and the IMRAD patterns of full-texts are examined together within Figs. 5b, 3, it is seen that they can have the same IMRAD section numbers as the full-texts with only "I" or "R" IMRAD patterns. For these two relatively small groups, it is not possi-

ble to choose sentences from another structural section. Thus, it has been demonstrated that ACAs are far from being fully compatible with the structural order of full-texts.

Figure 6 shows a graph that displays the distribution of sentences based on their output type and IMRAD patterns. The x-axis represents the abstract type, while the y-axis represents the IMRAD label. The grids at the top show the relationships between different groups of outputs based on the count of IMRAD sections, while the right outer edges of the figure show the relationships between different groups formed based on the IMRAD pattern of the related articles. The labels on the right outer edge represent the abbreviation of IMRAD pattern in the source articles, and the numbers at the top indicate the count of IMRAD in each output group. Each point in the graph shows the distribution of automatic abstract sentences based on the IMRAD count of each output group and IMRAD patterns of the articles from which they are produced.

The grids on the top and right side of the Fig. 6 show how the outputs are grouped based on the number of IMRAD sections and IMRAD pattern, respectively, helping to examine the full-text representativeness between these groups. The projection of each point on the x-axis determines the type of automatic summary in which the relevant sentence is from. Figure 6 displays the distribution of sentences to each output type and IMRAD section.

The distribution of ACAs and ASAs in full-text sentences, as shown in Fig. 6, indicates that they are completely different. ACAs are generated without considering the IMRAD structure of the full-text, while ASAs are generated from each IMRAD section. This results in the count of IMRAD sections in ACAs being independent of the count of IMRAD sections in the full-text. For example, ACAs from full-texts with two (I,M), three (I,M,R), and four (I,M,R,D) IMRAD sections may consist of a single (I) IMRAD section.

On the other hand, ASAs are compatible with the full-text and output patterns since they are generated by selecting relevant sentences from the full-text for a specific IMRAD section.

The content-based performance of the SAG is evaluated with n-gram co-occurrences between the system outputs and ideal summaries by ROUGE 2.0 package. At this stage, we used the original summaries as the ideal summaries. It should be noted that the abstracts are relatively short texts that may limit the overlap between the author's abstracts and the system outputs. On the other hand, the difference between the author's abstracts and the system outputs may be due to meaning and content, or synonymous words and concepts. Evaluating synonyms in automatic summarization is a difficult task as different synonyms can have different meanings and a word's meaning can change based on the context in which it is used. Since our study focuses on structural layouts that influence the performance of automatic summarization systems, we have limited our scope to exclude the evaluation of synonyms. As a result, synonyms were not evaluated in the study.

Table 5 shows the mean F-score values for each ROUGE measure, grouped by the count of IMRAD sections in the articles in the corpus. The line labeled "All" refers to the

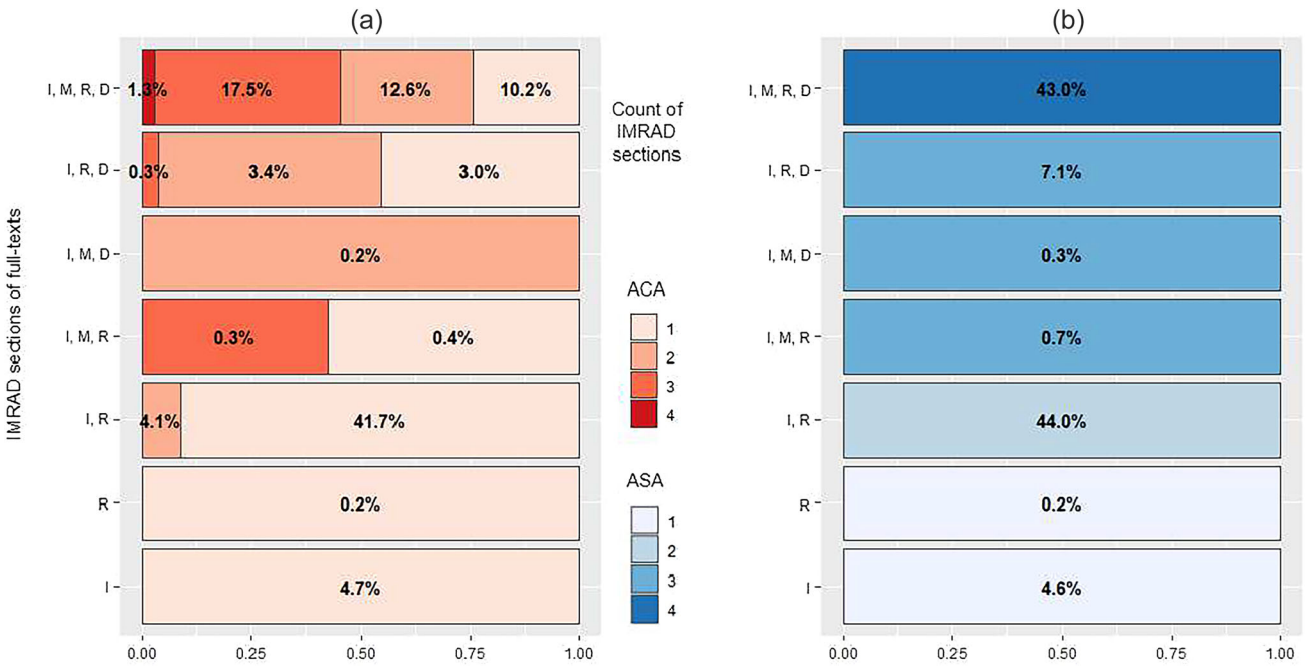

Fig. 5 Distribution of ACA (**a**), ASA (**b**) sentences according to the structural formats determined in the corpus. The color code darkens as the count of sections that are compatible with IMRAD increases

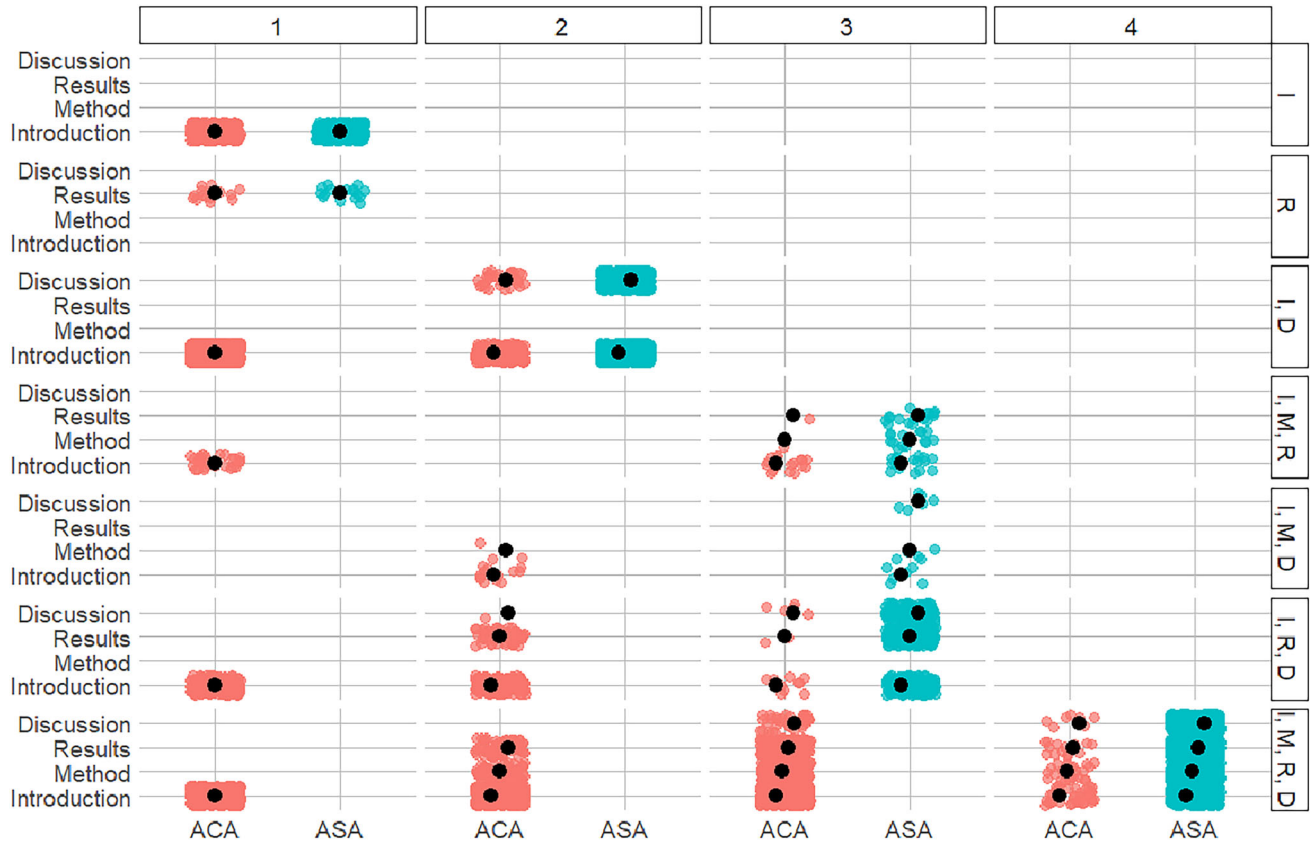

Fig. 6 ACA, ASA distribution of IMRAD patterns by their respective full-text IMRAD patterns which are presented on the right-side vertical edge. The numbers above indicate count of IMRAD sections of each output group

Table 5 Average F-score of ROUGE measures according to count of full-text IMRAD sections

Count of IMRAD	ROUGE-1		ROUGE-2		ROUGE-L		ROUGE-SU4	
	ACA	ASA	ACA	ASA	ACA	ASA	ACA	ASA
4	0.34480	0.38540	0.11080	0.13619	0.06021	0.06687	0.15076	0.17809
3	0.31640	0.32330	0.10244	0.10776	0.06300	0.05719	0.13947	0.14524
2	0.29663	0.31421	0.09198	0.10083	0.05016	0.05471	0.12808	0.13795
1	0.22790	0.22790	0.06265	0.06265	0.05639	0.05639	0.10350	0.10350
All	0.31659	0.34250	0.09967	0.11493	0.05561	0.06014	0.13733	0.15392

values without grouping the corpus based on IMRAD count. The mean F-score is consistently highest for the count of four IMRAD section groups compared to all other output groups. Additionally, ASAs performed better than ACAs for all F-scores at both four and two IMRAD sections, which are the dominant IMRAD patterns in the corpus.

The highest values of n-gram overlapping with the authors' abstracts are the ROUGE-1 in all cases. It is also suggested for very short outputs, such as abstracts of scientific articles, that ROUGE-1 alone may be sufficient for evaluating text quality [44]. The lower values of n-gram overlapping with the abstracts are those in the ROUGE-L. The ROUGE-L deals with the sentence-level structure similarity and identifies the longest string of n-gram associations that occur among the texts it compares. Therefore, it can be argued that short outputs and authors' abstracts may affect the size of the n-gram association sequences between the sentences. The overall decrease in ROUGE-L scores can also be explained in this way.

In Fig. 7 the results of content-based evaluation are presented. Since the majority of articles in the corpus had two or four IMRAD sections, the performance of the dominant group was compared to better illustrate the effect of IMRAD count on output. The boxplots in each section show the F-scores of the developed system outputs, based on the ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-SU4

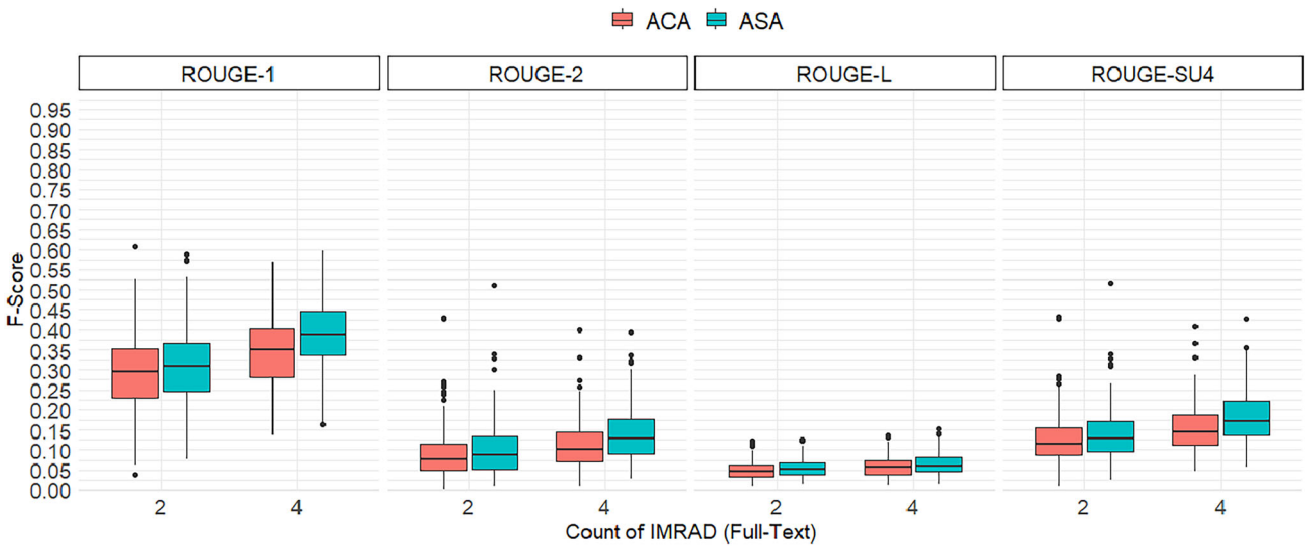

Fig. 7 Boxplots of F-scores of the developed system outputs, according to the ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-SU4

scores. The distributions of the ACA and ASA output groups show similar characteristics in all four score types. It is understood from the graphs that as the count of IMRAD sections in the full-texts increases, the ROUGE scores of both output groups of SAG also increase, and the ASAs have better performances in contrast to the ACAs in all cases.

6 Discussion and conclusion

In this paper, we introduced a Structured Abstract Generator which depends on a simple model for generating high-quality structured abstracts of scientific articles. The purpose of employing such automated methods in extracting abstract sentences from relevant full texts while considering the article structure was to improve the quality of the abstracts. Our system generates structured abstracts (ASAs). We evaluated the impact of considering structural features on the performance of an extractive-based automatic text summarization system with automatically generated classical abstracts (ACAs) without using structural features.

We also present a database that enables the efficient processing of the corpus of 421 Turkish LIS articles in full-text sentences where each component is assigned to the corresponding structural section of the document, as well as the document's metadata.

First, we explored any factors that could prevent the creation of structured abstracts and showed that our corpus is formatted in a way that enables the automatic generation of structured abstracts. 89.8% of the sentences in our corpus come from articles with an acceptable IMRAD pattern of all four (43.5%) IMRAD sections, or at least two (46.3%) IMRAD sections (Introduction and Discussion). Further research is needed to determine whether having only two IMRAD sections is a domain-specific format or a sign of incomplete content. The other problematic articles were completely incompatible with academic writing are remained in the minority. Our study only examined article structural arrangements with a focus on the sentence selection processes. We leave in-depth studies of articles with missing sections in their structural order according to IMRAD for future work.

Second, the readability levels of the full-texts of articles published in the field of Turkish LIS were calculated, and the corpus was largely classified as "difficult" according to the readability scale. However, the readability value of the abstracts produced by the same authors was significantly at the "very difficult" level. We observed that authors deliberately choose difficult-to-read language features in their abstracts, regardless of the language features they use in full-texts. Both ACA and ASA abstracts were calculated at the same readability level as full-text articles showing that selecting important sentences from full-text articles to generate automatic abstracts improves readability. Despite the reasons that lead authors to write difficult-to-read abstracts, widespread use of tools to select important sentences from the structural sections of full-texts may help to break this habit, which hinders scientific communication, over time.

After assessing the quality of SAG outputs, we found that having a well-organized full text improves the quality of both two output groups of SAG. It was observed that ASAs performed significantly better than ACAs. However, interestingly, ACAs also performed better as the number of structured sections increased, despite being produced without taking into account the structure of the full-text. This could be due to an increase in the structured content of original abstracts, resulting in greater similarity between structured

and non-structured automatic abstracts and author abstracts. Alternatively, in the context of information retrieval, it means that authors can produce abstracts that convey information more accurately and have higher recall and precision scores when full-texts structural layout improves. We conclude that it is possible to argue that focusing on structural writing in full-texts alone can contribute to improving the content of the original abstracts produced by the author.

In the near future, we can expect to see various systems such as LLMs (Large Language Models), knowledge graphs, NER (Named Entity Recognition systems) systems, QA (Question Answering) systems, MT (Machine Translation) systems, and text summarization systems being used together to produce high-quality structured abstracts. We may also see the emergence of new tools that are specifically designed to assist researchers in communicating their findings more effectively.

Future research should explore more efficient and effective features for automatic summarization methods to generate summaries of scientific records in different languages and domains. Additionally, future research should investigate how the structure of the full-text can be further optimized to improve the quality of automatic summarization methods. Training domain-specific dictionaries would help to improve the accuracy, readability, and effectiveness of generated abstracts. We plan to train a model to classify structural sections of Turkish articles by employing our data for future research. Thus, we can fully automate the process of producing structured abstracts by learning systems. Different summarization approaches and algorithms should be applied to obtain more readable, high-quality structured abstracts. We also plan studies to train our data to predict the structural order of abstracts. A detailed analysis of user opinions on the readability issue can also be conducted. User studies can also reveal the best sentence weights depending on the structural sections of articles.

Finally, we verified that using structural sentence selection, abstract-generating systems can support scholarly communication as a supplementary tool for authors and editors.

Acknowledgements This article is based on Özkan Çelik's [50] Ph.D. dissertation and was supported in part by a research grant from The Scientific and Technological Research Council of Türkiye (Project No: SOBAG 115K440) [51].

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK).

Data availability Data are available at: https://github.com/esraozzz/SAG/.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Dowling, M., Hammami, H., Tawil, D., Zreik, O.: Writing energy economics research for impact. Energy J. (2021). https://doi.org/10.5547/01956574.42.3.mdow
2. Fages, D.M.: Write better, publish better. Scientometrics **122**(3), 1671–1681 (2020). https://doi.org/10.1007/s11192-019-03332-4
3. Day, R.A.: Bilimsel Makale Nasıl Yazılır Ve Yayımlanır? [How to Write and Publish a Scientific Paper?]. TÜBİTAK, Ankara (1996)
4. Gazni, A.: Are the abstracts of high impact articles more readable? Investigating the evidence from top research institutions in the world. J. Inf. Sci. **37**(3), 273–281 (2011). https://doi.org/10.1177/0165551511401658
5. Hartley, J., Pennebaker, J.W., Fox, C.: Abstracts, introductions and discussions: How far do they differ in style? Scientometrics **57**, 389–398 (2003)
6. Jamar, N., Šauperl, A., Bawden, D.: The components of abstracts: The logical structure of abstracts in the areas of materials science and technology and of library and information science. New Libr. World **115**(1/2), 15–33 (2014). https://doi.org/10.1108/nlw-09-2013-0069
7. Dewan, P.: Are books becoming extinct in academic libraries? New Libr. World **113**(1/2), 27–37 (2012). https://doi.org/10.1108/03074801211199022
8. Meadows, A.J.: The scientific paper as an archaeological artefact. J. Inf. Sci. **11**(1), 27–30 (1985). https://doi.org/10.1177/016555158501100104
9. Carr, N.: Is Google Making Us Stupid? Yale University Press, New Haven (2009). https://doi.org/10.12987/9780300156508-009
10. Issa, T., Isaias, P.: Internet factors influencing generations Y and Z in Australia and Portugal: a practical study. Inf. Process. Manag. **52**(4), 592–617 (2016). https://doi.org/10.1016/j.ipm.2015.12.006
11. Merzenich, M.: Going Googly - "On the Brain" with Dr. Michael Merzenich. http://onthebrain.com/2008/08/going-googly/. Accessed 14 Jun 2023
12. Singer, L.M., Alexander, P.A.: Reading on paper and digitally: What the past decades of empirical research reveal. Rev. Educ. Res. **87**(6), 1007–1041 (2017). https://doi.org/10.3102/0034654317722961
13. Wästlund, E.: Experimental Studies of Human-computer Interaction: Working Memory and Mental Workload in Complex Cognition. Department of Psychology, Göthenburg (2007)
14. Liu, Z.: Reading in the age of digital distraction. J. Doc. **78**(6), 1201–1212 (2021). https://doi.org/10.1108/jd-07-2021-0130
15. Atanassova, I., Bertin, M., Mayr, P.: Mining scientific papers for bibliometrics: A (very) brief survey of methods and tools. arXiv preprint arXiv:1505.01393 (2015)
16. Mabe, M.A., Amin, M.: Dr Jekyll and Dr Hyde: author-reader asymmetries in scholarly publishing. ASLIB Proc. **54**(3), 149–157 (2002). https://doi.org/10.1108/00012530210441692

17. Nicholas, D., Huntington, P., Jamali, H.R.: The use, users, and role of abstracts in the digital scholarly environment. J. Acad. Librariansh. **33**(4), 446–453 (2007). https://doi.org/10.1016/j.acalib.2007.03.004
18. Plavén-Sigray, P., Matheson, G.J., Schiffler, B.C., Thompson, W.H.: The readability of scientific texts is decreasing over time. eLife (2017). https://doi.org/10.7554/elife.27725
19. Wang, S., Liu, X., Zhou, J.: Readability is decreasing in language and linguistics. Scientometrics **127**(8), 4697–4729 (2022). https://doi.org/10.1007/s11192-022-04427-1
20. Atanassova, I., Bertin, M., Larivière, V.: On the composition of scientific abstracts. J. Doc. **72**(4), 636–647 (2016). https://doi.org/10.1108/jdoc-09-2015-0111
21. Bitri, E., Keseroğlu, H.S.: Türk kütüphaneciliği ve bilgi dünyası dergilerinin özlerine eleştirel bir bakış [A critical view to abstracts of Turkish Librarianship and Information World Journals]. Türk Kütüphaneciliği [Turkish Librarianship] **29**(2), 241–257 (2015)
22. Šauperl, A., Klasinc, J., Lužar, S.: Components of abstracts: Logical structure of scholarly abstracts in pharmacology, sociology, and linguistics and literature. J. Am. Soc. Inform. Sci. Technol. **59**(9), 1420–1432 (2008). https://doi.org/10.1002/asi.20858
23. Hartley, J., Betts, L.: The effects of spacing and titles on judgments of the effectiveness of structured abstracts. J. Am. Soc. Inform. Sci. Technol. **58**(14), 2335–2340 (2007). https://doi.org/10.1002/asi.20718
24. Sharma, S., Harrison, J.E.: Structured abstracts: Do they improve the quality of information in abstracts? Am. J. Orthod. Dentofac. Orthop. **130**(4), 523–530 (2006). https://doi.org/10.1016/j.ajodo.2005.10.023
25. DuBay, W.H.: The Principles of Readability. ERIC Clearinghouse, Costa Mesa, CA. (2004). https://books.google.com.tr/books?id=Aj0VvwEACAAJ
26. Ufnalska, S., Hartley, J.: How can we evaluate the quality of abstracts. Eur. Sci. Ed. **35**(3), 69–72 (2009)
27. Meadows, A.J.: Communicating Research. Academic Press, New York (1998)
28. Luhn, H.P.: The automatic creation of literature abstracts. IBM J. Res. Dev. **2**(2), 159–165 (1958). https://doi.org/10.1147/rd.22.0159
29. Altmami, N.I., Menai, M.E.B.: Automatic summarization of scientific articles: a survey. J. King Saud Univ. Comput. Inf. Sci. **34**(4), 1011–1028 (2022). https://doi.org/10.1016/j.jksuci.2020.04.020
30. Vilca, G.C.V., Cabezudo, M.A.S.: A study of abstractive summarization using semantic representations and discourse level information. In: Ekštein, K., Matoušek, V. (eds.) Text, Speech, and Dialogue, pp. 482–490. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-64206-2_54
31. Moratanch, N., Chitrakala, S.: A survey on extractive text summarization. In: 2017 International Conference on Computer, Communication and Signal Processing (ICCCSP), pp. 1–6. IEEE, Chennai, India (2017).https://doi.org/10.1109/icccsp.2017.7944061
32. Mridha, M.F., Lima, A.A., Nur, K., Das, S.C., Hasan, M., Kabir, M.M.: A survey of automatic text summarization: Progress, process and challenges. IEEE Access **9**, 156043–156070 (2021). https://doi.org/10.1109/access.2021.3129786
33. Baykara, B., Güngör, T.: Abstractive text summarization and new large-scale datasets for agglutinative languages Turkish and Hungarian. Lang. Resour. Eval. **56**(3), 973–1007 (2022). https://doi.org/10.1007/s10579-021-09568-y
34. Tsonkov, T., Lazarova, G.A., Zmiycharov, V., Koychev, I.: A comparative study of extractive and abstractive approaches for automatic text summarization on scientific texts. In: ERIS, pp. 29–34 (2021)
35. Güran, A., Arslan, S.N., Kılıç, E., Diri, B.: Sentence selection methods for text summarization. In: 2014 22nd Signal Processing and Communications Applications Conference (SIU). IEEE, Trabzon, Turkey (2014). https://doi.org/10.1109/siu.2014.6830198
36. Song, N., Chen, K., Zhao, Y.: Understanding writing styles of scientific papers in the IS-LS domain: evidence from abstracts over the past three decades. J. Inform. (2023). https://doi.org/10.1016/j.joi.2023.101377
37. Akın, A.: Zemberek-NLP, Natural Language Processing Tools for Turkish. (2018). https://github.com/ahmetaa/zemberek-nlp
38. Tunali, V., Bilgin, T.T.: Türkçe metinlerin kümelenmesinde farklı kök bulma yöntemlerinin etkisinin araştırılması [Examining the impact of different stemming methods on clustering Turkish texts]. In: ELECO'2012 Electric-Electronic and Computer Engineering Symposium, pp. 598–602 (2012)
39. Binwahlan, M.S., Salim, N., Suanmali, L.: Fuzzy swarm diversity hybrid model for text summarization. Inf. Process. Manag. **46**(5), 571–588 (2010). https://doi.org/10.1016/j.ipm.2010.03.004
40. Erkan, G., Radev, D.R.: LexRank: graph-based lexical centrality as salience in text summarization. J. Artif. Intell. Res. **22**, 457–479 (2004). https://doi.org/10.1613/jair.1523
41. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manag. **24**(5), 513–523 (1988). https://doi.org/10.1016/0306-4573(88)90021-0
42. Lin, C.-Y.: ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (2004). https://aclanthology.org/W04-1013
43. Saggion, H., Radev, D.R., Teufel, S., Lam, W., Strassel, S.M.: Developing infrastructure for the evaluation of single and multi-document summarization systems in a cross-lingual environment. In: LREC, pp. 747–754 (2002)
44. Ganesan, K.: Rouge 2.0: updated and improved measures for evaluation of summarization tasks. arXiv preprint arXiv:1803.01937 (2018)
45. Mani, I., House, D., Klein, G., Hirschman, L., Firmin, T., Sundheim, B.: The TIPSTER SUMMAC text summarization evaluation. In: Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, USA (1999). https://doi.org/10.3115/977035.977047
46. Crochemore, M., Rytter, W.: Text Algorithms. Oxford University Press, Oxford, UK (1994)
47. Özdemir, E.: Eleştirel Okuma [Critical Reading]. Bilgi Publishing, Ankara (2000)
48. Flesch, R.F.: A new readability yardstick. J. Appl. Psychol. **32**(3), 221–233 (1948). https://doi.org/10.1037/H0057532
49. Ateşman, E.: Türkçede okunabilirliğin ölçülmesi [Measuring readability in Turkish]. Dil Dergisi [J. Lang.] 58, 71–74 (1997)
50. Çielik, A.E.: Türkçe akademik yayınlar için yapısal öz çıkarım sistemi [Structured abstract extraction system for Turkish academic publications]. PhD Thesis, Hacettepe University (2021)
51. Al, U., Sezen, U.: Türkçe atıflar için içerik tabanlıatıf analizi modeli tasarımı [Designing a model for content-based citation analysis for Turkish citations]. TÜBİTAK Sosyal Bilimler Araştırma Grubu-Proje No: SOBAG 115K440). Hacettepe Üniversitesi Bilgi ve Belge Yönetimi Bölümü[Hacettepe University Department of Information Management] (2018)