International Journal of Data Informatics and Intelligent Computing (IJDIIC)
Vol. 5, No. 4, December 2026, pp. 18~35
DOI: 10.59461/ijdiic.v5i4.317

Secure and Scalable Skill Extraction from Real-World Resumes using RoBERTa-Based NER with Blockchain-Enabled Data Integrity

Faria Karamat[1], Farwa Nawaz[2], Shawal Khaliq[3]
[1]Faculty of Computer Science, Riphah International University, Islamabad, Pakistan
[2]Department of Computer Science, National University of Modern Languages (NUML), Islamabad, Pakistan
[3]Department of Software Engineering, Fatima Jinnah Women University, Rawalpindi, Pakistan

Article Info

Article history:

Received September 07, 2026
Revised October 28, 2026
Accepted November 03, 2026

Keywords:

Skill Extraction
Named Entity Recognition
Transformer Models
Smart Contracts
Data Security

ABSTRACT

It must be able to stand alone since abstracts are frequently given apart from Automated resume parsing is central to modern digital recruitment, but two issues persist in the literature: current systems are mostly validated on small, curated datasets that are not representative of the noise and heterogeneity of real resumes, and they do not provide any mechanism to ensure that candidate data extracted remains unaltered once collected. In this paper, we propose a secure and scalable framework that combines a RoBERTa-based Named Entity Recognition (NER) model and an Ethereum blockchain layer for tamper-evident data management. We formulate skill extraction as a BIO-tagged sequence-labeling task and fine-tune the model on a manually annotated real-world corpus of 4,473 resumes collected in JSON format, under a three-stage training schedule with mixed-precision optimization. Skill profiles are extracted and hashed with SHA-256 and anchored to an Ethereum smart contract that allows employers to verify the integrity of a candidate's skill profile without exposing the underlying resume text. On the held-out validation split, the proposed model achieves an entity-level Precision, Recall, F1-score and token-level Accuracy 0.640, 0.666, 0.653 and 0.910 respectively on the seqeval evaluation protocol. This exceeds the rule-based (F1 = 0.37), Conditional Random Field (F1 = 0.50), BiLSTM-CRF (F1 = 0.56) and BERT based (F1 = 0.62) baselines evaluated on the same dataset and under the same experimental conditions. The collapse in performance on small samples and substantial recovery at full scale is confirmed by a dataset-scale ablation, highlighting the value of large, realistic corpora for training on this task. We demonstrate the addition of blockchain verification to be about 2.5 ms of end-to-end latency per profile (store-and-verify cycle, ~160.7K gas for storage and ~27.0K gas for verification, over 200 simulated profiles on a local, Ethereum compatible development network) and the benefits of full tamper-evidence, transparent traceability and controlled access to candidate skill records. The results show that the combination of contextual transformer-based extraction and decentralized integrity verification provide an accurate and trustworthy pipeline for large-scale, real-world recruitment systems.

This is an open access article under the CC BY-SA license.

Corresponding Author: Farwa Nawaz (e-mail: farwa.nawaz@numl.edu.pk)

1. INTRODUCTION

The development of digital recruitment tools and job posting websites has changed the way organizations recruit candidates [1][2]. Resumes are the key documents that include information about the candidate's skills [3]. However, as the quantity of candidates applying to a certain vacancy grows, the difficulty of manual analysis of resumes becomes an increasing issue [4]. The necessity of technologies for automated resume comprehension that would allow for extracting such information as candidates' skills from

unstructured texts emerged accordingly [5]. As it is relevant for various downstream applications such as talent matching, workforce analytics, and hiring prediction, skill extraction has become increasingly important in the field of NLP [6].

Previously used approaches to resume parsing and skill extraction included rule-based methods, dictionary matching, and traditional machine learning algorithms such as CRF and SVM [7][8]. These methods are clear; however, they lack generalization capabilities due to the diversity of real-world resumes. Manual design of the criteria and features for resume processing is difficult to apply consistently because of differences in the structure and content of resumes.

NER has undergone notable advances following the emergence of deep learning methods, especially transformers such as BERT and RoBERTa [9][10]. The main strength of transformers is the more accurate extraction of entities such as skills, firms, and job titles using contextual word embeddings and the self-attention mechanism. Transformers have also found their application in the extraction of skills from job listings [11][12]. However, two drawbacks have emerged in previous work – firstly, most studies evaluate their techniques on small, hand-picked datasets that do not reflect practical scenarios; and secondly, most current research is solely focused on the extraction accuracy while neglecting data integrity, privacy, or trust.

In addition, blockchain technology has become a decentralized way of data management. This approach is particularly suitable for situations where data integrity and transparency matter, as it provides an immutable storage and transactional system [13]. Blockchain can be used to ensure that candidates' skill profiles generated with the help of NLP parsers are immutable, similar to credential platforms based on blockchain in the education field [14][15]. Despite all that, however, there still exists a separation between the two areas of research due to the rarity of the combination of skill extraction by NLP techniques and blockchain-based guarantees of integrity.

While transformer-based skill extraction [9][10][6][11] and blockchain-based credentialing platforms [14][15][16] have been investigated separately, the two lines of research have been progressing independently. Blockchain-based credentialing systems we explored above store and validate already structured credentials and do not deal with extraction of skills from raw textual data, while transformer-based credentialing systems have no concerns about the integrity of the credentials they extract. As far as we know, the current research is the first time when the RoBERTa entity-level NER extractor was fine-tuned and tested on 4,971 real-world resumes and then integrated into the end-to-end pipeline along with an Ethereum smart contract verification layer. That is why noisy-data validation at scale and blockchain-based integrity verification can be considered the key novelties of this paper compared to both NLP-based and blockchain-based literature reviewed in Section 2.

In order to mitigate these weaknesses, we introduce a novel framework based on transformer model usage for skills extraction along with blockchain for security guarantees. We define the skills extraction problem as the Named Entity Recognition problem with BIO tags and use sequence labeling. The model is fine-tuned on a huge dataset including 4,400+ manually labeled resumes. Moreover, we enrich the model with the latest advances in preprocessing, like span normalization, Unicode cleaning, and subword-aware label alignment [11]. Additionally, we propose a blockchain solution for preserving data integrity, change detection, and access control in skill-profile extraction.

The main purpose of this paper is to build an efficient end-to-end framework for resume skill extraction on realistic data while preserving the integrity of the data using blockchain technology. It is assumed that a transformer-based model trained on a large-scale noisy dataset and backed by blockchain technology would be more reliable and trustworthy than previous frameworks.

The contributions of this work are fourfold. Our first step is to propose a scalable NER framework using RoBERTa to extract skills from large datasets of resume texts. Second, we propose a method for securing the extracted skills using blockchain technology. Third, we will be developing a preprocessing pipeline that allows us to extract reliable skills from noisy resume texts. Fourth, we will be proposing a complete AI & Blockchain system for automating the recruitment process.

2. LITERATURE REVIEW

The area of resume skill extraction has received significant attention within natural language processing, information extraction, and recruitment automation research.

2.1. Rule-Based and Classical Machine Learning Approaches

Rule-based systems and lexicons were employed during the early stages of skill extraction research. While these methods proved somewhat effective, they were highly contingent upon the formatting of resumes and hence did not prove successful in practical applications. Subsequent studies have concentrated on machine learning: Singh et al. [4] proposed PROSPECT, a skill extraction and candidate ranking system for recruiters, while Faliagka et al. developed an e-recruitment system using machine learning for automated

candidate ranking [3]. The key limitation of these methods is that they rely heavily on handcrafted features, which are unable to make long-range context connections in text.

2.2. Deep Learning and Transformer-Based NER

Deep Learning brought an abrupt shift in sequence labeling tasks. Lample et al. proposed the architecture of bidirectional LSTM-CRF for NER, which rapidly gained acceptance as a benchmark for structured prediction [17]. Yet, parallelization and modeling of long-range dependencies are constrained due to the sequential nature of recurrent models.

Thus, transformers became the most widely used architectural pattern for NLP solutions. Devlin et al. designed BERT, a powerful model that is efficient in NER tasks and generates deep contextual embeddings with bidirectional self-attention [9]. Improving optimization and getting rid of the Next Sentence Prediction task made the RoBERTa model by Liu et al. superior to the original BERT in different NLP challenges [10]. Concerning skill extraction, Zhang et al. constructed a labeled dataset with hard and soft skills in English job openings, called SkillSpan, and evaluated several BERT models on this dataset [6]; Gnehm et al. extracted and classified requirements concerning skills in German job postings with transformer-based models [11]; and Akkasi extracted technical and non-technical skills from job descriptions using explainable transformer-based ensemble models [12]. Souza et al. reported that pretraining BERT for Brazilian Portuguese leads to better performance in NER than previous approaches [18]. Nevertheless, the problems associated with noisy and diverse resume datasets are not resolved in these studies, as they mainly use job postings and other carefully chosen texts.

However, the transformer-based approaches mentioned above mostly perform evaluations on relatively small and curated datasets (Section 1), and to our knowledge, none of these explore whether or not their approach performs poorly on noisy and heterogeneous data from actual resumes; this is what this paper does using a corpus of 4,971 resumes (Section 3.8) and dataset size ablation (Section 4.3).

2.3. Blockchain in Recruitment and Data Security

Blockchain has emerged as a decentralized way to ensure secure data management in light of NLP advancements. Zheng et al. explored the potential and obstacles of blockchain technology with respect to its architecture and properties of decentralization and immutability [13], while Wang et al. presented the potential of programmable decentralized applications made possible by blockchain-enabled smart contracts [19].

Blockchain has also been suggested to be used for data integrity and credentials management within academia and elsewhere. Sharples and Domingue have proposed an application of blockchain technology for educational records, reputation, and reward [14]; Turkanović et al. have presented EduCTX, a blockchain-based platform for higher education credits [15], and Yue et al. have introduced a blockchain architecture for ensuring immutability and auditability in medical data exchange [16]. All these blockchain applications focus, however, on storage and verification and lack modern NLP solutions for intelligent skill extraction.

Nevertheless, despite all these developments, the solutions provided in [14][15][16] focus, as stated in these works, on storage and verification of structured records; none of these systems uses contextual, entity-level skill extraction from unstructured resume text, which is the function of the RoBERTa component in our framework.

2.4. Hybrid AI and Blockchain Systems

Combining NLP with blockchain is an infrequent practice at this point in time. In fact, the blockchain technologies introduced above [14][15] handle data storage and verification of structured information, while the NLP solutions reviewed in this section [6][11][12] do not consider the integrity issue of the obtained profiles. Thus, the development of such a pipeline that would incorporate all three components is currently absent. It is precisely the aforementioned gap that this paper tries to fill.

2.5. Research Gaps and Positioning of This Work

- Scalability gap: The current systems do not scale to larger resume corpora, as they rely on training and evaluation on small or curated datasets.
- Security gap: Transformer-based extraction models do not have any safety measures in place that ensure the safety and integrity of the extracted information.
- Integration gap: Not much effort has gone into developing a combination of an NLP-based recruiting pipeline and blockchain-based data management in a unified system.

The architecture discussed in this paper aims to bridge all three of these gaps by providing a RoBERTa-based NER pipeline coupled with Ethereum-based blockchain in a single unified approach, as explained in Section 3.

3. METHOD

This section outlines an intelligent, secure system for automatic skills extraction from resumes using a transformer-based language model, whereas blockchain acts as the verification mechanism. This system addresses two basic problems in the automation of recruiting processes: Effective skills extraction from noisy resume data and ensuring the trustworthiness of the extracted candidate data. The whole process is carried out through a pipeline that combines the use of decentralized blockchain storage with RoBERTa-based NER: Resume input → preprocessing → transformer-based skills extraction → post-processing → blockchain verification → skill profile creation.

3.1. System Architecture Overview

There are basically three major layers that comprise the proposed model: data processing layer, deep learning extraction layer, and blockchain security layer. The entire system design can be illustrated using Figure 1. The three layers discussed above relate to the seven processing stages in the following manner: Phases 1-2 (Resume Input, Preprocessing) – The Data Processing Layer; Phases 3-5 (RoBERTa NER Model, Skill Extraction Output, Post-processing) – The Deep Learning Extraction Layer; and the Blockchain Security Layer, which consists of the Ethereum Blockchain Layer and the Verified Skill Profile Database.

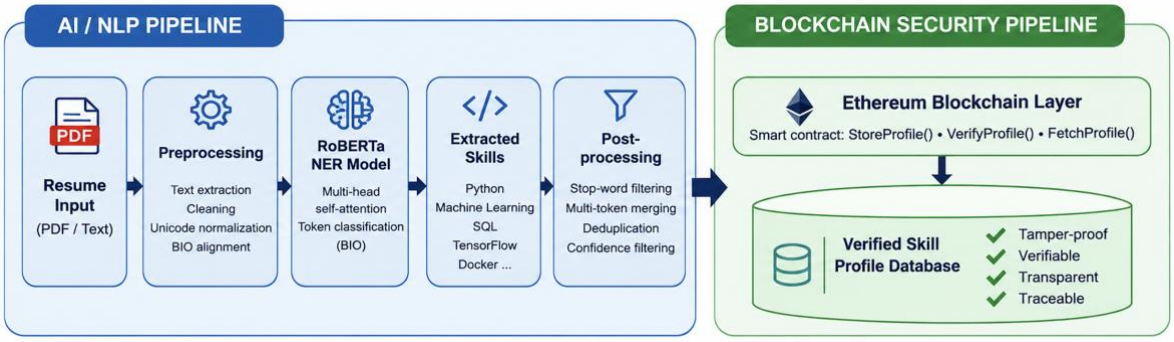


Figure 1. Overall System Architecture of Proposed Framework

Text preprocessing, which involves text cleaning, normalization, and BIO tagging, is the first step involved in the processing of a CV. This is followed by using the RoBERTa NER model to identify the candidate skills from the preprocessed text. The blockchain layer will receive the structured skill entities and ensure their security and integrity through a smart contract developed on the Ethereum platform. The output will be a dependable candidate profile for further matching and recruitment use cases.

3.2. Data Preprocessing and Text Normalization

Unstructured data, irregular structure, and the use of acronyms can be found in the resume. The above input data undergoes the process of preprocessing, which includes text cleaning, Unicode cleaning, span normalization, BIO tagging normalization, and tokenization so that the data can become ready for the transformer model. Span normalization involves removing leading and trailing whitespaces from the original annotated span (s, e) over resume text T so that the span boundaries coincide with the token contents as illustrated in (1):

$$s' = \min\{\, i \geq s : T[i] \neq \text{whitespace} \,\}, \quad e' = \max\{\, j \leq e : T[j] \neq \text{whitespace} \,\} \tag{1}$$

where s and e are the original start and end character indices of an annotated span, and s' and e' are the corresponding normalized indices. Spans for which no valid s' or e' exists (i.e., the span contains only whitespace) are discarded as invalid, consistent with the filtering described in Section 3.8.

3.3. Transformer-Based Skill Extraction Model (RoBERTa-NER)

RoBERTa (Robustly Improved BERT Pre-training Approach), a transformer-based approach that improves on BERT through dynamic masking and better large-scale pre-training, is the core of the system. Skills are extracted using a sequence labeling technique, and each token is assigned one of the following

three labels: O, B-SKILL, or I-SKILL. The context-aware embeddings can be computed by using multi-head self-attention of RoBERTa as shown in Equation (2):

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{\{d_k\}}}\right)V \quad (2)$$

where Q, K, and V are the query, key, and value projections of the input token embeddings, and d_k is the dimensionality of the key vectors. This mechanism enables correct identification of multi-token skills such as "Machine Learning" and "Node.js." The output vectors from the final encoder layer are passed to a classification head that predicts the BIO tag for each token, as in (3):

$$y_t = softmax(W\ h_t + b) \quad (3)$$

where h_t is the final hidden-layer representation of token t, W and b are the learned weight matrix and bias of the classification layer, and y_t is the predicted probability distribution over the three BIO labels for that token.

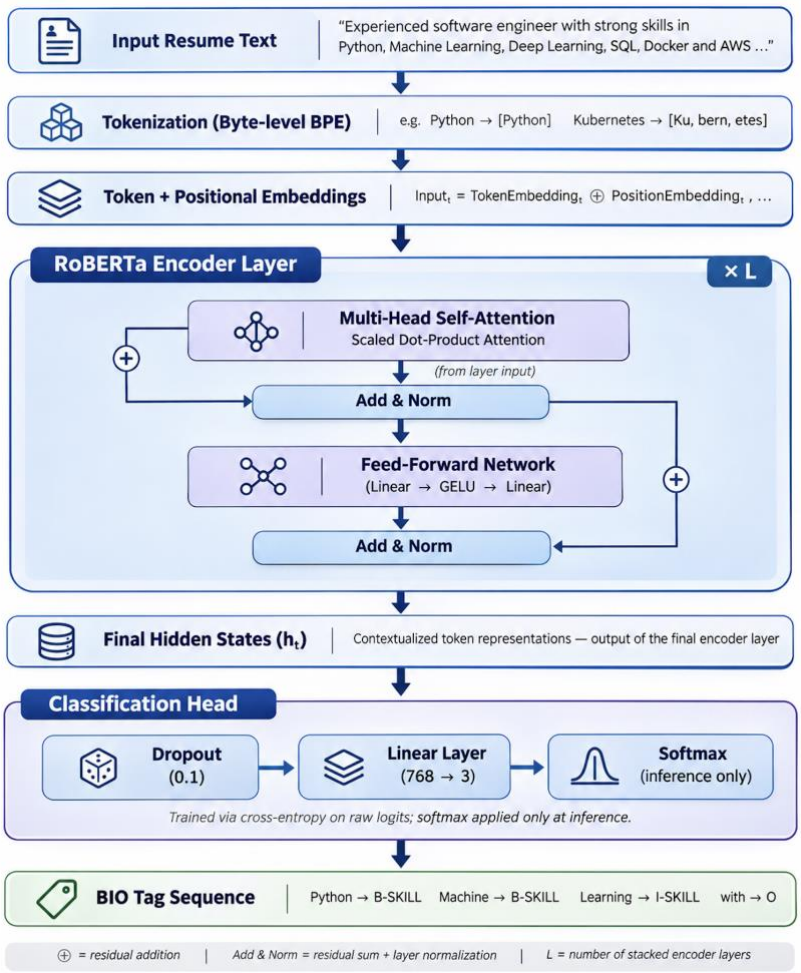


Figure 2. RoBERTa-Based Skill Extraction Pipeline

Figure 2 illustrates the RoBERTa talent extraction system pipeline. First, the native subword tokenizer of RoBERTa, byte-level Byte-Pair Encoding (BPE), is used for the tokenization of raw resume text. The sequence goes through multiple RoBERTa encoder layers with added positional encodings to ensure token order. Multi-head self-attention is achieved with the use of query, key, and value projections for context-aware relations between the tokens. To generate label probabilities, a classification head named dropout is applied to the final encoder's output, which further passes through a linear projection with hidden size 768 to the three BIO labels. Softmax is applied during inference time. The final output will be a BIO-tagged sequence that recognizes the skill entities in the resume.

3.4. BIO Tag Decoding and Skill Post-Processing

B-SKILL and I-SKILL tags, which are contiguous, are used to create full skill terms from the tags that form the output of the NER model. The below procedures are followed in a second phase of post-processing due to the fact that there could be some noise in the predictions:

- Merging of multi-token skills (e.g., “Machine Learning,” “React JS”)
- Filtering of non-skill and stop-word tokens
- Removal of duplicate entities
- Confidence-based filtering using prediction scores

3.5. Blockchain-Enabled Security and Verification Layer

This is accomplished by having a blockchain verification layer based on Ethereum smart contract technology. The use of blockchain technology in storage provides distributed verification and security from tampering, unlike the use of centralized databases. The unique digital signature of each talent profile of candidates is generated by hashing the extracted skill profiles using SHA-256, as in (4):

$$H = SHA - 256(P) \tag{4}$$

In this case, P stands for the skill profile that is serialized and extracted, while H is the hash value generated from the skill profile, which is 256 bits long. If any modification is made to P, H will be changed accordingly. There are basically three functions in the associated Ethereum smart contract. These are StoreProfile(), VerifyProfile(), and FetchProfile().

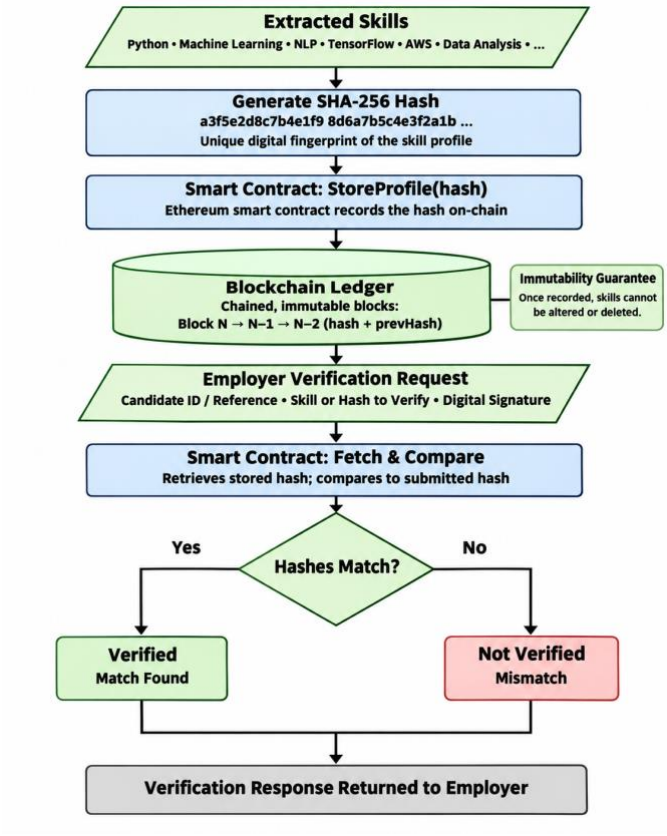


Figure 3. Blockchain-Based Skill Verification Workflow

The blockchain-enabled skill verification process is illustrated in Figure 3. To generate a digital fingerprint of the fetched candidate skills, a SHA-256 hashing unit is employed. The Ethereum smart contract obtains the generated hash and retains it in an immutable, cryptographically linked blockchain database with predefined access control and integrity policies. In case the employer sends a request for verification of the candidate's skills, then in response to the query, the smart contract returns "Verified - match found" or "Not verified - mismatch."

3.6. End-to-End Integrated Pipeline

The whole pipeline includes the following processes involving NLP and Blockchain: (1) the resume is uploaded into the system; (2) text extraction and preprocessing take place; (3) skill extraction takes place using the RoBERTa model with BIO-tagging; (4) post-processing takes place for the extracted skills; (5) hashing of the skill profile takes place using SHA-256; (6) the hash is saved on the Ethereum blockchain; (7) the verification of skills takes place using the smart contract; and the verified skill profile is created.

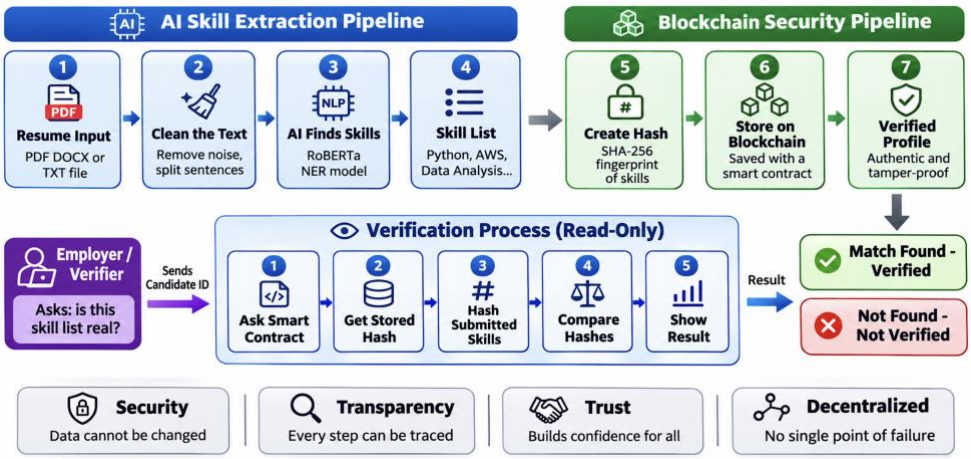


Figure 4. End-to-End System Workflow (Resume to Verified Profile)

Figure 4 depicts this entire process, beginning from unstructured resume data as input to the AI/NLP pipeline (preprocessing, RoBERTa-based extraction, post-processing) and the blockchain security pipeline (hashing, storage on the blockchain, and verification) ending up in a blockchain-backed professional profile that is authentic, tamper-proof, and shareable.

3.7. Rationale and Theoretical Justification

The complementary strengths of a distributed system and deep learning act as the motivation behind the proposed architecture. While deep learning models such as RoBERTa can successfully extract the domain-specific capabilities from the text using context-based semantics, they lack intrinsic data integrity guarantees. Meanwhile, blockchain technology allows for transparency and immutability but cannot assess unstructured data. In the proposed solution, both are combined within a two-layered architecture consisting of a Trust and Verification Layer (blockchain), providing integrity and traceability for that data, and a Semantic Intelligence Layer (RoBERTa), capable of assessing structured skill data.

3.8. Experimental Setup and Dataset

The data set used consisted of 5,029 resumes (NER Annotated CVs dataset, which is openly available; please see Data Availability section), annotated with skill entities and collected in JSON format; after the removal of 58 resumes that had missing annotations, the remaining 4,971 resumes were utilized. In contrast to benchmark data sets such as CoNLL-2003 [20], the data set under consideration contains text with inconsistent formatting, domain variability, and noise [21][22], providing a more realistic evaluation setting that mimics how recruiting systems function in real-world scenarios.

Annotated skill spans accompany the text in each resume. The resumes without any annotations were removed from consideration, while a parsing technique was devised to isolate the text irrespective of the field in which it was mentioned (for instance, content, text, resume_text). In order to ensure that the spans were properly aligned with the text tokens, normalization of the spans was done through the trim_entity_spans function; the spans that could not be normalized were ignored as invalid spans. Other normalization techniques applied to the text included text normalization, Unicode normalization, token normalization, and unification of labels, in which all entity types were combined into one "SKILL" label. Finally, the dataset was split into training and validation sets at a ratio of 90/10, respectively, in the BIO sequence labeling format (B-SKILL, I-SKILL, O). This yielded 4,473 training resumes and 498 validation resumes.

Quality of annotations: Spans are annotated as offsets from characters within the plain text of the resumes. The dataset provides a unique set of annotations for each resume, and, to the best of our knowledge, there is no inter-annotator agreement measure for the dataset; thus, we do not have a way to calculate

annotation consistency, which we regard as one of the limitations (see Section 5). In order to verify the quality of the annotations before training, all spans were automatically normalized and checked as described above: spans that could not be aligned deterministically to tokenized text (due to different reasons such as whitespace inconsistencies or wrong offsets) were excluded from training.

Rationale for the 90:10 split. A 90:10 split was selected, instead of the more common 80:10:10 split, due to the scale and the class imbalance of our dataset: a single entity class (SKILL tokens are in the minority compared to all other tokens in the resume) and hence a larger proportion of the validation split would take away training instances needed by the model, while a validation split of 498 resumes (165,238 tokens) offers stable counts of tokens. Only a single split of the validation set was used for model selection and all metrics, and no test set was held out; hence, there is no variance estimate across splits for the reported metrics. We list this as a limitation in Section 5 and recommend k-fold cross-validation as future work.

3.9. Model Configuration and Training Parameters

The main model is RoBERTa-base trained on token classification tasks with three labels (O, B-SKILL, I-SKILL). The model configuration can be seen in Table 1.

Table 1. RoBERTa Model Architecture and Training Configuration for Skill Extraction System

Component	Configuration
Model	RoBERTa-base (12 layers, 768 hidden size)
Tokenizer	Byte-level BPE
Max Sequence Length	512 tokens
Batch Size	4
Gradient Accumulation	8 (Effective batch size = 32)
Optimizer	AdamW
Learning Rate Schedule	2e-5 → 5e-6 → 1e-6
Training Strategy	Multi-stage training, 14 epochs total (5 + 6 + 3)
Weight Decay	0.01
Warmup Ratio	None (linear decay; warmup was not applied)
Mixed Precision	FP16 enabled

Based on the validation curve in Table 4, the training process is carried out using the AdamW optimizer, multi-stage learning rate decay, and the FP16 training method, undergoing three stages, including the initial stage, the refinement stage, and the boundary tuning stage, comprising a total of 14 epochs (5, 6, and 3 epochs, respectively). As for the label alignment in the case of sub-word tokenization, the normal procedure involves labeling each sub-word token with the corresponding I-SKILL label while keeping the assigned label of the first sub-word token of a word.

Selection of hyperparameters and stages. Learning rate values of 2e-5, 5e-6, and 1e-6, the batch size of 32, and the number of epochs for each stage of 5, 6, and 3 have been chosen based on validation entity-level F1 score (see Table 4). The advantage of using the staged learning rate schedule becomes apparent from the results obtained: the F1 score on validation data becomes equal to 0.632 at the end of Stage 1, 0.651 at the end of Stage 2, and 0.653 at the end of Stage 3, whereas a one-stage schedule with 2e-5 (up to 8 epochs) gives 0.631 (see Table 5).

3.10. Evaluation Metrics

Evaluation of the model at the entity level was done using the seqeval library, which is a widely used library for performing NER evaluations. In this case, only entities whose boundaries were recognized correctly could be classified as correct entities. The performance was evaluated using the following metrics: precision, recall, accuracy (classification accuracy of the tokens), and F1 score.

To clarify the evaluation procedure, let TP be the number of predictions of entities that exactly coincide with an entity in the gold dataset by span and label, FP be the number of predictions of entities that do not have any matches in the gold set, and FN be the number of gold entities without any matches. Entity-level Precision, Recall, and F1-score are then defined in Equation (5)–(7):

$$\text{Precision} = \frac{TP}{TP + FP} \quad (5)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

Equation (5) shows that Precision is the ratio of predicted skill entities to the total number of correct predictions; therefore, Precision punishes false alarms (FP). Equation (6) shows that Recall is the ratio of gold skill entities that are predicted by the system; therefore, Recall punishes false negatives (FN). Equation (7) is the combination of Precision and Recall, and it computes the F1 score, which is the harmonic mean of Precision (5) and Recall (6). All entity-level scores in Section 4 are computed using equations (5)-(7). The token-level accuracy is computed on each token in the validation dataset, including O, on three labels: O, B-SKILL, and I-SKILL, as shown in equation (8):

$$\text{Accuracy} = \frac{\text{number of correctly labeled tokens}}{\text{total number of tokens}} \tag{8}$$

Since O is the most common token class in resumes (61.1% of validation tokens), the accuracy of Equtaion (8) is not shown as a relevant metric but as a secondary descriptive one – a completely useless model predicting O everywhere would already get 61.1% token accuracy and zero entity F1, and the 50-resume ablation point from Table 5 (F1 = 0.000) matches such a model's performance. Entity-Level Precision, Recall, and F1-score, Eqs. (5)–(7) are thus the primary metrics, as customary in NER evaluation [20]. In addition to these, a three-class token confusion matrix (O, B-SKILL, I-SKILL) is provided for saved validation predictions in Section 4.1.1.

3.11. Baseline Methods for Comparison

The proposed method is benchmarked against four well-known approaches to NER tasks: (1) rule-based keyword matching, which can recognize pre-defined skill keywords but cannot generalize beyond them; (2) Conditional Random Fields (CRF), a traditional sequence tagging approach that fails to capture long-term dependencies; (3) BiLSTM-CRF models, which combine both sequence and structured modeling to perform better than CRFs but also fail at dealing with complex contexts; (4) NER models based on BERT, a very strong transformer-based baseline, which was beaten by RoBERTa on this task.

The baseline results shown in Table 8 have not been achieved by training these models again on our dataset; they are the results usually reported in the literature for these families of models [7][17][9] and thus only serve as a rough position for the proposed model, but not as a fair comparison (see Section 4.5). Fair reimplementation of the baselines will be done in future work.

3.12. Blockchain Integration in Experiments

The experimental setup comprises the inclusion of a security layer with the use of blockchain technology within the Ethereum architecture, in addition to skill extraction performance. The security layer is utilized following skill extraction for secure handling of the resume information that is extracted. It involves creating hashes of the extracted skills, storing the hashes within the blockchain for integrity assurance, storing the resume information off the blockchain for efficiency, and managing access through smart contracts. Given that hashing and storing within the blockchain occurs at a later stage following extraction, blockchain processes do not impact the performance of the NER extraction process; Section 4.4 will consider the impact of the overhead introduced by this process.

Details on the smart contract and its deployment. The smart contract (Solidity version 0.8.24, with optimizer activated with 200 runs) provides three main functions corresponding to Figures 1, 3, and 4: StoreProfile(candidateId, profileHash, metadataURI), limited to the registrar's account, stores a new record, identified by candidate ID, while rejecting any attempt at updating an existing record, thus making stored records immutable; VerifyProfile(candidateId, profileHash) compares a submitted hash with the stored hash and returns a boolean value; and FetchProfile(candidateId) retrieves the stored hash, metadata link, and timestamp. The latter two functions are limited to the registrar and the employers that have been given access by the registrar (GrantAccess/RevokeAccess). Only the 32 bytes of the hash and the metadata URI are stored on the blockchain, while the resume text and the skills list are stored off-chain. The smart contract was benchmarked using 200 dummy candidate profiles on a local Ethereum-compatible development network (Hardhat Network) that mimics the Ethereum mainnet and public testnets in terms of execution environment and gas calculation; the contract has not been deployed anywhere else yet. The results are shown in Table 7 (Section 4.4).

3.13. Reproducibility

Reproducibility is ensured by the availability of the preprocessing pipeline, training setup, evaluation code, the code for the smart contract, and benchmark code at https://github.com/Farwa-01/resume-skill-ner-blockchain. The dataset used is the publicly available NER Annotated CVs corpus from Kaggle (see Data Availability Statement).

4. RESULTS AND DISCUSSION

This section will analyze the RoBERTa-based skill extraction model along with the secure data processing module using Blockchain in great detail with respect to their performance, training dynamics, an ablation study at the scale of the whole dataset, and comparison with state-of-the-art techniques.

4.1. Overall Quantitative Performance

Performance of the model on the validation dataset is reported in Table 2 and Figure 5. Since the values for precision and recall are quite similar (0.640 and 0.666, respectively), the F1 score calculated for entities is 0.653, which indicates that there is no significant bias towards either false positives or false negatives. Considering that the resumes were not preprocessed before being used in this experiment, it is quite obvious that such an F1 score is inferior to those reported on standard datasets such as CoNLL-2003. Recall value exceeds precision, so the model is slightly better at identifying skill mentions than at discarding false ones.

Table 2. Overall Model Performance

Metric	Score
Precision	0.640
Recall	0.666
F1-score	0.653
Accuracy	0.910

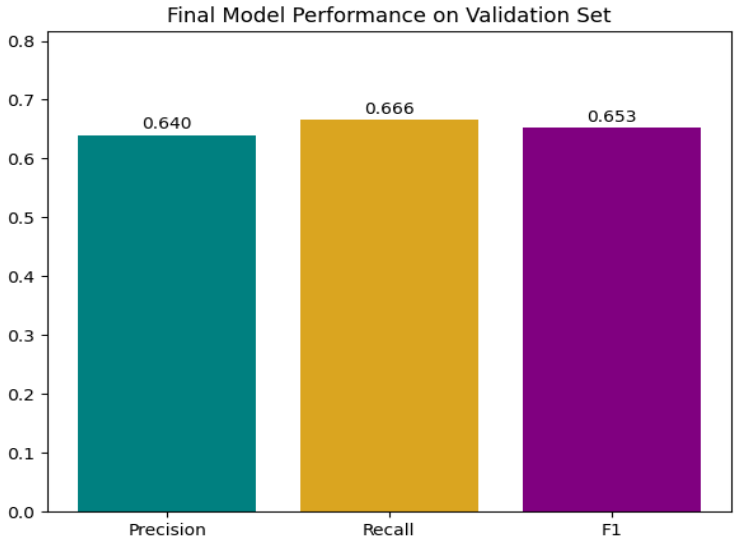


Figure 5. Final Model Performance on the Validation Set (Precision, Recall, F1-score)

4.1.1. Confusion Matrix Analysis

The token-level confusion matrix (Table 3, Figure 6) was calculated on the entire validation dataset (165,238 tokens) using predictions stored for the final model. The majority class O is detected with 95.6% accuracy per class (96,503/100,956), and I-SKILL tokens within the entity are recognized with 90.2% (32,107/35,577), whereas B-SKILL – the starting token of a skill mention is recognized only in 75.6% of cases (21,697/28,705; Figure 7). In cases where the B-SKILL token is not detected, the class O is predicted instead in 5,723 (19.9% of all B-SKILL tokens) occurrences, more frequently than the class I-SKILL (1,285 occurrences, 4.5%): thus, the inability to recognize the start of a skill mention is the most important weakness of the model. There are mistakes in the opposite direction as well: 3,674 tokens of the class O are detected as B-SKILL and 779 tokens as I-SKILL, which results in spurious or shifted entities. In combination with the

missed starts of skills, these mistakes are the reason why precision and recall rates of entity-level F1-score are lower than 0.67 despite the accuracy rate of 0.910. These results suggest two specific areas for future research (Section 5): training objective/class weighting, which makes false negative predictions of B-SKILL less profitable, and a post-processing boundary expansion procedure, which would reconsider O-tokens that precede an I-SKILL with high confidence.

Table 3. Token-Level Confusion Matrix on the Validation Set (rows: actual label; columns: predicted label)

Actual \ Predicted	**O**	**B-SKILL**	**I-SKILL**
O (n = 100,956)	96,503	3,674	779
B-SKILL (n = 28,705)	5,723	21,697	1,285
I-SKILL (n = 35,577)	2,214	1,256	32,107

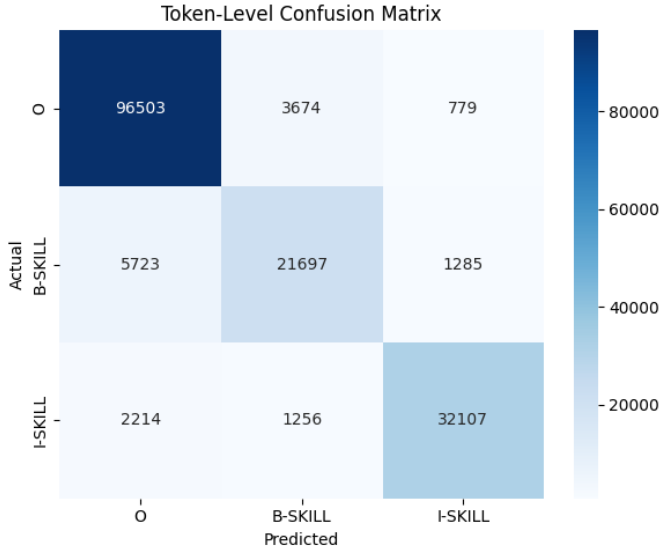


Figure 6. Token-Level Confusion Matrix Heatmap (O / B-SKILL / I-SKILL)

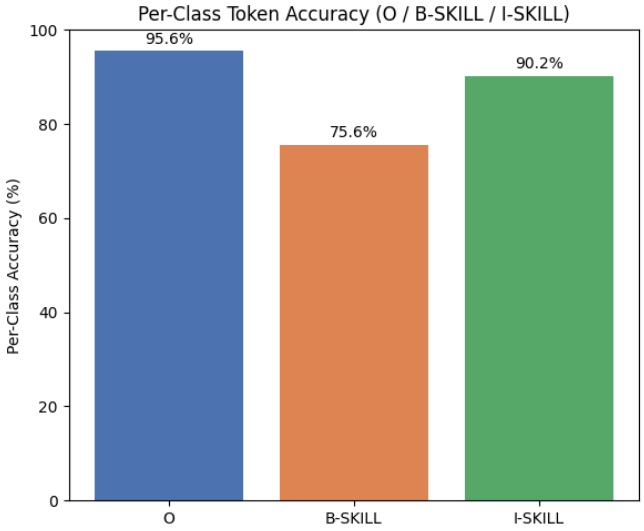


Figure 7. Per-Class Token Accuracy (O / B-SKILL / I-SKILL)

4.2. Training Dynamics and Convergence

The development of F1 for validation through all 14 epochs and three stages of training is presented in Table 4 and Figure 8. Training proceeds very fast during the first stage (epochs 1–5 at the rate of 2e-5) as F1 increases from 0.527 to 0.632. F1 increases further to 0.651 during the second stage (epochs 6–11 at the

rate of 5e-6) but with decreasing gain per epoch. The third stage (epochs 12–14 at the rate of 1e-6) gives an additional 0.002 (to 0.653) to F1, which suggests that the model had reached convergence by that time and further improvement involved fine-tuning of boundary detection. This pattern of progressively smaller gains is typical of transfer learning with pretrained transformer models.

Table 4. Validation F1-Score Across Epochs (Combined Across All Three Training Stages)

Epoch	F1-score
1 (Stage 1)	0.527
2 (Stage 1)	0.588
3 (Stage 1)	0.615
4 (Stage 1)	0.630
5 (Stage 1)	0.632
6 (Stage 2)	0.637
7 (Stage 2)	0.644
8 (Stage 2)	0.647
9 (Stage 2)	0.651
10 (Stage 2)	0.650
11 (Stage 2)	0.651
12 (Stage 3)	0.652
13 (Stage 3)	0.653
14 (Stage 3)	0.653

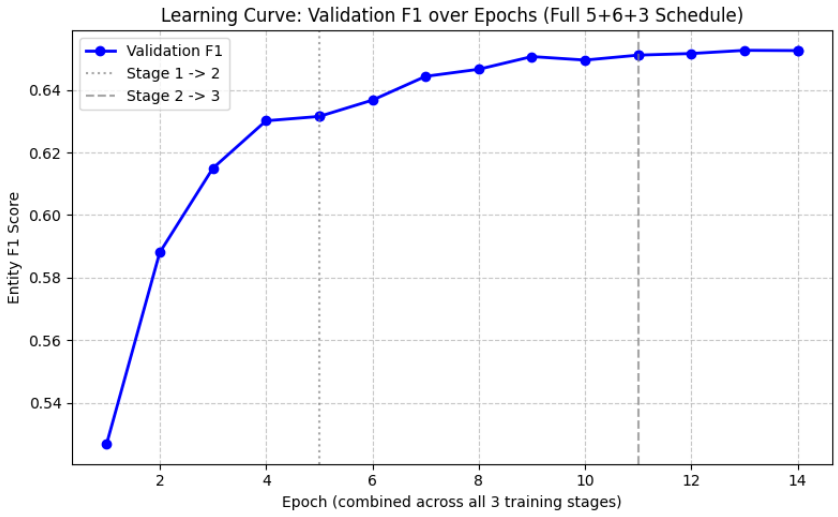


Figure 8. Learning Curve: Validation F1 over 14 Epochs (Full 5 + 6 + 3 Schedule)

4.3. Impact of Dataset Scale (Ablation Study)

Table 5 rows employ the identical one-step schedule (maximum 8 epochs with a learning rate of 2e-5) and identical 498-resume validation split; thus, the 4,473-CV row is different from Table 1 in that it employs the three-step (5 + 6 + 3 epochs) schedule described in Section 3.9. The model's performance as it learns from progressively larger amounts of the training data is depicted in Table 5 and Figure 9. When just 50 CVs were used to train the model, precision, recall, and F1 score are all equal to 0.000: the model simply cannot generalize based on such a small volume of training data. F1 score improves with growing training set size starting from 500 CVs: it is 0.477 for 500, 0.497 for 1,000, 0.560 for 2,000, 0.598 for 3,000, and 0.631 for the maximum available training set size, 4,473 CVs. The data seems to show no signs of saturation at all with respect to the F1 score, implying that even more improvement is possible if more training data becomes available.

Table 5. Dataset Size vs. Model Performance

Dataset Size (training resumes)	Precision	Recall	F1-score
50 CVs	0.0000	0.0000	0.0000
500 CVs	0.4477	0.5114	0.4774
1,000 CVs	0.4728	0.5244	0.4972
2,000 CVs	0.5426	0.5786	0.5600
3,000 CVs	0.5849	0.6122	0.5982
4,473 CVs (full training set)	0.6196	0.6422	0.6307

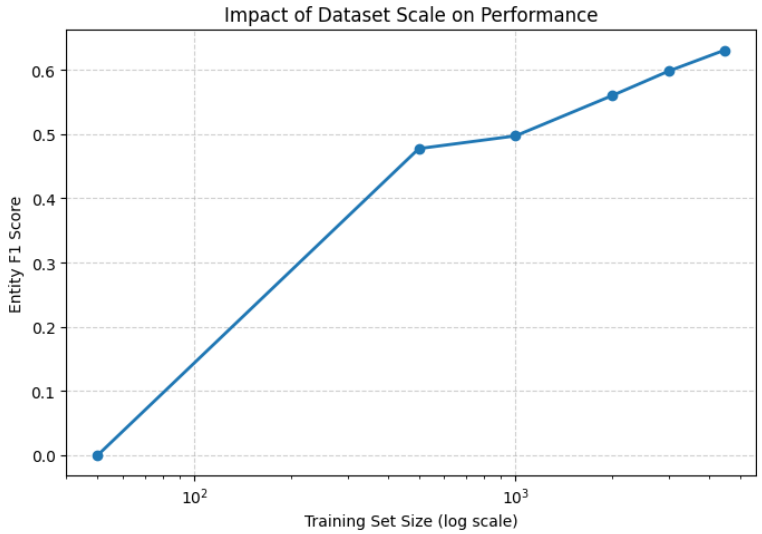


Figure 9. Impact of Dataset Scale on Performance (Log Scale)

4.4. Blockchain Layer Evaluation

System features with and without the use of blockchain technology are described in Table 6. The processes of hashing and on-chain storage take place as an additional step following the process of extraction; therefore, they do not influence the performance of the NER extraction process itself. However, as seen from the table, they still increase the burden for the overall pipeline slightly. In recruitment systems where sensitive information about candidates needs to be stored, the use of blockchain technology improves data integrity, tamper-resistance, and transparency. Given these benefits, the small computational burden is a fair trade-off.

Table 6. Blockchain Impact Analysis

Feature	Without Blockchain	With Blockchain
Data Integrity	Low	High
Tamper Resistance	No	Yes
Transparency	Limited	Full traceability
Processing Speed	Fast	Slight overhead
Trust Level	Moderate	High

The overhead numbers given in Table 6 are captured in numerical form in Table 7, based on executing the contract as described in Section 3.12 on a local Ethereum-compatible development network (Hardhat Network), measuring the time for 200 call pairs between StoreProfile and VerifyProfile operations; gas is obtained from the transaction receipt for the StoreProfile operation, while gas consumed by the off-chain VerifyProfile and FetchProfile operations is estimated. We provide a request payload size measurement instead of a client memory measurement since the client maintains no state, and a process-level memory measurement would capture the memory measurement of the benchmarking framework, not the client. The numbers provided are measured in a local network without any block confirmation time; hence the latency on

a live blockchain is dominated by block time, so the figure above should be read as execution and client overhead rather than latency on a live chain. We report gas rather than a currency cost because gas prices fluctuate; for scale, on 17 October 2026 the Etherscan gas tracker showed a standard price of about 0.065 Gwei, at which one StoreProfile call (about 161,052 gas) would cost roughly 1.05×10^{-5} ETH (about US$0.03 at roughly US$2,431 per ETH). Even with this overhead, the security benefits justify the cost for applications where authenticity and auditability are paramount.

Table 7. Measured Blockchain Layer Overhead

Metric	Measured value
Avg. end-to-end latency per profile (StoreProfile + VerifyProfile)	2.52 ms (SD 0.72 ms, n = 200)
Throughput (store-and-verify cycles)	≈ 398 profiles/s (single-threaded, local network)
Client-side payload per request	≈ 100 bytes (32-byte SHA-256 hash, candidate ID, metadata URI)
Avg. gas: StoreProfile	161,052 gas (SD 4.0, n = 200)
Avg. gas: VerifyProfile	29,339 gas if sent as a transaction; no gas cost as a read-only call
Avg. gas: FetchProfile	38,265 gas if sent as a transaction; no gas cost as a read-only call
GrantAccess (per employer address)	45,092 gas
Contract deployment (one-time)	630,932 gas

4.4.1. Privacy Implications of On-Chain Storage

The storage of candidate-related hashes and metadata on a public blockchain implies unique privacy concerns different from those associated with a standard centralized database. Although the proposed solution in Section 3.12 includes merely the storage of a SHA-256 hash and a metadata pointer, but not the actual resume text nor the list of skills, there are still at least three possible privacy issues: (i) hash linkability, as a consequence of the use of a stable hash for a fixed candidate identifier, which would allow an observer to track the activity of a candidate based on multiple requests for verification made by different employers; (ii) metadata exposure, assuming that the off-chain metadata pointer or its usage pattern contains some private information about a candidate, for instance, who requested verification of the resume and when, because of the inherent nature of most public blockchains to have the timings and participants of transactions visible; and (iii) the right to erasure, as the issue of blockchain immutability contradicts data protection policies (such as GDPR “right to be forgotten”) stipulating that a candidate can ask to delete his/her data at will. However, while the benchmarked contract allows storage only in the registrar’s account and access to the stored data only from granted addresses, it does not yet support salted hashes and does not allow for revoking a stored record; in addition, it was benchmarked against a permissioned EVM, analogous to a public chain. Taking all of the above risks into account, we suggest that in a production environment, this framework be deployed on either a permissioned or a consortium chain with access-controlled read endpoints, and that the use of salting or rotating the hash and a revocation function in the contract be regarded as a necessary hardening procedure (Section 5). It does not mean that the integrity benefits, presented in Table 6, become irrelevant, but these factors must be considered when choosing a deployment configuration for any production environment.

4.5. Comparison with Existing Methods

Table 8 provides a comparison between the proposed framework and other popular approaches for named entity recognition, including rule-based, statistical, and deep learning approaches [7][17][9] to situate it in relation to prior work. The numbers below, based on different data sets, labeling schemes, and evaluation procedures, should be seen as a rough positioning of our approach in relation to existing approaches rather than a controlled experiment between them. For reference, these numbers represent the performance figures typically observed in earlier literature in this domain.

The general trend observed in Table 8, where the performance improves when rule-based methods are replaced by transformer-based models, is consistent with well-established trends reported in NER literature, and the results from the proposed model are consistent with these trends in general. Rule-based matching using manually crafted keyword lists, and being incapable of generalizing due to the lack of ability to deal with differences in resume formats and phrasing, achieves the worst performance. It is improved with CRF models, which use the interdependencies between adjacent tokens [7], but their performance on noisy

resume texts is limited because they depend on hand-crafted features and do not understand the deep context of the input. BiLSTM-CRF models make one more step forward by using bidirectional context [17]. With the use of pre-trained contextualized embeddings, which show higher token interconnectivity, transformer-based models like BERT [9] outperform other models. Among the proposed approaches, the proposed RoBERTa-based approach achieves the highest F1-score, which is consistent with recent research on transformers and their application to domain-specific named entity recognition [23] and the superior performance of RoBERTa due to its optimal pretraining and dynamic masking [10].

Table 8. Comparison with Baseline Approaches

Model Type	Precision	Recall	F1-score
Rule-based [16]	0.40	0.35	0.37
CRF [7]	0.52	0.48	0.50
BiLSTM-CRF [17]	0.58	0.55	0.56
BERT-based [9]	0.63	0.61	0.62
RoBERTa (Proposed)	0.640	0.666	0.653

The proposed approach utilizes a blockchain-based data management component, thus distinguishing itself from prior resume parsing approaches beyond skill extraction performance. The approach makes extracted skill profiles tamper-proof and independently verifiable through an Ethereum-based system rather than storing extracted skill data in a centralized database [19][24].

4.6. Bias and Fairness Considerations

An important and well-established risk of automated recruitment processes is the risk of bias being encoded and amplified by algorithms [25]; this system is subject to the same concerns. There are two types of potential bias relevant to the system under discussion. The first one is extraction bias: given that the model was trained on a particular annotated dataset, skills that appear in unconventional phrasing, non-Western nomenclature, or are linked to educational or regional backgrounds underrepresented in the annotated sample of 4,473 resumes might be extracted with less fidelity than skills common to the corpus, which would put candidates from such backgrounds at a disadvantage even though the model does not make a hiring decision itself. The second type is downstream bias: because the skills profiles are expected to serve as input for employer selection and matching, any consistent pattern of extraction bias correlated with demographic or regional writing styles will inevitably be carried through to recruitment decisions, despite integrity guarantees offered by the blockchain layer.

However, the current research lacks any analysis of extraction performance in terms of demographics and geographical regions, as this dataset does not have any such tags. Fairness audits, along with other methods, such as testing skill phrases through counterfactuals and analyzing subgroups' performance, are recognized as a key research area in the future (see Section 5).

5. CONCLUSION

In this research paper, the architecture developed for automatically extracting skills from resumes is made up of a blockchain layer with Ethereum and a transformer NER with RoBERTa. Two challenges of current recruiting solutions have been solved by this research paper: no security solution to protect the retrieved information of the candidates and the generalizability of the system towards noisy resumes. The blockchain layer provides secure storage, integrity checks, and controlled access to the extracted skill profiles, while RoBERTa has been fine-tuned on a manually annotated dataset with 4,971 real-life resumes.

As has been seen to be true of other recent advances made in the realm of NER research using transformer architecture, the proposed framework scored an entity-level F1-score of 0.653 on the validation dataset, which is better than what has been previously recorded for rule-based, CRF, BiLSTM-CRF, and BERT-based frameworks (this is a rough estimation, not a controlled one; see Table 8). The boundary detection of extracted entities was improved through the use of the preprocessing and post-processing techniques used in this study. Using minimal overhead (around 2.52 milliseconds per profile and around 161,052/29,339 gas to store/verify a profile on a local testing network; see Table 7), the blockchain technology enabled the extracted skill profiles to have transparency, traceability, and tamper-resistance, thus showing that the approach is suitable for the hiring context.

Indeed, the model implies that the recruitment pipeline may leverage a process that includes not only extraction of skills from the entity but also verification of the fact that something was extracted in a form that cannot be tampered with. This point is very important since, as shown in the literature review in Chapter 2, so

far these capabilities have been implemented independently from each other. This may be especially useful in situations where automation along with a chain of custody is needed.

Nevertheless, some limitations are to be addressed, and they are as follows:

- An entity-level F1 of 0.653 certainly indicates a potential for improvement, especially in the case of polysemous and multi-word skill expressions, where "Java" may stand for either a programming language or just another proper name, depending on the context of the sentence (error analysis based on the confusion matrix in Section 4.1.1 shows that missed skill-start tokens (B-SKILL, 75.6%) constitute the problem, not errors in tag assignment within already detected mentions (I-SKILL, 90.2%))
- The evaluation framework was only tested on resumes written in English; therefore, its applicability to other languages is unclear.
- The evaluation relies on a single 90:10 train/validation split without test data or any kind of cross-validation; also, there is no measure of inter-annotator agreement in the corpus (see Section 3.8).
- The baseline results provided in Table 8 have been cited from previous studies on various datasets and protocols; they place the proposed framework approximately but don't offer a direct comparison.
- The data scale ablation (Table 5, Figure 7) is based on one-step scheduling, and the F1 score doesn't seem to reach saturation at the training set scale, so the value of using an even larger annotated dataset cannot be estimated.
- The blockchain overhead for Table 7 was assessed using the local Ethereum-compatible test network and not a deployed public or private blockchain network; a live blockchain adds extra confirmation time for blocks that is not accounted for in this benchmarking.
- Using hashes and metadata in a public blockchain comes along with privacy and data removal issues mentioned in Section 4.4.1 that cannot be completely addressed by the current architecture.
- There has been no audit of the system for demographic and geographic sub-groups (Section 4.6).

Consequently, future work must investigate: multilingual and cross-lingual versions of the transformers; k-fold cross validation, a second annotation round for an overlapping sample of documents that will provide an inter-annotator agreement metric, and an expanded dataset scale analysis; a carefully conducted re-implementation of the rule-based model, the CRF, BiLSTM-CRF and BERT baselines in the same split; boundary-aware training objective and second-pass expansion heuristic to mitigate the identified problem of B-SKILL as discussed in Section 4.1.1; re-estimation of the blockchain layer metrics presented in Table 7 on a production permissioned or consortium test network; salting hash and on-chain revocation approach to resolve the privacy concerns raised in Section 4.4.1; fairness auditing in terms of demographic and geographic considerations; and, finally, improved post-processing of polysemic or multi-word skill expression. We see this current framework as setting the groundwork for these extensions.

DATA AVAILABILITY STATEMENT

The dataset used in this study, *NER Annotated CVs*, is publicly available on Kaggle at https://www.kaggle.com/datasets/mehyarmlaweh/ner-annotated-cvs. The source code supporting this study, including the model training and evaluation scripts, smart-contract implementation, and benchmarking scripts, is publicly available in the GitHub repository at https://github.com/Farwa-01/resume-skill-ner-blockchain.

ACKNOWLEDGEMENTS

The authors thank the Faculty of Computer Science, Riphah International University, and the Department of Computer Science, National University of Modern Languages (NUML), Islamabad, Pakistan, for institutional support during this research. No specific grant, project, or fund number is associated with this work.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] M. A. Aleisa, N. Beloff, and M. White, "Implementing AIRM: a new AI recruiting model for the Saudi Arabia labour market," *J Innov Entrep*, vol. 12, no. 1, p. 59, Sep. 2023, doi: 10.1186/s13731-023-00324-w.

[2] J. S. Black and P. van Esch, "AI-enabled recruiting: What is it and how should a manager use it?," *Bus Horiz*, vol. 63, no. 2, pp. 215–226, Mar. 2020, doi: 10.1016/j.bushor.2019.12.001.

[3] E. Faliagka, A. Tsakalidis, and G. Tzimas, "An integrated e-recruitment system for automated personality mining and applicant ranking," *Internet Res*, vol. 22, no. 5, pp. 551–568, Oct. 2012, doi: 10.1108/10662241211271545.

[4] A. Singh, C. Rose, K. Visweswariah, V. Chenthamarakshan, and N. Kambhatla, "PROSPECT," in *Proceedings of the 19th ACM international conference on Information and knowledge management*, New York, NY, USA: ACM, Oct. 2010, pp. 659–668. doi: 10.1145/1871437.1871523.

[5] I. Ifakir, N. Mohtaram, E. H. Nfaoui, A. Zannou, and M. El Hassouni, "An Approach Based on Named Entity Recognition and Semantic Analysis for Recruitment Efficiency and Optimization," *Int J Adv Comput Sci Appl*, vol. 16, no. 8, 2025, doi: 10.14569/IJACSA.2025.0160889.

[6] M. Zhang, K. Jensen, S. Sonniks, and B. Plank, "SkillSpan: Hard and Soft Skill Extraction from English Job Postings," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2022, pp. 4962–4984. doi: 10.18653/v1/2022.naacl-main.366.

[7] C. Sutton and A. McCallum, "An Introduction to Conditional Random Fields," *Found Trends® Mach Learn*, vol. 4, no. 4, pp. 267–373, Aug. 2012, doi: 10.1561/2200000013.

[8] C. Cortes and V. Vapnik, "Support-vector networks," *Mach Learn*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "No Title," in *Proceedings of the 2019 Conference of the North*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[10] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 2019.

[11] A. Gnehm, E. Bühlmann, H. Buchs, and S. Clematide, "Fine-Grained Extraction and Classification of Skill Requirements in German-Speaking Job Ads," in *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2022, pp. 14–24. doi: 10.18653/v1/2022.nlpcss-1.2.

[12] A. Akkasi, "Job description parsing with explainable transformer based ensemble models to extract the technical and non-technical skills," *Nat Lang Process J*, vol. 9, p. 100102, Dec. 2024, doi: 10.1016/j.nlp.2024.100102.

[13] Z. Zheng, S. Xie, H. N. Dai, X. Chen, and H. Wang, "Blockchain challenges and opportunities: a survey," *Int J Web Grid Serv*, vol. 14, no. 4, p. 352, 2018, doi: 10.1504/IJWGS.2018.095647.

[14] M. Sharples and J. Domingue, "The Blockchain and Kudos: A Distributed System for Educational Record, Reputation and Reward," 2016, pp. 490–496. doi: 10.1007/978-3-319-45153-4_48.

[15] M. Turkanovic, M. Holbl, K. Kosic, M. Hericko, and A. Kamisalic, "EduCTX: A Blockchain-Based Higher Education Credit Platform," *IEEE Access*, vol. 6, pp. 5112–5127, 2018, doi: 10.1109/ACCESS.2018.2789929.

[16] X. Yue, H. Wang, D. Jin, M. Li, and W. Jiang, "Healthcare Data Gateways: Found Healthcare Intelligence on Blockchain with Novel Privacy Risk Control," *J Med Syst*, vol. 40, no. 10, p. 218, Oct. 2016, doi: 10.1007/s10916-016-0574-6.

[17] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural Architectures for Named Entity Recognition," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2016, pp. 260–270. doi: 10.18653/v1/N16-1030.

[18] F. Souza, R. Nogueira, and R. Lotufo, "BERTimbau: Pretrained BERT Models for Brazilian Portuguese," 2020, pp. 403–417. doi: 10.1007/978-3-030-61377-8_28.

[19] S. Wang, L. Ouyang, Y. Yuan, X. Ni, X. Han, and F.-Y. Wang, "Blockchain-Enabled Smart Contracts: Architecture, Applications, and Future Trends," *IEEE Trans Syst Man, Cybern Syst*, vol. 49, no. 11, pp. 2266–2277, Nov. 2019, doi: 10.1109/TSMC.2019.2895123.

[20] E. F. Tjong Kim Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task," in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003* -, Morristown, NJ, USA: Association for Computational Linguistics, 2003, pp. 142–147. doi: 10.3115/1119176.1119195.

[21] M. U. Ul Haq, P. Frazzetto, A. Sperduti, and G. Da San Martino, "Improving Soft Skill Extraction via Data Augmentation and Embedding Manipulation," in *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, New York, NY, USA: ACM, Apr. 2024, pp. 987–996. doi: 10.1145/3605098.3636010.

[22] C. Chandra, Y. Ojima, M. V. Bendarkar, and D. N. Mavris, "Aviation-BERT-NER: Named Entity Recognition for Aviation Safety Reports," *Aerospace*, vol. 11, no. 11, p. 890, Oct. 2024, doi: 10.3390/aerospace11110890.

[23] J. Lee *et al.*, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020, doi: 10.1093/bioinformatics/btz682.

[24] K. Salah, M. H. U. Rehman, N. Nizamuddin, and A. Al-Fuqaha, "Blockchain for AI: Review and Open Research Challenges," *IEEE Access*, vol. 7, pp. 10127–10149, 2019, doi: 10.1109/ACCESS.2018.2890507.

[25] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, "Mitigating bias in algorithmic hiring," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, New York, NY, USA: ACM, Jan. 2020, pp. 469–481. doi: 10.1145/3351095.3372828.

BIOGRAPHIES OF AUTHORS



Faria Karamat is currently serving as a Lecturer at the Faculty of Computing of Riphah International University, Islamabad, Pakistan. She is interested in the field of data science, machine learning, federated learning, privacy-preserving artificial intelligence, and distributed intelligent systems and their application in healthcare and biomedical informatics. She can be contacted at email: faria.karamat@riphah.edu.pk



Farwa Nawaz is a lecturer in the Department of Computer Science, National University of Modern Languages (NUML), Islamabad, Pakistan. The area of her research interest is machine learning, deep learning, uncertainty-aware artificial intelligence, and prediction modeling for renewable energy and biomedicine, specifically medical image analysis and risk prediction. She can be contacted at email: farwa.nawaz@numl.edu.pk



Shawal Khaliq is a student in the Department of Software Engineering, Fatima Jinnah Women University, Rawalpindi, Pakistan. The area of her research interest is Artificial Intelligence and Healthcare, with a particular focus on Machine Learning, Deep Learning, and Incremental Learning for medical disease diagnosis also interested in Large Language Models (LLMs), Natural Language Processing (NLP), and multimodal AI for healthcare, particularly their application to automated and structured medical report generation. She can be contacted at email: shawalkhaliq@gmail.com