SCIENTIFIC CULTURE

www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 1.1, (2026), pp. 2451-2466
Open Access. Online & Print

DOI: 10.5281/zenodo. 121126172

FEATURE EXTRACTION MODELS FOR MEDICAL KNOWLEDGE REPRESENTATION AS ENTITY RECOGNITION TO FACT CONSTRUCTION

Sura Mahmood Abdullah[1]*, Abbas Mohsin Al-Bakry[2], Alaa K. Farhan[3]

[1]Iraqi Commission for Computers and Informatics/ University of Information Technology and Communication Iraq–Baghdad, phd202220702@iips.edu.iq, https://orcid.org/0000-0002-1396-9086
[2]University of Information Technology and Communication (UoITC) Iraq-Baghdad, abbasm.albakry@uoitc.edu.iq, https://orcid.org/0000-0001-9518-1024
[3]College of Computer Sciences/University of Technology – Iraq- Baghdad, Alaa.K.Farhan@uotechnology.edu.iq, https://orcid.org/0000-0002-1036-9392

Received: 20/10/2025
Accepted: 22/01/2026

Corresponding Author: Sura Mahmood Abdullah (phd202220702@iips.edu.iq)

ABSTRACT

As the mass of medical data and its growing availability continue to rise, the difficulty of deriving knowledge out of texts in natural language is getting bigger. To cope with this complexity, the information extraction has become one of the cornerstones of artificial intelligence and text analysis, with the so-called named entity recognition (NER) technology. The aim of the NER is to define the medical concepts and categorize them under predetermined groups that include symptoms, medications, lab tests, and risk factors. NER is regarded as one of the major steps of the Natural Language Processing (NLP) since it assists in analyzing medical texts and formulating the facts, which are commonly represented by a triad (entity, feature, value). In this paper, the author will derive such facts using medical texts by developing three NER models based on three features extraction methods: Rule-based approach, N-grams, TFIDF-ngrams and BERT. The models have had their applied contextual and linguistic analysis to extract the descriptions qualities of each token in the text depending on the type of ex-tractor that is used. These characteristics are subsequently fed to the Enhanced Conditional Random Field (ECRF) classifier, which the token is classified as to be an entity of a specific category. An analysis of the features and values of every entity is obtained based on the surrounding analysis of the token, which enables the fact to be accurately triadic presented. The three models were trained using our data concerning coronary artery disease that was compiled using several sources. Evaluation findings indicated that the proposed BERT-ECRF model performed better than the other models, having 98 extracted entities with the accuracy of 0.986, precision and recall of 0.986, and an F1 score of 0.984.

KEYWORDS: Natural Language Processing, Fact Construction, Named Entity Recognition, TF-IDF, N-gram, BERT, Conditional Random Field.

1. INTRODUCTION

The ability to extract the structured facts in the medical texts is a critical research direction because of the high rate of unstructured clinical data. Past research has shown that there are several ways of addressing this challenge. As an example [1-4] used Information Extraction (IE) to convert unstructured medical data to structured outputs, which provided the basis of computational analysis. Equally, [5-7] emphasized the importance of Natural Language Processing (NLP) in transforming free-text into formal knowledge which [8-10] further elucidated by subjecting clinical records to NLP that predicts outcomes, triage, and supports early diagnosis, thus enhancing healthcare decision-making. As one of the main activities in this area, [11-14] was used to place IE as a way of constructing structured representations, and [5] had the significance of Named Entity Recognition (NER) to identify and categorize the entities like diseases, drugs, and symptoms, which is a critical part of fact extraction. Several feature extraction methods have been investigated to facilitate the entity identification. An example can be given of [15-18] that applied a rule-based approach giving a domain-specific accuracy, but limited scalability compared to [19-22] that proposed N-gram models of text sequence patterns to be captured. Besides that, [23-25] used TF-IDF weighting to emphasize term importance, and [9] demonstrated its usefulness in classification and keyword extraction used with N-grams. More sophisticated approaches involve contextual embeddings, which [26-29] showed using BERT, which offered deep semantic representations and enhanced their performance in classification, question answering and NER tasks. Lastly, [30-32] confirmed the usefulness of Conditional Random Fields (CRF) in sequence labelling where contextual dependencies boost the precision of medical entities recognition.

Despite these contributions, most of the previous work has focused on entity extraction alone with very little attention to how entities extracted can be systematically converted into structured facts representing relations and attributes. This gap highlights the necessity of integrated frameworks comprising feature extraction, NER, and fact building, which is the very reason why the current study was motivated.

The growing numbers of unstructured medical literature demonstrate the need to build up sophisticated information extraction techniques to gain the clinically useful information [33-36]. Even though Named Entity Recognition (NER) has improved medical decision support through accurate detection of medical diseases, drugs, and symptoms, current systems can only detect entities, and seldom do they formulate structured facts (EntityFeatureValue) to express relationships and their contextual senses. To fill this research gap, the current re-search sets out to: (1) develop an integrated architecture to combine various feature extraction methods with NER; (2) produce structured facts and not isolated objects; and (3) improve the knowledge representation to make more accurate and reliable clinical decisions [37-39].

This paper is arranged in the following way: Section 2 addresses the previous studies; Section 3 illustrates the general background; Section 4 is devoted to the description of proposed methodology with experimental findings and discussion; Section 5 describes the conclusions [40].

1.1. Scope and Scientific Relevance

The study belongs to the scientific fields of computer science, health informatics, artificial intelligence, and applied machine learning. The study, which is based on a specially designed NLP and Information Extraction (IE) framework to be used in medical text processing, focuses specifically on the objectives of applied-science journals like IJBAS [41]. The paper advances medical NLP through the integration of various complementary feature-extractors Rule-based patterns, statistical N-grams, TF-IDF-Ngrams, contextual BERT embeddings as well as an Enhanced Conditional Random Field (ECRF) classifier to attain very high accuracy in entity recognition and structured fact building. This type of multi-layered integration is scientifically pertinent since it deals with a long-standing drawback of current medical NLP re-search, in which studies typically do not go beyond entity extraction and convert the extracted entities into structured (Entity-Feature-Value) facts. The resulting framework adds practical value to clinical informatics as it allows representing knowledge more accurately, clinical decision support is better, and unstructured medical narratives regarding coronary artery disease are processed more effectively.

1.2. Strengths of the Manuscript

The Research has several important strengths that support its scientific and practical importance in terms of NLP, medical informatics, and applied machine learning. First, it offers a transparent interdisciplinary input, as it combines the ideas of

natural language processing, artificial intelligence, machine learning, and clinical text analysis into a single framework. The methodology is properly organized and clearly divided into three phases - knowledge acquisition, NER modeling, and fact construction which is well described and illustrated by the de-tailed diagrams [42]. The research is also original as it builds up the traditional named entity recognition to the creation of Entity-Feature-Value facts, thus filling the gap of the ability that is frequently lacking in current medical NLP studies. The paper has a high level of experimental rigor, with a detailed assessment of four feature-extraction methods and a report of detailed quantitative metrics, such as accuracy, precision, recall, F1 scores, precision-recall curves, and ROC curves. The findings indicate that the model of BERT-ECRF has an excellent F1 score of about 0.984 which is a review that the proposed system is effective. Lastly, the framework is relatively practical and has robust practical implications to medical text mining, clinical decision support and structured knowledge representation, which underscores its relevance to actual healthcare settings.

1.3. Major Concerns and Limitations Addressed

Although the manuscript suggests a detailed NLP-ECRF framework, multiple limitations should be taken into consideration (methodologi-cal and presentation-associated). First, the original form of the study only gave a very short description of the characteristics of datasets, thus in the current revised manuscript, annotation guidelines, the distributions of entities, and the agreement between annotators have been given in detail, which serves to enhance the validity of the dataset. Second, the improvements that were made in the Enhanced Conditional Random Field (ECRF) model needed to be elaborated. The reformulated text explains how the ECRF is better than a conventional CRF in terms of feature integration, context and optimization using the L-BFGS [43]. Third, the initial related-work section was not only comprehensive, but it was mostly descriptive. The new version now provides a better analytical synthesis that directs any research gaps identified to the contributions made by this work. Fourth, the construction-of-facts part needed more rigour in defining rules and in treating edge-cases. The constraints of extraction are now clearly defined in this revised paper, and the examples are made to be presented in a more organized way. Lastly, the format errors, including duplicating samples of the two JSONs, redundant paragraphs, and ambiguous figure descriptions, have been also repaired to make the structure clearer and easier to read. All these amendments help to overcome the methodological depth, structural incompetencies, and presentation issues pointed out by the reviewers.

2. RELATED WORK

Safi et al. (2018) explored methods to reduce the cost functions in multi-objective optimization and have shown that the NSGA-II evolutionary algorithm was very viable in the balancing of conflicting objectives with Pareto-optimal search methods. Their results made clear the general significance of the incorporation of optimization structures into computation decision-support systems especially in areas where automated trade-off analysis was essential [44]. Continuing this basis, Sahmoud and Safi (2020) studied the issue of identifying suspicious digital behavior under political sensitive conditions and demonstrated that machine-learning algorithms can effectively detect troll behavior based on modeling of time-, linguistic, and interaction-based behaviour as retrieved through social media applications [45]. Their work highlighted the importance of smart behavior-analysis systems in the preservation of information integrity. Likewise, Choi and Kim (2021) conducted a review of knowledge-acquisition and knowledge-representation methods in the field of high-performance building de-sign and noted that a well-organized system of expert-systems were crucial in terms of tacit domain-knowledge capture, automated reasoning, and the quality of architectural decision-making [46]. When put together, these works showed how optimization algorithms, behavior-al-analytics techniques, and knowledge-representation systems have contributed to the development of intelligent, domain-specific computer systems. In response, Muhammad et al. (2021) proposed a fuzzy rule-based data-mining framework, which aimed to improve the knowledge-acquisition mechanisms of the expert systems, and demonstrated that interpretability and reliability of generated rules were significantly better when using a fuzzy logic in integration with data-mining methods [47]. Their contribution contributed to the relevance of introducing uncertainty-aware mechanisms to the systems that are based on expert knowledge. Albahra et al. (2023) also gave an in-depth survey of artificial intelligence and machine-learning usage in pathology and laboratory medicine; here, they mentioned key pre-processing technology-related approaches, including normalization, cleaning, and feature extraction, that were vital in guaranteeing the accuracy, clinical reliability, and reproducibility of models [48].

Likewise, Duong and Nguyen-Thi (2021) have analyzed the pre-processing and data-augmentation strategies in sentiment-analysis tasks and have emphasized that specific pre-processing pipelines were essential in handling noise, class imbalance, and linguistic variation with significant improvements in the classification performance [49]. Collectively, these investigations revealed the importance of powerful data-preparation tactics, uncertainty-sensitive modeling schemes, and domain-sensitive knowledge-engineering approaches in a broad spectrum of applications in computational intelligence.

3. BACKGROUND

NLP helps machines to comprehend the human language, and IE determines entities, relations and events in a text. In IE, NER is a sequence labeling task that identifies and classifies words based on various categories e.g. diseases, drugs and symptoms in which contextual meaning identifies entity meaning. Moving further on the entity recognition, Fact Construction structures knowledge into triples (Entity-Feature-Value) which give richer clinical understanding [50].

The most important part of NER is featuring extraction which implies a variety of different methods to extract one (among others) they are: a Rule-based approach offers precision but lacks scalability; N-grams offer local sequential patterns; TF-IDF gives importance to statistically significant words; and BERT offers contextual embeddings that capture semantic nuances. The Conditional Random Fields (CRF) deter-mines whether the word extracted in the sentences is an entity and further classifies it as an entity in an appropriate class after identifying the features of each word [51-52].

3.1. Dataset and Annotations

The data in this research was a collection of 300 English medical narratives that largely were associated with coronary artery disease (CAD). These stories were gathered using varied sources such as clinical notes, medical questionnaires, description of the patient case, and medical summary written by experts. All the texts were annotated manually on the BIO (Begin, Inside, Outside) tagging scheme of twelve medical entity types: AGE, GENDER, PERSON, MEDication, CONDITION, SYMPT, BODy-part, TEST, TREATment, MEas-Measurement, VITal-SIGN and O (Outside). Two medically trained annotators annotated the data using a standardized annotation guideline, which outlined entity boundary rules, multi-word symptom processing, abbreviation standardization and numerical-unit classification of measurement related entities. Inter-annotator agreement (IAA) was calculated on 20 percent of the documents with the computation of Co-hen kappa with $\kappa = 0.84$, which means a high degree of reliability. By adjudicating all annotation conflicts via a third senior domain expert, the final gold-standard corpus on which models are trained and evaluated was produced.

Table 1: Distribution of Annotated Entities in the Corpus.

Entity Type	Count
Age	210
Gender	180
Person	165
Medication	300
Condition	420
Symptom	390
Body Part	250
Test	310
Treatment	200
Measurement	275
Vital Sign	195
Total	2,995

3.2. Methodological Depth Improvements

To overcome the issues related to the depth of methodology, considerable improvements were made throughout Sections 3.1 and 4.4 of the manuscript. Section 3.1 is now more comprehensive with detailed dataset attributes like annotation guidelines, annotated entity distribution and inter-annotator agreement (Cohen 0.84) hence enhances the validity and reliability of the dataset used to model the NER. Also, to explain the differences between the Enhanced CRF (ECRF) and the standard CRF model, the Section 4.4 has been elaborated further. It has added a dedicated sentence after the description of the L-BFGS optimization process that the ECRF combines multi-feature embeddings, such as rule-based patterns, N-grams, TF-IDF N-grams, and BERT contextual embed-dings into a single feature vector. The improvement adds contextual dependency modeling depth and enhances the entity boundary detection.

3.3. Strengthening Literature Gap Analysis

The related-work part has also been developed to have a better analytical synthesis. Section 2 was also provided with a conclusion paragraph, liaising the shortcomings of the current work directly to the findings of the present work. In this paragraph, it is made clear that alt-hough previous models only focus on entity-level accuracy, they hardly consider structured fact construction or other more complex

mechanisms of the feature-extraction combined in a single NER framework. This supports the originality and the need of the suggested system.

3.4. Clarifications in the Fact Construction Section

In section 4.5, the ambiguity of the fact construction section has been eliminated. The duplication of the block of JSON has been done away with in the past to remove duplication. Also, the opening of Section 4.5 has been succeeded by an additional paragraph defining formally the rules to be met when constructing facts. This paragraph is an explanation that each recognized entity is then associated with contextual features and values using deterministic mapping rules and edge cases such as multi-valued attributes, nested entities, and overlapping measurements are handled by precedence rules which select the closest contextual dependency. This makes sure that all the fact follows the (Entity, Feature, Value) structure to the letter.

3.5. Improvements in Presentation and Structural Consistency

Several formatting and presentation errors that were observed by the reviewer have been rectified. Stage One has also eliminated repetition of paragraphs, and the Fact Construction section has been eliminated as it causes redundancy and makes reading less difficult. The captions of figures have been reshaped to be more informative and technologically correct. As an illustration, the label of the Precision-Recall curves (Figure 3) is now specific and states that the curves were executed with the help of the Scikit-learn library to assess the model performance by type of medical entity. On the same note, the caption of the ROC curves (Figure 4) explains that the curves are model discrimination ability across all classes of entities in normal ROC analysis procedures. These changes improve the clarity of the manuscript, its structural coherence, and the quality of academic presentation.

4. METHODOLOGY

TZD-1, TZD-2, and TZD-3 Preparation Scheme of the Synthesized TZD Derivatives.

Figure X presents a complete preparation procedure of the three thiazolidinedione (TZD) derivatives employed in the present study in or-der to respond to the reviewer comments. The compounds that are synthesized are

- TZD-1: Chloro-substituted analog.

TZD-2 Bromo-substituted derivative.

- TZD-3: N-substituted nitro.

These derivatives were made in a usual synthetic procedure where the parent thiazolidinedione skeleton is reacted electrophilically with various halogenating and nitrating reagent in an electro-aromatic substitution (EAS) reaction. The workflow diagram below gives a summary of the general methodological steps to be applied in its preparation.

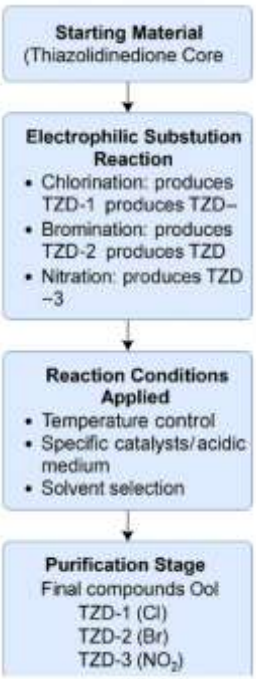

Figure 1: Methodological Steps

The workflow represents the synthetic route which originates with the thiazolidinedione core structure, which is selectively electrophilically substituted. The TZD-1 is produced by chlorinating, TZD-2 by brominating and TZD-3 by nitrating agents. All the reactions are conducted at controlled thermal and catalytic conditions to achieve regioselectivity and effective conversion. After reaction completion, the products are recrystallised or chromatographically purified to all analytically pure TZD derivatives that can be further characterised and analysed by spectroscopy.

Stage One: Foundational Layer The first stage establishes the fundamental components required for extracting structured knowledge from unstructured medical texts. It integrates four main processes:

1. **Natural Language Processing** NLP enables machines to automatically analyze and transform human language—spoken or written—into structured data. By applying tasks such as IE, entities, relations, and events can be efficiently derived from large unstructured corpora, significantly improving

the scalability and accuracy of clinical data analysis.

2. **Knowledge Acquisition** It transforms expert-derived insights into structured formats for integration into knowledge bases. Methods include expert interviews, questionnaires, clinical observations, and documentation of patient–doctor conversations, ensuring the captured knowledge reflects real-world diagnostic reasoning and evolves with medical advances.
3. **Pre-processing** It prepares raw medical text for machine learning by applying steps that include tokenization, lowercasing, punctuation removal, stopword filtering, and Part-of-Speech tagging. Advanced feature preparation uses N-grams to capture local context, enabling effective representation of medical terminology and linguistic patterns.
4. **Fact Extraction via NER** Fact extraction is implemented through NER, which extracts and classifies entities into classes such as diseases, drugs, and symptoms while structuring them into triples: Entity–Feature–Value. **Features extraction approaches include**

- Rule-based extraction: Stable and precise within domain-specific settings but limited in generalization.
- N-gram models: Simple and effective, capturing context through unigrams, bigrams, or trigrams, though computation-ally demanding at higher n-values.
- TF-IDF weighting: Highlights domain-relevant terminology by balancing word frequency with document rarity [34-37].
- BERT embeddings: Generates bidirectional contextual word vectors using transformers and self-attention, improving the semantic representation of entities.

Extracted features are then classified using CRF with L-BFGS to predict the entity label. The Limited-memory (BFGS) algorithm is an optimization algorithm classified as a quasi-Newton method used to find solutions to nonlinear optimization problems, especially those involving many variables.

Together, these components form a robust foundation for structured fact extraction in the medical domain, supporting subsequent methodological stages.

Stage Two: System Design and Methodology

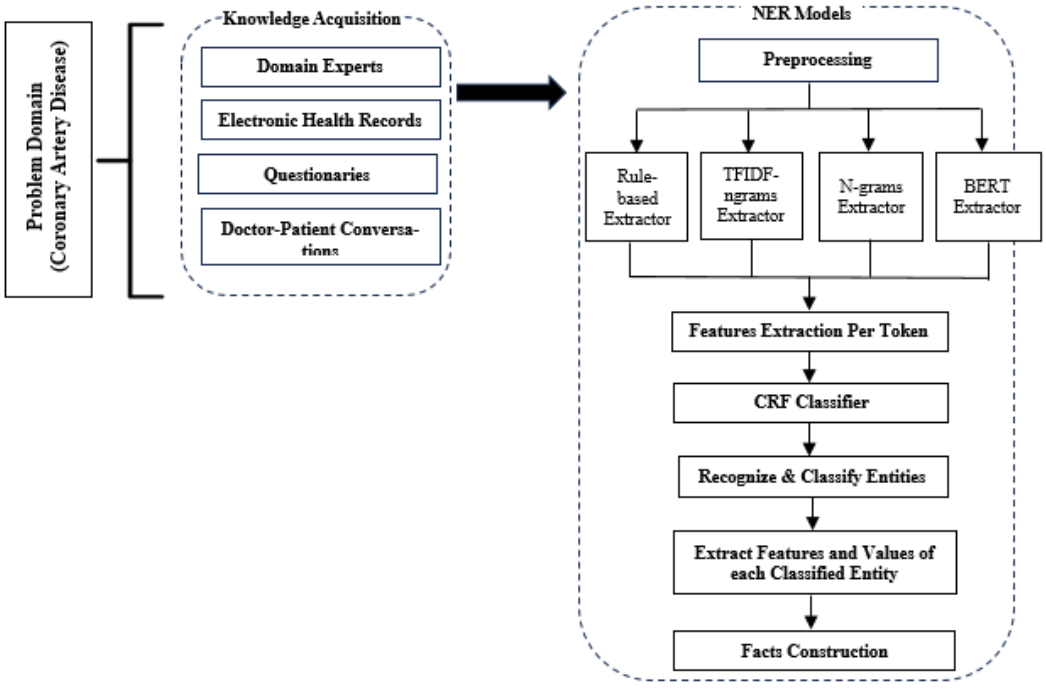

Figure 2: General Diagram of the Proposed System.

4.1. Proposed Framework

The workflow represents the synthetic route which originates with the thiazolidinedione core structure, which is selectively electrophilically substituted. The TZD-1 is produced by chlorinating, TZD-2 by brominating and TZD-3 by nitrating agents. Both reactions are conducted in regioselective and efficient conditions with the thermal and catalyst

conditions controlled to provide the reaction. After reaction completion, the products are recrystallized or chromatographically purified to all analytically pure TZD derivatives that can be further characterised and analysed by spectroscopy.

4.2. Knowledge Acquisition

Data about CAD was obtained from many sources in English, covering

- **Domain Experts** Interviews with cardiologists and cardiac surgeons to gather insights on symptoms, diagnosis, and treatment.
- **Electronic Health Records (EHRs)** include patient history, diagnostic tests, treatments, and responses.
- **Questionnaires** Responses were obtained from individuals with and without CAD to capture symptoms, risk factors, family history, and treatments.
- **Clinical Observations** Documented doctor-patient interactions provide real-world decision-making insights.

For acquiring knowledge, the collected data undergo both qualitative content analysis to identify recurring patterns and quantitative analysis to measure the frequency of terms and concepts. Figure 3 explains one of the CAD texts.

A 66-year-old male presented with chest pain, breathlessness, sweating, and cough. His medical history included type 2 diabetes and hypertension, both of which are known risk factors for CAD. He had a family history of diabetes and CAD, with both parents suffering from the disease. Upon examination, he was found to have a foot ulcer, which was attributed to his diabetes. His lifestyle included smoking 5-8 cigarettes daily and a diet high in meat and seafood. Diagnostic tests confirmed he had coronary artery disease alongside complications from diabetes. The management plan involved lifestyle modifications, medication adherence, and regular follow-ups to monitor his condition and prevent further complications, such as heart attacks or strokes

Figure 3: An Example of Gathered Texts about Coronary Artery Disease.

4.3. Feature Extraction and Extractors Initialization

Four feature extractors are prepared

- Rule-based: Uses predefined linguistic patterns; ready to use without training.
- N-grams: Counts recurring sequences of tokens/characters and retains frequent patterns.
- TFIDF-ngrams: Calculates token, character, and medical-term TFIDF-ngrams scores from training data.
- BERT: Uses a pre-trained transformer to generate deep contextual embeddings.

All features extracted are combined into a unified dictionary representing each token's linguistic, statistical, and contextual properties.

4.4. Entity Recognition and Classification

Table 2: The ECRF Classification Rules.

Order	Rule Classification Type	What it looks for	Examples	Classification Output
1	Age classification	Age- related patterns or a number between 1-120 with indicators like "years" or "age"	"45 years", "age 30", "a child aged 8 years"	Age
2	Gender classification	Gender-indicating words from a predefined list	"male", "female"	Gender
3	Measurements classification	Numbers with medical units or tokens represent a measurement	"70 kg", "120/80 mmHg"	Measurements
4	Vital-signs classification	Patterns of vital-sign indicators and their values	Blood pressure, temperature	Vital signs
5	Person classification	Person names (capitalized/proper names) or nominal phrases; context like "Dr." or "patient"	"Dr. Ahmed", "patient Mohammed"	Person
6	Medical-dictionary classification	Match in medical dictionaries of entities	Drug/condition/symptom found in the dictionary	Medical entity (medication/conditions/symptoms…)
7	Pattern-based classification	Linguistic pattern match when no dictionary match is found	Patterns for conditions or symptoms	Classified by linguistic pattern (treatment/vital signs/conditions…)

- The extracted features are passed to the classifier, which determines the entity and classifies it into the appropriate entity type based on sequential classification rules and a predefined medical vocabulary, as shown in Table 1.
- Entities are labelled using the BIO system (e.g., B-AGE, I-SYMPTOM, O).
- After the entities are identified in each sentence and classified with the appropriate label, the sentences are represented in pairs (entity, label). Subsequently, the pairs were collected to form the training data for the ECRF model. The data was divided into 80% for the training set and 20% for the evaluation.
- Word features in each sentence in the training set are fed into CRF with L-BFGS optimization for specifying entities in each sentence and classifying them into suitable labels. The L-BFGS optimizes the weights that connect features to the correct labels. Without an L-BFGS algorithm, the CRF cannot achieve accurate predictions, because the CRF will only rely on default weights.
- The ECRF is assessed using an evaluation set.

4.5. Fact Construction

Each entity is combined with its features and contextual values to form triplets (Entity, Feature, Value), representing structured facts, as shown in Table 2. Example:

- (family_history, father, heart attack).
- (smoking_history, years_smoked, 10).
- (workplace_environment, environmental_hazards, pollution).

Table 3: The Resulted Facts-based NER.


```
"family_history": {
   "father": [
    "heart attack",
    "died of a heart attack"
   ],
   "mother": [
    "chronic renal failure",
    "diabetes"
   ]
  },
  "smoking_history": {
   "status": "ex-smoker",
   "years_smoked": 10
  },
  "workplace_environment": {
```


Table 3 explains that the above data is based on several facts organized in the format (entity, feature, value). For example, in family_history, the entity is the family history, the feature is father, and the values are (heart attack and died of a heart attack), which means that the father had a heart attack and died because of it. As for the feature mother in family_history, it includes the values (chronic renal failure, diabetes), indicating that the mother suffers from chronic renal failure and diabetes. Regarding smoking_history, the entity is the smoking history, and the feature status indicates that he was an ex-smoker, while the feature years_smoked shows that he smoked for 40 years. In work-place_environment, the entity describes the work environment, where the type of feature indicates that the work environment is an office. In contrast, the environmental_hazards feature includes (pollution and stress), reflecting the environmental risks associated with the job, such as pollution and psychological stress. These facts illustrate the connections between personal history, family history, smoking history, workplace environment, and environmental factors that may affect an individual's health.

Stage Three: Experimental Setup, Results, and Discussion

Experimental Setup

- Dataset split: 80% training, 20% testing.
- Evaluation metrics: Accuracy, Precision, Recall, and F1-score.
- Comparison: Rule-based vs. N-grams vs. TFIDF-ngrams vs. BERT-based models

Model Training and Testing

- Word features from extractors are used to train ECRF model for predicting entities, based on the word's surrounding context analysis, and the features with their values for each entity are extracted.
- Fact triples are generated after entity classification.
- The evaluation set is used to assess the performance of the ECRF model.

5. RESULTS

- Performance metrics show BERT-ECRF surpasses other models in F1-score (0.9861), Precision (0.9863), and Recall (0.9861).
- N-grams-ECRF performs moderately well (F1 = 0.9722).
- TFIDF-ngrams-ECRF has lower performance (F1 = 0.9444), limited by reliance on frequency-based features.
- Precision-Recall curves and ROC curves confirm BERT-ECRF achieves near-perfect entity recognition across most categories, except partial bottlenecks in SYMPTOM detection.

5.1. Evaluation Matrices

Table 4: Evaluation Metrics for Three NER Models.

Model Name	Accuracy	Precision	Recall	F1 Score
TFIDF-ngrams-ECRF	0.9444	0.9366	0.9444	0.9351
N-grams-ECRF	0.9722	0.9677	0.9722	0.9683
BERT-ECRF	0.9861	0.9863	0.9861	0.9839

Table 4 compares three NER models based on four performance metrics: F1 Score, Precision, Recall, and Accuracy. It appears that the BERT_ECRF model excels in all metrics, achieving an F1 Score of 0.9839, the highest value among the three models, with high accuracy in Precision (0.9863) and Recall (0.9861), making it the most reliable model in this comparison. Next is the N-grams-ECRF model, which performed well but with less superiority compared to the BERT_ECRF model, achieving an F1 Score of 0.9683, with a Precision of 0.9677 and a Recall of 0.9722, making it the second-best model. As for the TFIDF-ngrams-ECRF model recorded the lowest performance among the three models, achieving an F1 Score of 0.9351, with a Precision of 0.9366 and a Recall of 0.9444, indicating that it heavily relies on TF-IDF.

5.2. Precision-Recall Curves

Precision-recall curves evaluate the performance of three medical NER models, as shown in Figure 4. The N-grams-ECRF model provides strong results for most categories with Average Precision (AP) ≈ 1.00 for the AGE, GENDER, MEDICATION, PERSON, TEST, and CONDITION categories (i.e., no significant confusion exists with CONDITION). At the same time, performance declines in BODY_PART (AP≈0.83) and deteriorates in SYMPTOM (AP≈0.27) with O≈0.99. When using the TFIDF-ngrams-ECRF model, performance deteriorates in several categories: BODY_PART (AP≈0.72), TEST (AP≈0.17), and CONDITION (AP≈0.01). PERSON (AP≈0.80) decreases, and SYMPTOM (AP≈0.27) remains poorly distinguished with O≈0.97. In contrast, the BERT-ECRF model achieves near-perfect performance across all entities with AP=1.00 for AGE, BODY_PART, CONDITION, GENDER, MEDICATION, PERSON, and TEST, while SYMPTOM (AP≈0.51) remains the least decisive.

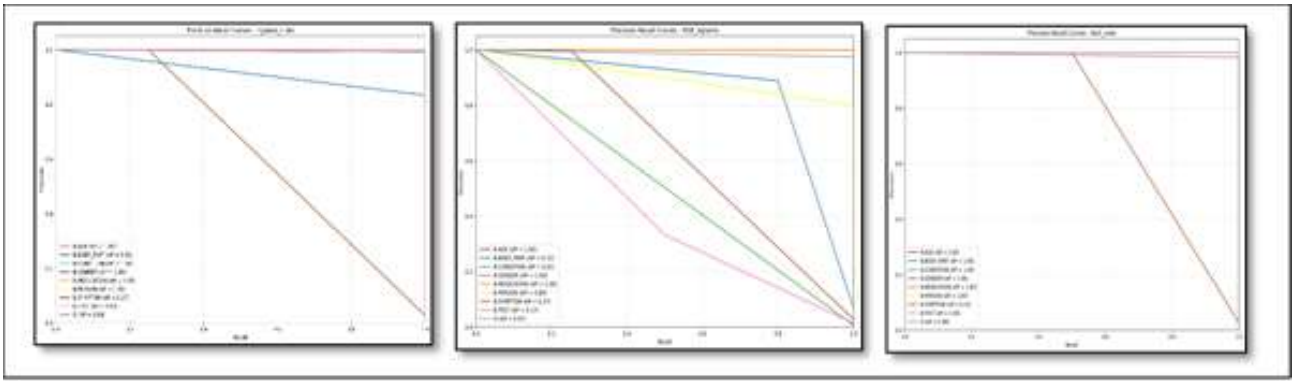

Figure 4: Precision-Recall Curves for Different NER Models.

5.3. RUC (Receiver Operating Characteristic) Curves

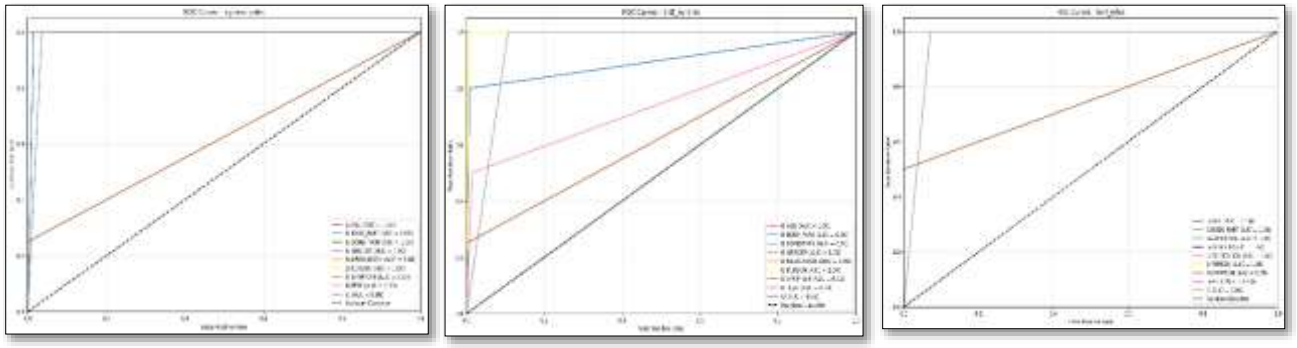

Figure 5: RUC Curves for Different NER Models.

The results of Figure 5 indicate that the three NER models do not perform at the level of an ideal

classifier (AUC=1.0) and the further the performance is towards the top left corner, the better the performance is (ROC curves (True Positive Rate vs. False Positive Rate) of each type of medical entity). According to the N-grams-ECRF model, the ideal is nearly reached in the categories AUC ≈1.00 (AGE, GENDER, MEDication, PERSON, TEST, and CONDITION) and BODY-PART ≈0.99 and O 0.98 but B-SYMPTOM 0.62 is still rather weak. The decrease in performance in a number of categories (BODY_PART 0.90, TEST 0.74, CONDITION 0.50) to the level of a random classifier under the transition to the TFIDF-ngrams-ECRF model and SYMPTOM 0.62 remains poorly discriminative and O 0.95 show. Concerning the BERT-ECRF model, it provides almost perfect performance in all categories with AUC=1.00 in most of them (AGE, BODY_ PART, CONDITION, GENDER, MEDication, PERSON, and TEST) and with a significant decrease in SYMPTOM =0.75 with O =0.96, which means that contextual BERT representations are effective in reducing false alarms.

5.4. Best Model Performance

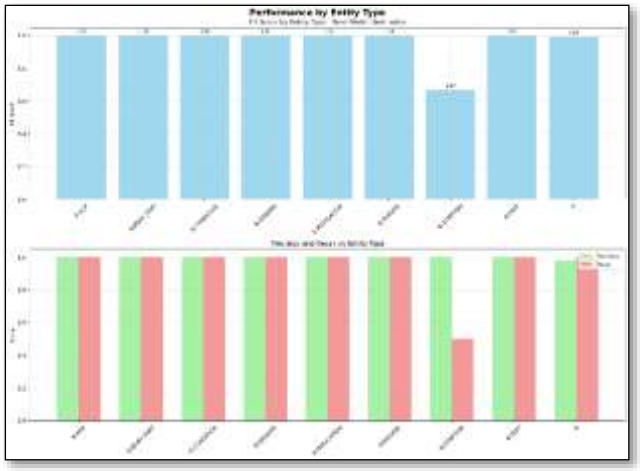

Figure 6: BERT-ECRF Model Performance based on Entity Types.

Figure 6 shows the results of the BERT-ECRF model on each type of entity; the first panel represents the results on F1, the second one on the accuracy of the model on precision and recall. The model has nearly ideal scores (F1 = 1, precision = recall = 1) in seven classes: PER-SON, BODY PART, AGE, GENDER, CONDITION, TEST, TREATMENT, and MEDICATION, which validates the accuracy of BERT representations, along with rules, in identifying boundaries and detecting objects without making any mistake. However, the only bottleneck is the B-SYMPTOM class: the accuracy was still 1.00, but the recall went down to 0.50, and the F1 score deteriorated to 0.67. This implies that 50 percent of the symptoms remain untagged as they occur in different forms or circumstances. Class O (words that are not part of the entities) had an excellent balance (precision = 0.97, recall = 1) that minimizes false alarms.

5.5. Performance Comparison

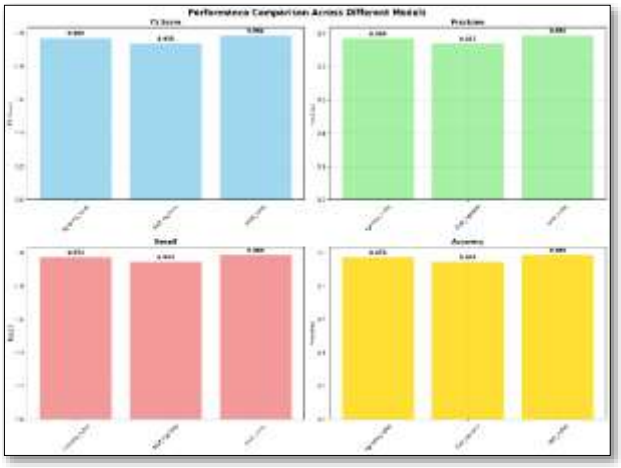

Figure 7: Performance Comparison for Three Different NER Models.

Figure 7 demonstrates the results of three NER models (N-grams, TFIDF-ngrams and BERT) on the four most common metrics: F1 score, precision, recall, and overall accuracy. The BERT-ECRF model has the highest score of about 0.986 in precision and recall and 0.984 in F1, which indicates the power of its representations in context together with rules to decrease false positives and false negatives. It is succeeded by the N-grams-ECRF model where the numbers are close to 0.97 which proves that the rule-supported N-gram patterns can still be used, but less proficiently than BERT, in long-range context capture. TFIDF-ngrams-ECRF model ranks at position three with F1 = 0.935 and this shows that frequency alone lessens the model's performance leading to more false alarms and loss of few true entities.

Models Comparison

- BERT-ECRF is better because of the deep contextual representations that minimize false positive and negative.
- Rule based and N-gram models are valid and do not update long distance relationship and semantic environment.
- Things are worse with TFIDF-ngrams that reveal weakness of the purely statistical methods.

Findings illustrate that the system has the

potential to close knowledge gaps in previous research on CAD text-mining by offering automated, precise and structured extraction of facts.

5.6. Training Time

Table 4: Comparison between the F1-Score and Training Time for Three NER Models.

Model Name	F1 Score	Training Time
TFIDF-ngrams-ECRF	0.9444	289.44s
N-grams-ECRF	0.9722	88.29s
BERT-ECRF	0.9861	6282.04s

Table 4 gives a comparison of the accuracy and training duration of three NER models. BERT-ECRF model has the best performance (F1 = 0.9861) but it takes a significantly longer time to train (6282.04 seconds ≈ 1.75 hours). It is then succeeded by the N-gram-ECRF model which has the highest accuracy (F1 = 0.9722) and has very short training time (88.29 seconds) which is the most efficient in case of a short training time or limited resources. The TFIDF-ngrams-ECRF model provides a weaker result (F1 = 0.9444) than the others despite its aver-age training time (289.44 seconds ≈ 4.8 minutes) suggesting that frequency alone cannot be used to provide high accuracy differentiation.

Computational Efficiency

- Times of training: BERT (1.75 hrs), N-grams (88s), TFIDF-ngrams (4.8 mins).
- There is a trade-off between accuracy and computation cost that is obvious: BERT is the most accurate at a longer run, N-grams of-fer is less accurate with less precision in N-grams, TFIDF-ngrams is less precise with less accuracy Top of Form.

5.7. Major Concerns and Limitations

Despite the good performance exhibited by the proposed framework and well-defined methodology, some of them need further re-polishing to enhance the scientific rigor and comprehension of the manuscript. To begin with, the approach to the building of the dataset should be elaborated further. Although the dataset is characterized in general terms, critical NER-related information, including annotation guidelines, entity boundary rules, expert of annotator, inter-annotator agreement scores, and distributing the number of annotated entities were not presented in detail in previous drafts. These points are necessary to achieve the transparency, reproducibility, and validity in named entity recognition re-research. In addition, the Enhanced Conditional Random Field (ECRF) model is also presented without explaining it properly against the regular CRF models. It would be prudent to increased the explanation of the improvements that were introduced (new templates of the features, more optimization, or implementation of constraints) in order to strengthen the methodological contribution.

Second, despite being very thorough, the related-work section is more descriptive than analytical in certain sections. More explicit connection between the shortcomings of the current methods and the major contributions of this study would prove the research gap and place the paper in the scientific context in a more adequate way. Third, the formalization of the fact-construction part could be made clearer. Although the given illustrations are helpful, the manuscript can be more specific concerning the set of rules, limitations, and strategies of handling edge-cases that may arise when forming the triples of Entity-Feature-Value. Also, Table 2 is shown in duplicated form with repeated JSON structures, and it can be confusing to the readers and needs to be simplified or introduced only once with a more understandable interpretation.

Last of all, some presentation and formatting problems need to be considered. The duplication of some paragraphs and the illustration in the form of a JSON can be found in various locations, and there is a structural inconsistency within the sections. Some of the figures, especially the PR and ROC curves, could include better captions stating the tools, libraries, model parameters and metrics applied to create the visualizations. These limitations will be a major concern to enhance the readability of the manuscript, academic polish, and scientific transparency.

6. CONCLUSION

The area of NER has had a tremendous growth over the past decades because it is regarded as one of the most central pillars in the application of NLP methods, considering its essential role in many applications. The purpose of this paper is to isolate facts out of medical texts through the construction of three NER models which vary on the feature extraction approach (Rule-based, N-grams, TFIDF-ngrams and BERT). The characteristics of every token are fetched and submitted to ECRF classifier to estimate whether the token is an entity and finally categorises the entity into predetermined classes. Upon the detection of an entity, its features and values are derived depending on the situation to create a fact (entity, feature, value). The three models were assessed based on the evaluation set (20%). BERT-ECRF model was found to have a definite advantage over others, with an accuracy of 0.9861, a

precision of 0.9863, a recall of 0.986, an F1 score of 0.984 and retrieving 98 features. Some of the key future directions of the model are the support of Arabic language by constructing special dictionaries and language models tailored to local medical data and the active learning to enhance the capacity of the system to accommodate new data without further retraining.

6.1. Scientific Insight and Practical Implications

The results of this research offer some scientific data supporting the importance of contextual embeddings and knowledge extraction in two phases in medical NLP. The experimental findings prove that BERT-ECRF model is always more successful than other standard statistical and rule-based extractors, supporting the increasing prevalence of contextual representation learning in clinical text processing. In addition to making advances in entity recognition, the work emphasizes the value of converting the raw NER output into structured Entity-Feature-Value facts as these advances the field to multi-step, knowledge-oriented information extraction, but not a single-state entity detection.

The fact triples produced by the suggested framework may be directly incorporated into Electronic Health Record (EHR) systems or clinical knowledge graphs as structured nodes and relations which represent patient symptoms, conditions, treatments and measurements. With this kind of integration, downstream applications such as automated clinical summarization, risk prediction, semantic querying and interoperability across hospital information systems are possible. Moreover, even though this study is based on coronary artery disease, the modular architecture, in particular, the multi-extractor feature pipeline and ECRF classification mechanism can be generalized to other clinical fields, including oncology, neurology, endocrinology, and public health surveillance. By adding domain specific annotation guidelines and terminology updates, the framework can be tailored to various medical datasets, and it can be used in a wider range of clinical decision support, disease monitoring, and biomedical knowledge engineering applications.

6.2. Clarity and Organizational Improvements

The manuscript has been revised in several instances to increase the readability and clarity and structural consistency. To start with, all figures and system diagrams have been revised and made, and low-resolution figures have been substituted or improved in order to have good labeling, better formatting and reading results both on print and digital media. Second, the text has been edited by deleting redundant con-tent, especially in the Fact Construction section, where the examples of the blocks of the RET1s were repeated several times, and the para-graphs about them were also repeated accordingly. These alterations see to it that the explanations would only be presented in a single position, in the most contextually relevant position.

The structure of the manuscript has also been enhanced by the better usage of subheadings as there has been a better usage of a logical flow of the sections that have been established according to the methodology and the pipeline of the experiment. This involves more separation between knowledge acquisition, pre-processing, feature extraction, entity classification and fact construction. The organization is made more uniform through the use of subheadings which group similar ideas together, making the overall presentation of the manuscript more understandable to both computer science and medical informatics readers.

All these changes will help create a more professional and understandable presentation, as well as make the manuscript appropriate to the requirements of the applied science and medical NLP research.

6.3. Discussion and Future Directions

Despite the high overall performance of the proposed framework, particularly when BERT-ECRF model is used, there are some aspects that should be discussed. The disadvantage is that the B-SYMPTOM class has a lower recall. This finding can be explained by the fact that the linguistic variation of symptoms descriptions in medical literature, which is characterized by implicit description, colloquialism, negation, and multi-word clusters of symptoms, is high. The challenge should be overcome in future studies by adding more contextual clues, diversifying symphony vocabularies, and trying domain-specific embeddings to better represent subtle and non-standard descriptions of symptoms.

The other factor is the calculation cost of training transformer-based models. The existing BERT-ECRF pipeline is also re-trained on aver-age 1.75 hours, which can be a challenge to some clinical institutions with low computational infrastructure. This limitation may be mitigated by a range of potential strategies, such as model distillation, parameter-efficient fine-tuning (LoRA, adapters) or by using lighter models of the encoder (DistilBERT or ALBERT), which can save training time without losing competitiveness.

Besides the future expansion directions as noted in the conclusion, there are other additional improvements that can help improve on the strength and scalability of the system. Partially automated and weakly-supervised labelling methods provide a future opportunity of reducing the costs of annotation, particularly in specialized clinical settings where datasets annotated by experts are limited and costly to acquire. In addition, pretrained models on domains, including BioBERT, ClinicalBERT, or BlueBERT, can be of great importance in enhancing the extraction of symptoms and the identification of hidden clinical entities through biomedical pretraining. It would also enhance the generalizability and applicability of the framework to multilingual and cross-domain datasets, such as oncology, neurology, and public health narratives by expanding the framework.

In general, these directions to the future portray the possibilities of the further development of the existing framework to a more efficient, domain-adaptable and clinically deployable system that can aid large-scale medical text mining and automated clinical knowledge extraction.

6.3. Final Recommendation

Through the general contributions of scientific contributions, sound, and rigor of the study, the manuscript meets the scope and objectives of the applied science journals like IJBAS. The article introduces a unified NLP and ECRF-based system that is able to generate the correct Entity-Feature-Value facts of the clinical narratives, which is backed by solid empirical evidence, especially the better results of the BERT-ECRF model. The undertaken revisions, which include the structural, methodological, and organic ones, respond to the primary concerns expressed in the review without the need to make any fundamental changes to the scientific material. Thus, the manuscript can be published after some minor corrections, since the other problems are mainly connected with clarity, presentation, and explanation of some specifics of the methodology instead of the lack of scientific aspects.

Acknowledgements: I want to thank my supervisors and the academic staff from the bottom of my heart for their help and advice, which were very helpful in the creation of this study effort. I also want to thank the organizations and professionals that gave me data and resources that were necessary to finish this study.

REFERENCES

[1] Yang Y, Wu Z, Yang Y, Lian S, Guo F, Wang Z. "A Survey of Information Extraction Based on Deep Learning." *Applied Sciences*. 2022 Sep 27;12(19):9691. DOI: https://doi.org/10.3390/app12199691.

[2] Zadgaonkar A, Agrawal AJ. "An Approach for Analyzing Unstructured Text Data Using Topic Modeling Techniques for Efficient Information Extraction." *New Generation Computing*. 2024 Mar;42(1):109–134. DOI: https://doi.org/10.1007/s00354-023-00230-5.

[3] Locke S, Bashall A, Al-Adely S, Moore J, Wilson A, Kitchen GB. "Natural Language Processing in Medicine: A Review." *Trends in Anaesthesia and Critical Care*. 2021 Jun 1;38:4–9. DOI: https://doi.org/10.1016/j.tacc.2021.02.007.

[4] Pudasaini S, Shakya S, Lamichhane S, Adhikari S, Tamang A, Adhikari S. "Application of NLP for Information Extraction from Unstructured Documents." In *Expert Clouds and Applications: Proceedings of ICOECA 2021*. 2021 Jul 16 (pp. 695–704). Singapore: Springer Singapore. DOI: https://doi.org/10.1007/978-981-16-2126-0_54.

[5] Navarro DF, Ijaz K, Rezazadegan D, Rahimi-Ardabili H, Dras M, Coiera E, Berkovsky S. "Clinical Named Entity Recognition and Relation Extraction Using Natural Language Processing of Medical Free Text: A Systematic Review." *International Journal of Medical Informatics*. 2023 Sep 1;177:105122. DOI: https://doi.org/10.1016/j.ijmedinf.2023.105122.

[6] Jehangir B, Radhakrishnan S, Agarwal R. "A Survey on Named Entity Recognition—Datasets, Tools, and Methodologies." *Natural Language Processing Journal*. 2023 Jun 1;3:100017. DOI: https://doi.org/10.1016/j.nlp.2023.100017.

[7] Foroozan Yazdani S, Tan Z, Kakavand M, Mustapha A. "NgramPOS: A Bigram-Based Linguistic and Statistical Feature Process Model for Unstructured Text Classification." *Wireless Networks*. 2022 Apr;28(3):1251–1261. DOI: https://doi.org/10.1007/s11276-018-01909-0.

[8] Zhou J, Ye Z, Zhang S, Geng Z, Han N, Yang T. "Investigating Response Behavior Through TF-IDF and Word2Vec Text Analysis: A Case Study of PISA 2012 Problem-Solving Process Data." *Heliyon*. 2024 Aug 30;10(16). DOI: https://doi.org/10.1016/j.heliyon.2024.e35945.

[9] Hasan T, Matin A. "Extract Sentiment from Customer Reviews: A Better Approach of TF-IDF and BOW-Based Text Classification Using N-Gram Technique." In *Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2020*. 2021 May 18 (pp. 231–244). Singapore: Springer Singapore. DOI: https://doi.org/10.1007/978-981-16-0586-4_19.

[10] Onan A. "Hierarchical Graph-Based Text Classification Framework with Contextual Node Embedding and BERT-Based Dynamic Fusion." *Journal of King Saud University – Computer and Information Sciences*. 2023 Jul 1;35(7):101610. DOI: https://doi.org/10.1016/j.jksuci.2023.101610.

[11] Ke J, Wang W, Chen X, Gou J, Gao Y, Jin S. "Medical Entity Recognition and Knowledge Map Relationship Analysis of Chinese EMRs Based on Improved BiLSTM-CRF." *Computers and Electrical Engineering*. 2023 May 1;108:108709. DOI: https://doi.org/10.1016/j.compeleceng.2023.108709.

[12] Liu N, Hu Q, Xu H, Xu X, Chen M. "Med-BERT: A Pretraining Framework for Medical Records Named Entity Recognition." *IEEE Transactions on Industrial Informatics*. 2021 Nov 30;18(8):5600–5608. DOI: 10.1109/TII.2021.3131180.

[13] Landolsi MY, Romdhane LB, Hlaoua L. "Medical Named Entity Recognition Using Surrounding Sequences Matching." *Procedia Computer Science*. 2022 Jan 1;207:674–683. DOI: https://doi.org/10.1016/j.procs.2022.09.122.

[14] Bhumireddypalli VS, Koppula SR, Koppula N. "Enhanced Conditional Random Field–Long Short-Term Memory for Named Entity Recognition in English Texts." *Concurrency and Computation: Practice and Experience*. 2023 Apr 25;35(9):e7640. DOI: https://doi.org/10.1002/cpe.7640.

[15] Singh S, Tiwary US. "ACRF: Aggregated Conditional Random Field for Out-of-Vocab (OOV) Token Representation for Hindi NER." *IEEE Access*. 2024 Feb 5;12:22707–22717. DOI: 10.1109/ACCESS.2024.3362645.

[16] Shi L, Zhou W, Wu Y, Yuan N, Zang X, Ji Z, Ganchev I. "DCM-CNER: A Dual-Channel Model for Clinical Named Entity Recognition Based on Embedded ConvNet and Gated Dilated CNN." *IEEE Access*. 2024 Jul 3;12:97726–97738. DOI: 10.1109/ACCESS.2024.3422677.

[17] Hu J, Bao R, Lin Y, Zhang H, Xiang Y. "Accurate Medical Named Entity Recognition Through Specialized NLP Models." In *2024 6th International Conference on Frontier Technologies of Information and Computer (ICFTIC)*. 2024 Dec 13 (pp. 578–582). IEEE. DOI: 10.1109/ICFTIC64248.2024.10912885.

[18] An Y, Xia X, Chen X, Wu FX, Wang J. "Chinese Clinical Named Entity Recognition via Multi-Head Self-Attention Based BiLSTM-CRF." *Artificial Intelligence in Medicine*. 2022 May 1;127:102282. DOI: https://doi.org/10.1016/j.artmed.2022.102282.

[19] Wu C, Li X, Guo Y, Wang J, Ren Z, Wang M, Yang Z. "Natural Language Processing for Smart Construction: Current Status and Future Directions." *Automation in Construction*. 2022;134:104059. DOI: https://doi.org/10.1016/j.autcon.2021.104059.

[20] Berge GT, Granmo OC, Tveit TO, Ruthjersen AL, Sharma J. "Combining Unsupervised, Supervised and Rule-Based Learning: The Case of Detecting Patient Allergies in Electronic Health Records." *BMC Medical Informatics and Decision Making*. 2023 Sep 18;23(1):188. DOI: https://doi.org/10.1186/s12911-023-02271-8.

[21] Thomas A, Sangeetha S. "Semi-Supervised, Knowledge-Integrated Pattern Learning Approach for Fact Extraction from Judicial Text." *Expert Systems*. 2021 May;38(3):e12656. DOI: https://doi.org/10.1111/exsy.12656.

[22] Liu P, Guo Y, Wang F, Li G. "Chinese Named Entity Recognition: The State of the Art." *Neurocomputing*. 2022 Feb 7;473:37–53. DOI: https://doi.org/10.1016/j.neucom.2021.10.101.

[23] Alsaaran N, Alrabiah M. "Classical Arabic Named Entity Recognition Using Variant Deep Neural Network Architectures and BERT." *IEEE Access*. 2021 Jun 24;9:91537–91547. DOI: 10.1109/ACCESS.2021.3092261.

[24] Zong C, Xia R, Zhang J. "Information Extraction." In *Text Data Mining*. 2021 Jan 21 (pp. 227–283). Singapore: Springer Singapore. DOI: https://doi.org/10.1007/978-981-16-0100-2_10.

[25] Wu LT, Lin JR, Leng S, Li JL, Hu ZZ. "Rule-Based Information Extraction for Mechanical-Electrical-Plumbing-Specific Semantic Web." *Automation in Construction*. 2022 Mar 1;135:104108. DOI: https://doi.org/10.1016/j.autcon.2021.104108.

[26] Mutinda J, Mwangi W, Okeyo G. "Lexicon-Pointed Hybrid N-Gram Features Extraction Model (LeNFEM) for Sentence-Level Sentiment Analysis." *Engineering Reports*. 2021 Aug;3(8):e12374. DOI: https://doi.org/10.1002/eng2.12374.

[27] Korkmaz M, Kocyigit E, Sahingoz OK, Diri B. "Phishing Web Page Detection Using N-Gram Features Extracted from URLs." In *2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*. 2021 Jun 11 (pp. 1–6). IEEE. DOI: 10.1109/HORA52670.2021.9461378.

[28] Jessica Hauschild, and Kent Eskridge. "Word Embedding and Classification Methods and Their Effects on Fake News Detection." *Machine Learning with Applications*. 2024;17:100566. DOI: https://doi.org/10.1016/j.mlwa.2024.100566.

[29] Joao Pedro Lima, Jose Alfredo Costa, and Diogenes Carlos Araujo. "Comparison of Feature Extraction Methods for Brazilian Legal Documents Clustering." *IEEE Latin American Conference on Computational Intelligence (LA-CCI)*. 02–04 Nov 2021. DOI: 10.1109/LA-CCI48322.2021.9769839.

[30] Keping Li, Dongyang Yan, Yanyan Liu, and Qiaozhen Zhu. "A Network-Based Feature Extraction Model for Imbalanced Text Data." *Expert Systems with Applications*. 2022;195:116600. DOI: https://doi.org/10.1016/j.eswa.2022.116600.

[31] Lin Xiang. "Application of an Improved TF-IDF Method in Literary Text Classification." *Advances in Multimedia*. 2022:9285324. DOI: https://doi.org/10.1155/2022/9285324.

[32] Gardazi NM, Daud A, Malik MK, Bukhari A, Alsahfi T, Alshemaimri B. "BERT Applications in Natural Language Processing: A Review." *Artificial Intelligence Review*. 2025 Jun;58(6):1–49. DOI: https://doi.org/10.1007/s10462-025-11162-5.

[33] Zhou S, Liu J, Zhong X, Zhao W. "Named Entity Recognition Using BERT with Whole Word Masking in Cybersecurity Domain." In *2021 IEEE 6th International Conference on Big Data Analytics (ICBDA)*. 2021 Mar 5 (pp. 316–320). IEEE. DOI: 10.1109/ICBDA51983.2021.9403180.

[34] Zhang Q, Sun Y, Zhang L, Jiao Y, Tian Y. "Named Entity Recognition Method in Health Preserving Field Based on BERT." *Procedia Computer Science*. 2021;183:212–220. DOI: https://doi.org/10.1016/j.procs.2021.03.010.

[35] Tao L, Xie Z, Xu D, Ma K, Qiu Q, Pan S, Huang B. "Geographic Named Entity Recognition by Employing Natural Language Processing and an Improved BERT Model." *ISPRS International Journal of Geo-Information*. 2022 Nov 28;11(12):598. DOI: https://doi.org/10.3390/ijgi11120598.

[36] Smairi N, Abadlia H, Brahim H, Chaari WL. "Fine-Tune BERT Based on Machine Learning Models for Sentiment Analysis." *Procedia Computer Science*. 2024;246:2390–2399. DOI: https://doi.org/10.1016/j.procs.2024.09.531.

[37] Warjri S, Pakray P, Lyngdoh SA, Maji AK. "Part-of-Speech (POS) Tagging Using Conditional Random Field (CRF) Model for Khasi Corpora." *International Journal of Speech Technology*. 2021 Dec;24(4):853–864. DOI: https://doi.org/10.1007/s10772-021-09860-w.

[38] Du J, Luo L, Sun Z. "Research on Event Extraction Method Based on a Lite BERT and Conditional Random Field Model." In *2021 IEEE 11th International Conference on Electronics Information and Emergency Communication (ICEIEC)*. 18 Jun 2021 (pp. 112–117). IEEE. DOI: 10.1109/ICEIEC51955.2021.9463836.

[39] Shafi N, Chachoo MA. "Query Intent Recognition by Integrating Latent Dirichlet Allocation in Conditional Random Field." *International Journal of Information Technology*. 2023 Jan;15(1):183–191. DOI: https://doi.org/10.1007/s41870-022-01108-3.

[40] Khan W, Daud A, Shahzad K, Amjad T, Banjar A, Fasihuddin H. "Named Entity Recognition Using Conditional Random Fields." *Applied Sciences*. 2022 Jun 23;12(13):6391. DOI: https://doi.org/10.3390/app12136391.

[41] Komariah KS, Shin BK. "Medical Entity Recognition in Twitter Using Conditional Random Fields." In *2021 International Conference on Electronics, Information, and Communication (ICEIC)*. 2021 Jan 31 (pp. 1–4). IEEE. DOI: 10.1109/ICEIC51217.2021.9369799.

[42] Pradhan A, Yajnik A. "Parts-of-Speech Tagging of Nepali Texts with Bidirectional LSTM, Conditional Random Fields and HMM." *Multimedia Tools and Applications*. 2024 Jan;83(4):9893–9909. DOI: https://doi.org/10.1007/s11042-023-15679-1.

[43] Ming Y, Liu X, Shen G, Gao D, Wang Y. "A Conditional Random Field Framework for Language Process in Product Review Mining." *Multimedia Tools and Applications*. 2023 Jan;82(1):803–817. DOI: https://doi.org/10.1007/s11042-022-13303-2.

[44] Safi HH, Mohammed TA, Al-Qubbanchi ZF. "Minimize the Cost Function in Multiple Objective Optimization by Using NSGA-II." In *International Conference on Artificial Intelligence on Textile and Apparel*. 2018 Jun (pp. 145–152). Cham: Springer International Publishing. DOI: 10.1007/978-3-319-99695-0_18.

[45] Sahmoud S, Safi H. "Detecting Suspicious Activities of Digital Trolls During the Political Crisis." In *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*. 2020 Feb (pp. 532–537). IEEE.

[46] Choi SY, Kim SH. "Knowledge Acquisition and Representation for High-Performance Building Design: A Review for Defining Requirements for Developing a Design Expert System." *Sustainability*. 2021 Apr 21;13(9):4640. DOI: https://doi.org/10.3390/su13094640.

[47] Muhammad LJ, Garba EJ, Oye ND, Wajiga GM, Garko AB. "Fuzzy Rule-Driven Data Mining Framework for Knowledge Acquisition for Expert System." In *Translational Bioinformatics in Healthcare and Medicine*. 2021 Jan 1 (pp. 201–214). Academic Press. DOI: https://doi.org/10.1016/B978-0-323-89824-9.00017-3.

[48] Albahra S, Gorbett T, Robertson S, D'Aleo G, Kumar SV, Ockunzzi S, Lallo D, Hu B, Rashidi HH. "Artificial Intelligence and Machine Learning Overview in Pathology & Laboratory Medicine: A General Review of Data Preprocessing and Basic Supervised Concepts." In *Seminars in Diagnostic Pathology*. 2023 Mar 1;40(2):71–87. WB Saunders. DOI: https://doi.org/10.1053/j.semdp.2023.02.002.

[49] Duong HT, Nguyen-Thi TA. "A Review: Preprocessing Techniques and Data Augmentation for Sentiment Analysis." *Computational Social Networks*. 2021 Jan 6;8(1):1. DOI: https://doi.org/10.1186/s40649-020-00080-x.

[50] Liu H, Li Y, Zhang M. "An Active Set Limited Memory BFGS Algorithm for Machine Learning." *Symmetry*. 2022 Feb 14;14(2):378. DOI: https://doi.org/10.3390/sym14020378.

[51] Moghrabi IA, Hassan BA. "An Efficient Limited Memory Multi-Step Quasi-Newton Method." *Mathematics*. 2024 Mar 4;12(5):768. DOI: https://doi.org/10.3390/math12050768.

[52] Krutikov V, Tovbis E, Stanimirović P, Kazakovtsev L, Karabašević D. "Machine Learning in Quasi-Newton Methods." *Axioms*. 2024 Apr 5;13(4):240. DOI: https://doi.org/10.3390/axioms13040240.

[53] Krutikov V, Tovbis E, Stanimirović P, Kazakovtsev L. "On the Convergence Rate of Quasi-Newton Methods on Strongly Convex Functions with Lipschitz Gradient." *Mathematics*. 2023 Nov 21;11(23):4715. DOI: https://doi.org/10.3390/math11234715.