

## Optimizing resume information extraction through TSHD segmentation and advanced deep learning techniques

Anmar Abuhamdah<sup>1</sup>, Mohammed Al-Shabi<sup>1</sup>, Sana Jawarneh<sup>2</sup>

<sup>1</sup>Department of Management Information Systems, College of Business Administration, Taibah University, Madinah, Saudi Arabia

<sup>2</sup>Department of Computer Science, Applied College Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia

### Article Info

#### Article history:

Received Mar 22, 2025

Revised Jul 8, 2025

Accepted Oct 15, 2025

#### Keywords:

Deep learning

Information extraction

Natural language processing

Textual

Topic segmentation

Transformer models

### ABSTRACT

This research focuses on a significant factor in the natural language processing area, which is extracting information from unstructured textual data through efficient methods in order to pull useful insights and structured representations from this data. This research attempts to boost the effectiveness of information retrieval systems through computational analysis. This paradigm is explored in this work using question answering models in an extractive style, a modern information extraction approach, creating a new methodology combining the topic segmentation based on headings detection (TSHD) segmentation algorithm and deep learning methods. The TSHD algorithm breaks documents into sections in which certain topics are addressed. Refined extraction models are then used to process these disjoint segments leading to more accurate and context-judicious extraction compared to naive whole-document extraction approaches. We empirically validate this approach using the stanford question answering dataset (SQuAD) 1.1 dataset, with a specific adaptation to resumes. Experimental results show that the performance metrics increase by 7.4% in exact match (EM) and by 7.8% in F1-score. This can be concluded from these results illustrating the feasibility of the proposed approach in the automated information extraction frameworks such as resume processing.

*This is an open access article under the [CC BY-SA](#) license.*



### Corresponding Author:

Anmar Abuhamdah

Department of Management Information Systems, College of Business Administration

Taibah University

Al-Madinah, Al-Munawwarah, 42353, Medina, Kingdom of Saudi Arabia

Email: aabuhamdah@taibahu.edu.sa

## 1. INTRODUCTION

Modern information age is witnessing rapid growth of unstructured textual data, thus making textual information extraction a fundamental process in various domains including but not limited to scientific research, health care, business and education. This explosive growth underscores the critical role that natural language processing (NLP) plays in shaping unstructured data into a structured view. On the other hand, even in the era of state-of-the-art transformer-based models like BERT, context specific information extraction is difficult due to the inherent ambiguities and noise in real world texts [1]. This explosion of unstructured data highlights the necessity for sophisticated techniques, to glean actionable insights from it rapidly. That said, the complexities of natural language understanding are a significant barrier to accurate processing of nuanced or context-specific information. In spite of the recent surge in performance across various domains within NLP brought forth by advances in newly configured transformer-based models such as bidirectional encoder representations from transformers (BERT) and its respective derivatives [1], [2] due to

their extreme success in fields such as machine translation, question answering (QA) or information retrieval, such discovery methods still lack specificity towards the context.

Information extraction from unstructured text to structured data has always been a core problem in NLP, every now and then flooding a range of new text and data into the ever-growing digital space. A number of methods have been proposed through the years, ranging from traditional rule-based approaches to illage-based methods based on deep learning. Initially, early techniques were based on the manually crafted rules and statistical models [2], which have shown a potential but were limited regarding their scalability and their adaptability to complex real-world documents like resumes. These models have been effectively exploited in bio-medical text mining and legal document analysis. But more complex and diversified structures of text are less likely to be found in them, especially in unrestricted documents due to domain in specific vocabulary and polysemous words.

In the area of resume analysis, a common problem is to correctly extract appropriate information while excluding irrelevant information. For instance, in the task of extracting mentions of educational institutions from resumes, models would here and there wrongly bring out non-target entities, such as employer universities or publication venues. This phenomenon is related to vagueness in the definition of context boundaries and it may affect precision for information extraction systems.

Our main contribution is to improve the accuracy and the efficiency of resume content extraction by combining topic-aware segmentation and deep learning-based extractive question answering (EQA). Our approach leverages transformer-based QA models with the topic segmentation based on headings detection (TSHD) segmentation algorithm to segment the cognitively coherent segments and retrieve the relevant information without retraining the models.

Two fundamental limitations characterize conventional information extraction approaches, which indiscriminately process whole documents. The first, they have a tendency of throwing the precision out of the window since they are burdened with noise from irrelevant sections which obscure meaningful insights. Second, these methods are often computationally inefficient, especially on long documents, because they waste resources analysing text that is unlikely to contribute to the output. Such indeed are the weaknesses of previous works, which underscore the need for more targeted and efficient solutions to the task that constrain relevant segments of a document and minimize redundancy.

In order to work around the mentioned issues, this model proposes a new approach using the TSHD algorithm. TSHD has two steps: segmenting documents into semantically coherent sections, and using semantic matching to identify the topical relevance of those sections to specific user queries; this allows for precise and efficient extraction of relevant information. This not only helps in filtering out noise from unrelated parts but also enhances the use of computational resources, making it especially suitable for handling large-sized complex papers [3]. In recent research, the capabilities of segmentation techniques in improving information extraction models has been further confirmed; demonstrating that with segmentation, deep learning models may be educated to extract out only the information in segments of a document, thus reaching a balance between high accuracy and low resource consumption in tasks [4]–[10].

Information extraction systems have become much more accurate and efficient with the recent advances in deep learning. Models like Bert, Roberta, and Longformer have shown state of the art performance in problems like named entity recognition and EQA, especially in the case of structured or semi-structured documents. Transformers-based models like RoBERTa [11], BERT [2], [12], Longformer [1], and XLNet [13] realized some of the state-of-the-art solutions for many NLP tasks, such as document classification, sentiment analysis, or question answering. These models exploit attention mechanisms and can model long-distance dependencies and relations in text. For example, BERT Devlin *et al.* [3] for the first time, achieved state-of-the-art results on various NLP tasks, through large-scale pre-training and fine-tuning on task-specific data. RoBERTa and Longformer are post-Transformer architectures that resolved key performance issues raised by the original architecture (scalability, computational efficiency).

One of the clearest uses of transformer models is in EQA, where the objective is to extract relevant spans of text answering the question. EQA has transformed document retrieval systems by allowing for more fine-tuned extraction of information from structured or semi-structured text. The development of supervised models has heavily relied on benchmark datasets, e.g., the stanford question answering dataset (SQuAD), leading to great progress in EQA. SQuAD 1.1 [14] has become the standard for evaluation and training with 107,785 questions, and the (single) answer being guaranteed to exist in the provided context. SQuAD 2.0 [15], which contains 53,775 unanswerable questions, has added an additional layer of difficulty as models must now extract an answer or determine that a question does not have a valid answer. This has resulted in the creation of better models and RoBERTa has even better performance than BERT due to the improvements it brings. For instance, one model achieves an F1-score of 85.8 on SQuAD 2.0 but achieves an F1-score of only 66.3 on SQuAD 1.1, demonstrating the task's added complexity. Along with SQuAD, other datasets, for instance, NewsQA, QUAC, and CovidQA have been remarkably beneficial in training and

evaluating EQA models [16]–[18]. While QUAC has 14,000 dialogues with 100,000 questions in a form where a student asked Wikipedia texts and the teacher answers in short text excerpts, NewsQA includes 119,633 questions based on 12,744 news articles. CovidQA, by contrast, consists of 2,019 question-answer pairs drawn from 147 articles about COVID-19. F1-score is still one of the main metrics for evaluating models trained on such datasets like in Table 1 [19]. Table 1 shows a comparison of the F1-score of a variety of transformers on several datasets which shows performance differences on the basis of training size, question clarity and level of context complexity.

Table 1. Comparative performance of transformer-based models across datasets (F1-score)

Model	SQuAD 2.0	NewsQA	QUAC	CovidQA
XLNetBASE	64.9	53.2	30.1	44.9
BERTBASE	64.7	52.1	28.6	44.8
ROBERTABASE	68.2	57.0	31.3	44.5
ALBERTBASE	64.8	51.8	19.5	42.4
ConvBertBASE	67.4	55.7	31.5	44.9
BARTBASE	67.6	56.2	29.1	45.3
BERTBASE-BILSTM	65.0	52.6	28.9	45.6

In fact, most of the efforts in this area are based on models that operate at the document level to extract information. However, papers, drug leaflets, Wikipedia articles, and resumés typically include distinct parts that address specific matters [20]–[22]. Different evaluation results can be caused by (i) being affected by training set size, (ii) simplicity of the question-and-answer formulations and (iii) textual context complexity [20]. For instance, datasets which are well-structured, such as SQuAD 2.0, attain a much larger F1-score (85.8) compared to SQuAD 1.1 (66.3), indicating that the structure of the document has an impact on the performance of the model.

To alleviate the difficulty in processing long documents, greater models like longformer have shown promising results in text-based answer extraction tasks by using a scalable attention formulation [11]. 6RoBERTa RoBERTa is a pioneer in this category – it takes the classic BERT architecture and performs some massive upgrades on it. But these models are resource-intensive to be able to use on a resource-limited device. For computationally efficient alternatives, DistilBERT is a distilled variant of BERT which provide large speedups and reduced size with little quality loss [20].

However, these models are still not suitable for unstructured or overlooked data, indicating the requirement of novel pre-processing methods. Topic segmentation and identification can help the information extraction and summary tasks, because they allow models to focus on the relevant periphery. High quality segmentations are therefore needed to process natural language tasks effectively in these regimes. These trade-offs show that models have to be selected according to the application-specific requirements [3], [23]. Cross-document segmentation is a technique for splitting up a document into smaller, more manageable pieces, to address reading complexity issues faced with larger, complex documents.

One such recent algorithm belongs to this general framework: the TSHD that divides documents into coherent sections by topic. For retrieval task TSHD can achieve higher performance since it retains those segments whose meaning is coherent and consistent. This allows a more focussed and structured extraction method to proceed other than just processing whole documents. This titling of work is illustrated by TSHD for resume processing and legal document processing. For instance, TSHD achieved an F1-score of around 96% and a very low segmentation error rate of approximately 2% [23] thus, in this sense it can be seen as a useful tool to improve information extraction models. This variability is reduced when using TSHD algorithm on resume texts as the algorithm segments documents to topic coherent subsets reducing the effect of the unstructured nature of the text and the presence of domain dependent jargon, thereby letting models like Longformer to focus on improved window of context, rather than the unstructured raw text.

Several literature works has emphasized on automated extraction of useful data from resumes using NLP applications in the field of resume analysis. Methods such as keyword extraction, entity recognition, and syntactic parsing have been applied for identifying information such as skills, qualifications and experience. However, such simple strategies can face issues in a real scenario context where the resumes are high dimensional and unstructured or there may be advanced bio domain knowledge in the resume. This has led to the recent effort of integrating segmentation algorithms such as TSHD with powerful deep learning models that not only help to improve the accuracy of the resume information extraction, but also demonstrate a high generalization ability for the out-of-domain examples.

This research presents a new framework combining the TSHD segmentation algorithm with DL-based extractive QA models to improve the quality of information extraction from resumes. Such a hybrid method benefits from the advantages of both segmentation and question-answering approaches since it

ensures a more targeted and context-relevant manner of data extraction. Based on the results of our experimental study, which involved the use of the SQuAD 1.1 in the context of a resume, it is possible to conclude the feasibility of the proposed framework, as evidenced by the increase in exact match (EM) scores and F1-scores.

This research is specifically focused on resumes, which are among the building blocks of hiring. Resumes contain different sections (e.g., education, professional experience, skills), which require precise extraction mechanisms in order to ensure reliable candidate evaluation. Thus, the main purpose of this work is to establish a unified framework merging transformer models with TSHD based topic segmentation. The segmentation in this framework reduces information extraction to processing contextually small relevant segments only without the requirement of re-training models for document level data. This data is then formatted into tables and is used in downstream computations making the decision pipelines faster and more efficient. Novel text summarization and extraction frameworks have emerged recently exploiting pre-trained language models, which allow them to produce accurate and efficient output, especially for heterogeneous document sections, e.g., resumes [9], [10].

In this paper, we propose to tackle these challenges with a new framework that combines topic-aware segmentation with transformer-based models. This work extends previous work in the area of document segmentation and deep learning EQA. Prominent works of TSHD for topic-aware segmentation [3], RoBERTa for robust language modelling [11], and Longformer for long-document processing [3] have provided inspiration for our work. However, many previous approaches tend to handle the entire document in a very straightforward manner, which in turn hurts the prediction accuracy and runtime estimation. We build on these with our model, which adds topic-aware segmentation to transformer-based QA models, enabling more accurate context selection and focused extraction. On the other hand, the main novelties of this work are:

- A novel combination of TSHD segmentation algorithm and extractive QA models for resume presentation improvement.
- Empirical evaluation of the proposed method with adapted SQuAD 1.1 questions showing significant performance improvements in both EM and F1.
- Evidence that the computational cost is lessened by context filtering, without necessity to retrain models on segmented data.

These contributions are consistently shown throughout the method, empirical evaluation, and discussion sections in this paper.

This work is motivated by the problem of extracting structured information from unstructured resumes due to the growing requirement of efficient automated recruitment systems. Although NLP has made considerable progress, current methods do not usually deal well with the heterogeneous noisy nature of resume data. In this paper, we attempt to address the above issue by introducing a novel hybrid framework which integrates the TSHD segmentation model with transformer-based QA model. In this paper, we study whether we can enhance the accuracy of extraction by topic-aware segmentation at the low computational cost. The work is close to the focus of the journal, namely advanced computing techniques and their applications in real-life domains, in this case in the area of information extraction and document processing.

The paper is organized as follows: Section 2 presents the proposed approach the hybrid pipeline of transformer based EQA and TSHD. Section 3 contains the results and discussion, in which we evaluate the performance of the proposed framework against some baseline models. Concluding remarks are given in section 4 in which we present a list of merits of the research to the development of efficient and context-oriented information extraction system. References, constraints, implications and future study direction.

## 2. METHOD

In this section we describe the dataset used in this study, the evaluation metrics used, how the TSHD algorithm works, and the information extraction models that have been used to measure the effectiveness of our proposed method. Moreover, a further description of methodological considerations that underpinned the evaluation process to conduct a thorough and robust analysis is documented.

### 2.1. Dataset description

We collect 105 resumes in English from online portfolios, serving as our dataset. These resumes come in all types of formats and have a variety of structures to suit your needs. We manually labeled important fields (i.e., name, telephone number, email, education) to maintain consistency. Since resumes are unstructured text files, they all have different templates and formats in showing information with a significant difference in the number of sections and their order. There is also no consistency of font type, size or colour, which add to the structural non-homogeneity of the dataset. The dataset framework used to assess the proposed information extraction method was based on the SQuAD 1.1, which was introduced in [14].

The assessment consisted of asking certain questions to get the important details out of the resumes, in the way given below:

- Extracting the resume owner's name, and the question: what is my full name?
- Extract the email address, and the question: what is the email address?
- Obtaining the mobile phone number, and the question: what is the mobile phone number?
- Extracting the bachelor's major, and the question: for which major is the bachelor's degree?
- An extractor, which is used to extract the university awarding the bachelor, and the question: from which university bachelor degree?

It should be noted, the 444 questions in the test dataset all have corresponding answers. Now, the SQuAD dataset has the following structured format:

- Id: Unique identifier assigned for each of the question.
- Title: the name of the document file for the resume.
- Context: the text segment that the answer should be extracted from.
- Question: what you need to retrieve this information.
- Answers: correct answer(s), with up to 3 possible valid answers for each question.

The defined systematic approach helps to provide a consistent evaluation methodology for determining the suitability of information extraction models to different and potentially unstructured resume layouts.

## 2.2. Evaluation metrics

We evaluate information extraction models by two widely used metrics EM and F1-score. These evaluation metrics quantify how accurately the model retrieves the answer from textual data [2], [14].

- a) Evaluates whether the answer predicted by the model verbatim matches the ground truth answer. The metric gives a yes/no score:
  - A score of 100 is awarded if the predicted answer is identical to the actual answer. Otherwise, it is scored 0.
  - The overall EM score is calculated by averaging the EM scores over each evaluation question [24].
- b) F1-score metric computes the intersection of words in the candidate answer snippet and the ground truth snippet. It is computed using precision and recall, which can be defined as:
  - Precision: the number of correctly predicted words divided by the number of words in the predicted answer.
  - Recall: number of words that were predicted accurately in comparison with number of words in actual answer.
  - Using the (1), we compute the F1-score.

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (1)$$

Then, the overall F1-score across all evaluation questions is an average of the F1-scores to give a general idea of the model performance. However, this metric allows for a more precise evaluation that takes into account partial correctness of the predicted answer when it contains some, but not all words present in the actual answer.

## 2.3. General mechanism of the TSHD algorithm and its operation stages

TSHD algorithm consists of three main stages to extract information from unstructured text data. These steps are taken to facilitate the systematic performance of segmentation based on pre-processing, heading detection, content extraction, and so on:

- a) The initial step of the algorithm is on pre-processing, which refers to the transformation and refinement of raw textual data for subsequent processing. This phase is comprised of a sequence of successive steps:
  - Text pre-processing: basic transformation of text.
  - Lines tokenization: dividing text into multiple lines for precise analysis.
  - Data pre-processing: cleaning undesirable characters, spaces, extra symbols.
  - Normalization: adjusting text formats to a uniform standard, such as case normalization and font styles.
  - Enumeration of lines: identifying each line of text with unique line numbers.
  - Lines refinement: reformatting of the text in order to maintain coherence and improve segmentation
  - These steps ensure the increased quality of the data overall, leading to lesser discrepancies as the data proceeds to the following phase.
- b) The headings detection stage, at this stage the algorithm finds section headings within the document and brings consistency to the variations of heading terms. This consists of two linear scans one after another:

- Cue words scan: looks for commonly used keywords that are commonly found in section headers (like “Education,” “Experience”)
- Cue phrases scan: identifying multi-word expressions and header variations to combine section names across diverse document types.
- c) By labelling section headings, the algorithm ensures greater consistency in segment identification, which improves the subsequent segmentation performance.
- d) The last stage is segmentation which extracts content under each header and formats it in a machine-readable format. Sections are being extracted as key-value pairs, where each section heading will be the key, and the corresponding content will be the value in JSON format. By structuring the information in such a way, it becomes far more efficient for downstream processes to work with it, allowing for better querying to information and automating the analysis as well.

Instead, reworking documents through these three stages enables the TSHD algorithm to identify textual information more accurately and effectively, especially in circumstances of unstructured data.

## 2.4. Used Information extraction models

The mechanism was implemented and evaluated on three fine-tuned information extraction models trained on resume data using the PyTorch library. The models used in this study are described as follows.

### 2.4.1. Autotrain-resume\_parser model

For model evaluation, we chose three fine-tuned EQA models trained on resumes. The first model we use is the Autotrain-resume\_parser which is built on the Longformer model that can efficiently handle long document. This model is based on the longformer architecture [1], which scales self-attention -- a special case of the multi-head attention mechanism used in the transformer architecture -- linearly with respect to the length of the input sequence, and allows processing long documents more efficiently than regular transformer-based models. The base blocks of the model are visualized in Table 2, modified in a way to obtain maximum performance given to the task of extracting structured information from resumes. In addition, autotrain-resume\_parser model in more low-level side uses a sequence length of 4096 tokens with attention windowing to process long resumes quickly with context intact, as shown in Table 2.

Table 2. Configuration details of the autotrain-resume\_parser model based on longformer architecture

Feature Name	Value	Description
Hidden_act	gelu	Hidden layer activation function
Hidden_size	768	Hidden layer size (embedding dimensions)
Max_position_embeddings	4098	Maximum position embeddings limit
Model_type	longformer	Model type
Num_attention_heads	12	Number of attention heads
Num_hidden_layers	12	Number of hidden layers
Vocab_size	50265	Vocabulary size
Max_length	384	Maximum input length for the model

### 2.4.2. CV\_Custom\_DS model

This is based on [4], RoBERTa: a robustly optimized BERT pre-training approach. Unlike BERT, RoBERTa fixes some of its fractional improvement:

- a) Removal of the next sentence prediction (NSP) task to let the model train only on token-level and sentence-level representations without inter-sentence dependencies affecting representation learning.
- b) Dynamic masking during training, which introduces variability of masked tokens during pre-training for more effective model training.
- c) Further pre-training on larger and longer text sequences to allow this model to learn richer contextual representations and generalize well across a downstream of NLP tasks.

The adopted model has some differences with BERT on several settings (Table 3) to transform it into a model for information extraction tasks [25]. In contrast to BERT, this RoBERTa model eliminates the NSP task so that it can concentrate on the token-level and sentence-level representations without consideration of interference from inter-sentence learning. It also uses dynamic masking during training and pre-trained on larger text sequences which improves its context understanding and generalization across varied resume layouts. Table 3: CV\_Custom\_DS model our CV\_Custom\_DS model builds on BERT by eliminating the NSP task and using customizable dynamic masking during training to improve generalization on different resume formats.

Table 3. Architectural and training configurations of the CV\_Custom\_DS model based on RoBERTa

Feature name	Value	Description
Hidden-act	Gelu	Hidden layer activation function
Hidden_size	768	Hidden layer size (embedding dimensions)
Max_position_embeddings	514	Maximum position embeddings limit
Model_type	Roberta	Model type
Num_attention_heads	12	Number of attention heads
Num_hidden_layers	12	Number of hidden layers
Vocab_size	50265	Vocabulary size

#### 2.4.3. AQQ\_CV\_Squad model

AQQ\_CV\_Squad model: we base the model on DistilBERT, which is a lightweight version of BERT that was developed for low-resource contexts [16]. It provides faster speed and lighter computational burden which make it possible to be deployed in resource-limited applications. However, this optimization comes at the expense of a slight decrease in accuracy when compared to its larger siblings such as RoBERTa. Some particular details of the used model configuration are given in Table 4, upon the architecture for data extraction problems [26]. As depicted in Table 4, the AQQ\_CV\_Squad model is a lighter superset of BERT, slower to compute but with less precision, making it appropriate for more resource constrained environment.

Table 4. Configuration parameters of the AQQ\_CV\_Squad model built on DistilBERT architecture

Feature name	Value	Description
Hidden-act	Gelu	Hidden layer activation function
Hidden_size	768	Hidden layer size (embedding dimensions)
Max_position_embeddings	512	Maximum position embeddings limit
Model_type	Distilbert	Model type
N-heads	12	Number of attention heads
N-layers	6	Number of hidden layers
Vocab_size	30522	Vocabulary size

#### 2.5. Information extraction using TSHD algorithm and transformer models

The transformer-based models have revolutionized the field of NLP, and in particular in the domain of information extraction. Due to their capabilities in understanding context and semantics they had major improvements in question answering, summarization or document classification. These traits, and their property as good candidates to fuse with segmentation approaches like TSHD to improve the extraction accuracy.

In this work, we propose an innovative method which integrates segment-level information extraction based on TSHD algorithm. In contrast, as an alternative to processing the entire document, the proposed approach takes advantage of semantically coherent segments extracted by TSHD which are leveraged as the context input for extraction models. This approach improves the precision and saves computation resources by limiting the input to contextually relevant segments with clear topics.

First, the general mechanism of information extraction with transformer-based models is introduced. It then discusses an elaborate methodology for user TSHD algorithm integration for extracting information more effectively. As shown in Figure 1, this is the generic input/output structure of EQA models, including proposed segmentation-based improvement.

As shown in Figure 1, the EQA model consists of three main stages: tokenization and input encoding, prediction of start and end logits, and extraction of the final answer span. In our method the context is confined to semantically relevant parts, as addressed by the TSHD algorithm. Figure 1, also shows how to use the model in information extraction to the extract question (the information in context); the model follows the next steps to identify and return the answer from the provided text:

- Pre-processing: the input data is pre-processed, so it's suitable for the model. This means performing tokenization and transforming the textual input into the appropriate IDs in the lexicon of the model's vocabulary, then running the data through the model.
- Model prediction: the model predicts the start logit and end logit for each token (where a token is a language unit), specifying how likely (with respect to the textual context in which they are present) is that token to be the start/end of the answer.
- Post-processing: during this final stage, the model examines the logits it generated to find the tokens with the highest probability, thus determining where the answer starts and ends in the textual context. This skill provides an accurate answer.

Next, we present how the information extraction mechanism with TSHD algorithm is combined. Instead of sending the full document to be processed, only the extracted segments along with their corresponding topics are sent to the model for processing. This focused method maximizes accuracy and efficiency. Figure 2 illustrates how the model combines full resume segmentation by TSHD, section selection according to the alignment with the query topic, and finally QA model to extract the exact answers from the filtered context.

In the first step, the full document is fed into the TSHD model, which segments each passage and identifies its topic. After that, the question is matched with the relevant segment based on topic. The approach gives the model only the question, and the relevant part of the document so it does not have to deal with the entire document. Then, based on the retrieved snippet, the model extracts relevant information (i.e., the answer), saves the result in a structured format. By honing in on contextually relevant segments, this method improves the precision and efficiency of information retrieval, as opposed to parsing the full document. The experimental setup we use follows the steps:

- In this work, we collected 105 English resumes of different history from online portfolio, and manually labeled the fields (name, email, phone number, and education) in the resumes to obtain a labeled data set, and we adopted the SQuAD 1.1 way to arrange the questions and the answers.
- Pre-processing with TSHD algorithm: The resumes were pre-processed with the TSHD algorithm to extract key-value pair content stored in json format, where each key is the section header extracted and was assigned a value which is the text of the content found for that section.
- In model selection and configuration, three transformer based models i.e., autotrain-resume\_parser(Longformer), CV\_Custom\_DS(RoBERTa) and AQQ\_CV\_Squad(DistilBERT) were chosen, and all models were fine-tuned on the processed dataset using PyTorch.
- In Baseline without TSHD, original resumes were evaluated directly without being segmented, and models' prediction were checked against the ground truth using the metrics EM and F1-score.
- Evaluation with TSHD Integration For each question, the TSHD-segmented resume was analysed to determine which section is most relevant to resume content by topic matching and the selected portion of the resume served as the input instead of the whole document to the QA model, with resulting answers again evaluated using EM and F1-scores.
- Performance comparison involved the calculation and statistical analysis of the differences between baseline and Leukonychia TSHD-enhanced results.

The method as described above ensures the full reproducibility, and the developed models and datasets used are publicly available, which makes it possible to have reproducible, transparent and validation. In the next section, we examine the performance of our proposed method and the comparison between the performance of the model on TSHD with and without segmentation.

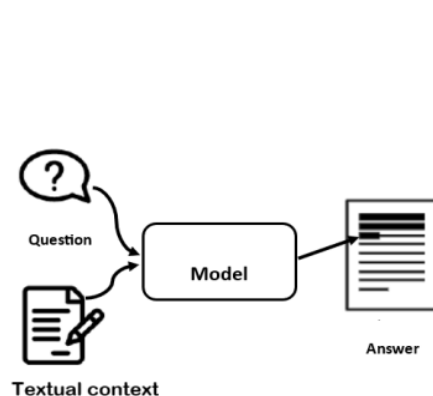


Figure 1. Input-output structure of EQA models with processing stages

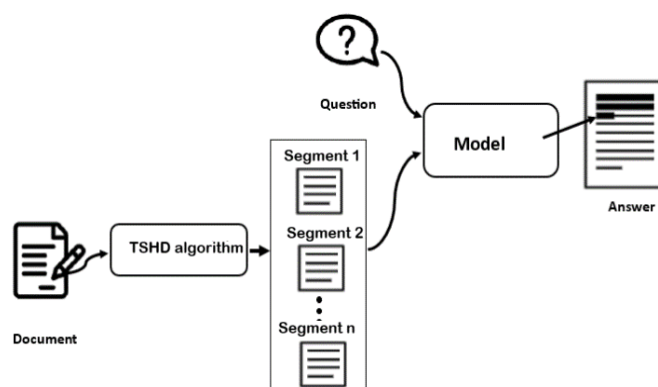


Figure 2. Workflow of integrating the TSHD algorithm with transformer-based extractive QA models

### 3. RESULTS AND DISCUSSION

In order to analyse the complementarity of our approach, we empirically evaluated the performances with 3 models based on transformer trained on resume corpus. Selected evaluation results of the proposed method are demonstrated, by focusing on the main metrics, including EM and F1-score signals. In the experimental setting we compare the models with and without TSHD segmentation to demonstrate the gain



brought by topic-dependent process. A demonstration of the proposed method evaluations is shown next, the proposed method was tested on top of a set of information extraction models as detailed on the research-tools section. We performed these models in the resume domain based on the SQuAD 1.1 framework to measure the effects of applying the TSHD to improve the accuracy over two phases of information extraction. Information extraction was evaluated on a dataset containing 444 questions, utilizing models such as CV\_Custom\_DS, autotrain-resume\_parser, and AQG\_CV\_Squad. Model evaluation was performed twice for each one:

- Without using TSHD algorithm: here, the model is fetching what it needs from the whole document.
- Using the TSHD algorithm: in this step the model provides the necessary information extracted from the relative segment based on the determined topic.

Evaluation results of all three models in terms of information extraction on the test dataset with and without TSHD algorithm are shown in Table 1, and also illustrated in Figures 3 and 4. As reported in Table 1, TSHD segmentation noticeably increases both EM and F1-scores for all models, which indicates the effectiveness of topic-aware context filtering.

Figure 3 shows a graphical summary of EM scores of all three models before and after the TSHD-based segmentation. The findings further demonstrate that focusing on the relevant sections of the documents improves the EM scores uniformly, in particular for RoBERTa and DistilBERT, for retrieval of answers.

Table 1. Performance comparison of three models with and without TSHD segmentation

Model	Without TSHD algorithm F1-score (%)	With TSHD algorithm EM (%)
Autotrain-resume_parser	91.55	80.63
CV_custom_DS	73.38	68.69
AQG_CV_squad	56.23	40.54

From Figure 4, when TSHD was introduced to the processing pipeline, all three models' F1-scores improved dramatically. Significantly, the RoBERTa-based model saw the greatest gain, supporting the effectiveness of the topic-aware segmentation in decreasing noise and increasing precision.

As shown in Figures 3 and 4 and Table 1, the used of the TSHD algorithm significantly improved the evaluation results of the three different models on different metrics. Specifically:

- a) For the autotrain-resume\_parser model:
  - The EM metric went from 80.63% (358 answers) up to 83.56% (371 answers), a 2.93% (13 answers) improvement.
  - The F1-score also increased from 91.55% to 93.18%, an improvement of 1.63%
- b) Model CV\_Custom\_DS
  - EM metric: 68.69% (305 answers) => 76.13% (338 answers) +7.44% (+33 answers).
  - The F1-score increased from 73.38% to 81.12% with a gain of 7.74%
- c) For AQG\_CV\_Squad model
  - EM metric increased from 40.54% (180 answers) to 45.95% (204 answers), a 5.41% (24 answers) improvement.
  - The F1-score metric was increased from 56.23% to 64.07%, a precise gain of 7.84%.

Overall, this significantly enhanced the information retrieval accuracy of transformer models when combined with the the TSHD algorithm. EM metric increased from 2.93% to 7.44% and F1-score metric increased from 1.63% to 7.84%.

Discrepancies between the performances of the models are related to their model architectures and learning methodologies [11]. This is for instance the model AQG\_CV\_Squad, based on DistilBERT, that had the lowest performance of the models tested [21], [27]. This result is consistent with prior work in which smaller models make an accuracy-computational cost tradeoff, especially under heavy tasks such as EQA [20]. This sacrifice, as Table 1 shows, is justified in resume parsing tasks by the lower F1 and EM scores. This is in accordance to our expectation, since DistilBERT is a small version of BERT tailored to be computationally efficient but not accurate to the best [22], [11]. Its smaller size and lower number of parameters make it suitable for running on resource constrained devices even when it is at the same time slightly lower performing on more complex NLP tasks like EQA. Previous works have verified that these compact models can be deployed with limited resources; however, their performance degrades when they process multi-segment or long-document tasks [22]. It is suited for running in resource-constrained environments due to its smaller size and fewer parameters, however with a trade-off being a lower performance in complex NLP tasks such as EQA [23]. This is an example of the-trade off between the task demand and the available resource, and that the choice of the model to be used could be determined according to the both task demand and the available resource. The advantage of doing this is that it is faster and requires less computation, though less accurate.

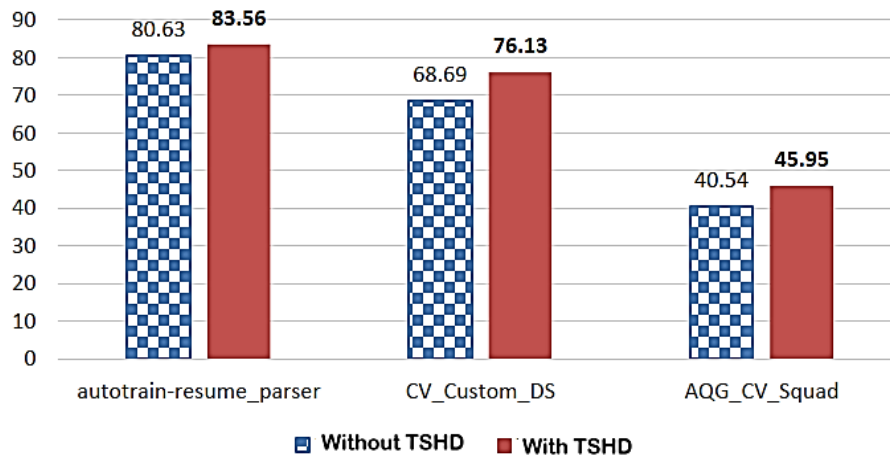


Figure 3. Comparative analysis of EM scores across three transformer-based models with and without TSHD segmentation

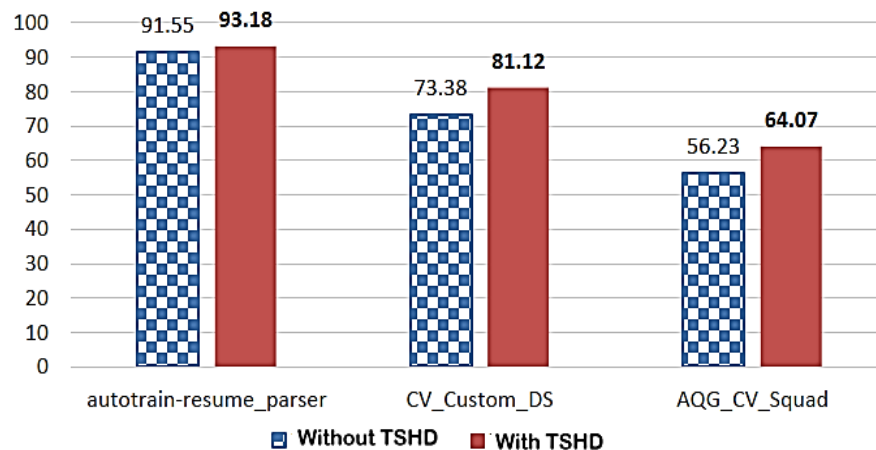


Figure 4. Comparative analysis of F1-scores across three transformer-based models with and without TSHD segmentation

#### 4. CONCLUSION

The proposed research proposes a new segmentation method based on TSHD to improve the information extraction mechanism of transformer models. The approach is to run the same algorithm to extract blocks of text, identify the topic for these blocks of text and then allow the models to extract information from these blocks instead of the whole document.

Three models were tested and evaluated on a collection of resumes with and without TSHD, replicating the unprecedented structure of the SQuAD 1.1 dataset. Experimental results show an obvious improvement after using the TSHD algorithm, achieving a maximal increase of 7.44% on the EM metric, and a maximal increase of 7.8% on the F1-score metric. Moreover, TSHD algorithm still contributes to the smaller amount of computational resources needed because of the decreased dimensionality of the textual context. Additionally, its visage does not demand retraining the models on the document segments after segmentation. The implications in terms of automated recruitment systems are clear: quick and accurate selection of candidates is of great importance. Our method enables scalable HR automation and enhanced decision-making pipelines by improving precision and reducing computational burdens.

Our findings suggest that combining topic-aware segmentation and deep learning based extractive QA (using TSHD) can help information extraction systems achieve the state of the art. Particularly, the better EM rate (+7.44%) and F1-score (+7.84%) indicate focusing the context into semantically coherent chunks can extract more with higher precision. The experimental results support that our TSHD-based segmentation remarkably improves the effectiveness of transformer models in resume extraction, proving the hypothesis.

This corroborates with the earlier work that documented how topicaware segmentation achieves a better precision in heterogeneous document analysis. It allows in addition to sit on recent methods such as SECTOR and Longformer-based segmentation that focus on structured context filtering. However, the study conducted is scoped with rather small dataset (105 resumes) for covering biases in template diversities. In the future, it would be interesting to extend the dataset and to explore multilingual settings.

An important implication of our results is that traditional whole-document processing in NLP tasks might not always be ideal, particularly when the input is heterogeneous or multi-topic documents as in resumes. Our method provides empirical support to the hypothesis that if the model pays attention only to semantically coherent segments, which we do not have to re-train the model on the segmented data, this methodology can be easily generalized to various domains and types of documents. This opens up the possibilities for applying similar techniques to be used in legal documents, research papers and technical reports. This suggests that adaptive segmentation approaches, which have been developed for specific document types, may further improve the performance of AR systems and other HRapplications.

Further work might include enlarging the architecture to work with multilingual resumes — and integrating un-supervised segmentation tools that empirically shows to perform well, that provide good complementary approaches when the amount of labeled heading data is limited, scaling up the system to larger test sets and comparing to (different) segmentation algorithms such as SECTOR or Text Tiling, and evaluate performance on SQuAD 2.0, which also contains questions that cannot be answered to see how it robustness works in the wild.

In general, this work is a step toward the development of more intelligent and context-sensitive information extraction systems capable of handling rich document structures while being efficient, and maintaining precision. In light of the increasing need for efficient and faithful document processing, this work provides insightful guidelines on how to improve NLP systems by means of segmentation. Due to its interdisciplinary nature (it integrates NLP, deep learning and document structure analysis) its context is highly switable for the journal's audience and its content is a valuable addition to the state of the art in intelligent information extraction.

## FUNDING INFORMATION

This research work is not funded by any specific project.

## AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Anmar Abuhamdah	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓	
Mohammed Al-Shabi	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓			
Sana Jawarneh	✓	✓	✓	✓	✓	✓	✓			✓	✓		✓	✓

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

## CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest to declare.

## DATA AVAILABILITY

Data availability does not apply to this paper as no new data were created or analyzed in this study.




## REFERENCES

- [1] A. Vaswani *et al.*, "Attention is all you need," in *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 5998–6008.
- [2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Minneapolis, MN, USA, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.




- [3] S. Arnold, R. Schneider, P. Cudré-Mauroux, F. A. Gers, and A. Löser, "SECTOR: a neural model for coherent topic segmentation and classification," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 169–184, 2019.
- [4] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 26th Conference on Neural Information Processing Systems (NeurIPS)*, vol. 26, 2013, pp. 3111–3119.
- [5] O. Koshorek, A. Cohen, N. Mor, D. Rotman, and J. Berant, "Text segmentation as a supervised learning task," in *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2018, pp. 469–473.
- [6] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.
- [7] C. D. Manning *et al.*, "The stanford CoreNLP natural language processing toolkit," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations (ACL)*, 2014, pp. 55–60.
- [8] G. Lample and A. Conneau, "Cross-lingual language model pretraining," in *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*, 2019, arXiv:1901.07291.
- [9] M. E. Peters *et al.*, "Deep contextualized word representations," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, New Orleans, LA, USA, 2018, pp. 2227–2237, doi: 10.18653/v1/N18-1202.
- [10] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 3730–3740, doi: 10.18653/v1/D19-1387.
- [11] Y. Liu *et al.*, "RoBERTa: a robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
- [12] M. E. Tannous, W. H. Ramadan, and M. A. Rajab, "TSHD: topic segmentation based on headings detection (Case Study: Resumes)," *Advances in Human-Computer Interaction*, vol. 2023, Art. no. 6044007.
- [13] Z. Yang *et al.*, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019, pp. 1–10.
- [14] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ Questions for machine comprehension of text," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Austin, TX, USA, 2016, pp. 2383–2392, doi: 10.18653/v1/D16-1264.
- [15] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: unanswerable questions for SQuAD," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, vol. 2, Short Papers, Melbourne, Australia, 2018, pp. 784–789, doi: 10.18653/v1/P18-2124.
- [16] A. Trischler *et al.*, "NewsQA: A machine comprehension dataset," in *Proceedings of the 2nd Workshop on Representation Learning for Natural Language Processing (ReplANLP)*, Vancouver, Canada, 2017, pp. 191–200, doi: 10.18653/v1/W17-2623.
- [17] E. Choi *et al.*, "QuAC: question answering in context," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Brussels, Belgium, 2018, pp. 2174–2184, doi: 10.18653/v1/D18-1241.
- [18] T. Möller, A. Reina, R. Jayakumar, and M. Pietsch, "COVID-QA: a question answering dataset for COVID-19," in *Proceedings of the 1st Workshop on NLP for COVID-19 at the Association for Computational Linguistics (ACL)*, 2020.
- [19] K. Pearce, T. Zhan, A. Komanduri, and J. Zhan, "A comparative study of transformer-based language models on extractive question answering," *arXiv preprint arXiv:2110.03142*, 2021.
- [20] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.
- [21] R. Ghosh *et al.*, "Topic segmentation of semi-structured and unstructured conversational datasets using language models," *arXiv preprint arXiv:2310.17120*, 2023.
- [22] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, vol. 2, Short Papers, Valencia, Spain, 2017, pp. 427–431.
- [23] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: the long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.
- [24] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proceedings of the Workshop on Text Summarization Branches Out*, Association for Computational Linguistics, 2004, pp. 74–81.
- [25] Sunitha, "CV\_Custom\_DS," *Hugging Face*. [Online]. Available: [https://huggingface.co/sunitha/CV\\_Custom\\_DS](https://huggingface.co/sunitha/CV_Custom_DS). Accessed: Feb. 27, 2025.
- [26] Sunitha, "AQG\_CV\_Squad," *Hugging Face*. [Online]. Available: [https://huggingface.co/sunitha/AQG\\_CV\\_Squad](https://huggingface.co/sunitha/AQG_CV_Squad). Accessed: Feb. 27, 2025.
- [27] S. Kumar, "Answer-level calibration for free-form multiple choice question answering," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, vol. 1, Long Papers, Dublin, Ireland, 2022, pp. 665–679, doi: 10.18653/v1/2022.acl-long.49.

## BIOGRAPHIES OF AUTHORS






**Anmar Abuhamdah**    received the B.Sc. degree in Computer Science from Princess Sumaya University for Technology, Jordan, in 2003, and his M.Sc. degree in Intelligent Systems from Utara University, Malaysia, in 2007, and the Ph.D. degree in Computer Science from the National University of Malaysia, Malaysia, in 2011. He is currently an associate professor at Taibah University, Kingdom of Saudi Arabia, since 2018. His research interests are mainly directed to metaheuristics and combinatorial optimization problems including timetabling, routing, quadratic, and rostering. His research interests include the applications of artificial intelligence, including deep learning, evolutionary and heuristic optimization techniques, operation and control, text classification, feature selection prediction models. He can be contacted at email: [aabuhamdah@taibahu.edu.sa](mailto:aabuhamdah@taibahu.edu.sa).



**Mohammed Al-Shabi**    is an associate professor in the Department of MIS, Faculty of Business Administration, at Taibah University, KSA. He received his bachelor's degree from the Computer Science Department at University at Iraq (1997). He received his master's degree in Computer Science from Putra Malaysia University at 2002, and Ph.D. (Computer Science) from Putra Malaysia University, Malaysia (2006). His research interests are mainly directed to Network security, data protection, and cloud security. In addition to data protection and privacy, developing new protocols, and using artificial intelligence in cybersecurity, wireless security, cryptography, UML, stenography multistage interconnection network, parallel computing, and apply mathematic. He can be contacted at email: mshaby@taibahu.edu.sa.



**Sana Jawarneh**    obtained her B.Sc. in computer engineering from Yarmouk University and her Ph.D. in computer science at University Kebangsaan Malaysia. Now, she is an assistant professor at Department of computer science, The applied college, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia. Her research interest falls under meta-heuristic algorithms in various optimization problems. She can be contacted at email: sijawarneh@iau.edu.sa.