Survey: Understand the challenges of Machine Learning Experts using Named Entity Recognition Tools

Florian Freund⬥, Philippe Tamla⬥, and Matthias Hemmje⬥

University of Hagen, Faculty of Mathematics and Computer Science
58097 Hagen, Germany

Abstract. This paper presents a survey based on Kasunic's survey research methodology to identify the criteria used by Machine Learning (ML) experts to evaluate Named Entity Recognition (NER) tools and frameworks. Comparison and selection of NER tools and frameworks is a critical step in leveraging NER for Information Retrieval to support the development of Clinical Practice Guidelines. In addition, this study examines the main challenges faced by ML experts when choosing suitable NER tools and frameworks. Using Nunamaker's methodology, the article begins with an introduction to the topic, contextualizes the research, reviews the state-of-the-art in science and technology, and identifies challenges for an expert survey on NER tools and frameworks. This is followed by a description of the survey's design and implementation. The paper concludes with an evaluation of the survey results and the insights gained, ending with a summary and conclusions.

Keywords: Expert Survey, Natural Language Processing, Named Entity Recognition, Machine Learning, Cloud Computing.

1 Introduction and Motivation

Named Entity Recognition (NER) tools, including libraries and frameworks, were introduced in the early 1990s [1] and have continued to evolve since then. Gudivada defined libraries and frameworks as follows: *"A software library is a set of functions that application can call, whereas a framework provides higher-level support in the form of some abstract design to speed up applications development"* [2]. Both libraries and frameworks are commonly used in NER applications, depending on the specific requirements and complexity of the task at hand. NER, a sub-discipline of **Natural Language Processing (NLP)**, plays a critical role in the extraction of knowledge from unstructured text, a particularly valuable task in healthcare where **Information Overload (IO)** is a persistent challenge [3, 4, 5, 6]. IO [7] complicates the development of **Clinical Practice Guidelines (CPGs)** [8, 9], as the search for evidence relies on vast amounts of unstructured text, such as clinical reports and findings of medical research [9]. By converting unstructured text into structured data, NER helps manage IO in the medical domain [6]. These data can be used to support **Information Retrieval (IR)** for the search for evidence as the basis for CPGs [10]. The use of CPGs can help reduce the risk for the patient in clinical decision-making [8]. Modern text analysis techniques have led to the development of **Machine Learning (ML)**-based methods for NER [11, 12, 13, 14]. These methods are highly efficient at processing unstructured natural language text [15]. The most commonly used ML methods for NER include supervised, unsupervised, and semi-supervised learning [15]. Supervised learning, which relies on manually annotated data to train models, is currently the most widely used method [16, 17]. Unsupervised learning, on the other hand, uses statistical algorithms to identify patterns in unlabeled data [17]. Semi-supervised learning combines

both approaches and requires only a small amount of annotated data [17]. The growing
interest in ML-based NER has led to a significant increase in research activity and the
number of available ML-based NER tools [18]. Consequently, it can be challenging for users
to keep up with the latest developments and stay informed about the state-of-the-art in
this field.

Having described that NLP, NER, and ML are essential techniques for effective IR in
the medical field, this research is now motivated by related research projects. The **Rec-
ommendation Rationalisation (RecomRatio)** project, launched by Bielefeld Univer-
sity in 2018 to develop a computational method to rationalize recommendations, uses
the medical literature to extract arguments for or against a particular medical treatment
[19, 20]. Arguments are made available in a knowledge base to support medical decisions.
CPGs can help reduce patient risk in medical decision making. However, domain experts
face the challenge of IO when developing CPGs [9] since they need to use large amounts
of unstructured text, such as clinical reports or medical research findings, as sources of
evidence [8]. To make the knowledge within these documents accessible, an automated
analysis and visualization of specific NLP features in natural language texts are essential,
such as NER, Entity Linking, or Relation Extraction [21]. The **Artificial Intelligence
for Hospitals, Healthcare & Humanity (AI4H3)** project builds upon the results of
RecomRatio and aims to support the transparency and explainability of medical decisions
using **Artificial Intelligence (AI)** [22]. For this purpose, AI4H3 proposes a layered
architecture with a central hub called "KlinGard Smart Medical Knowledge Harvesting
Hub". This hub serves as a central point for registering AI modules that can be used for
natural language text analysis. In addition to integrating various technologies, this hub
architecture allows heterogeneous data and AI modules to be integrated in a decentralized
manner. The **Cloud-based Information Extraction (CIE)** project deals with the hub
architecture of AI4H3 [23]. This concerns, among other things, the provision of cloud-
based resources (such as computing power and storage) for the automatic extraction of
natural-language texts using ML techniques [23]. These resources are intended to enable
end-to-end NER pipeline support in a cloud environment. To successfully use ML-based
NER in knowledge domains, such as medicine, domain-specific knowledge is necessary for
developing and training ML models. ML-based NER could be used more widely for in-
formation extraction if domain experts could train and use NER systems independently.
Framework-Independent Toolkit for Named Entity Recognition (FIT4NER),
also a project in the AI4H3 environment, aims to enable medical experts to use various
AI-based text analysis techniques [24]. For domain experts, the dynamic nature of NER
research presents several challenges. Firstly, NER users need to compare various tools
helping them to identify NLP features, such as like **Named Entities (NEs)** and Entity
Relations, before deciding which solution is best suited for their tasks. However, comparing
different solutions can be challenging because *"it remains difficult for NLP practitioners
to clearly and objectively identify what software perform(s) the best"* [25], along with deter-
mining which tools efficiently extract, analyze, and visualize NLP features for effective IR
to support CPG development. Existing studies have used different datasets and presented
their results heterogeneously, making it difficult to compare between them [18]. Secondly,
NER users face the challenge of selecting an appropriate tool for their specific task based
on the comparisons aforementioned. This step *"is critical in developing an NLP-based ap-
plication as it affects the accuracy of analysis tasks"* [26]. Third, domain users often lack
the computational and storage resources necessary to train high-quality NER models in
their knowledge domain. Although cloud computing could potentially address this issue,
domain experts often lack the knowledge and experience necessary to effectively leverage

this technology [27, 28]. The primary objective of this survey is to explore the challenges faced by ML experts when **comparing** and **selecting** NER tools for their projects. The findings from this research will be used to help non-experts make informed decisions when comparing and selecting NER tools. To achieve this goal, the following **Research Questions (RQs)** have been defined and will be addressed in this work: *(RQ1) How do ML experts evaluate NER tools, and which criteria are most important to them? (RQ2) What primary challenges do ML experts face when selecting a suitable NER tool?*

This study is structured using the Nunamaker research method [29], which consists of four phases: observation, theory building, system development, and experimentation. Chapter 2 belongs to the observation phase and focuses on analyzing the current state of the art and related work to this research. Chapter 3 is dedicated to survey modeling as part of the theory building phase and involves developing the survey questionnaire, which belongs to the system development phase. In Chapter 4, the results of experiments with ML experts are described as part of the experimentation phase. Finally, Chapter 6 summarizes the study findings.

2   State of the Art in Science and Technology

This chapter focuses on the observation phase and introduces the background of this work and related research activities. The objective is to identify and discuss **Remaining Challenges (RCs)** in the areas addressed in this article. First, NER and the challenges of dealing with various NER tools are described. Second, a short introduction to cloud technologies is given. Finally, related studies are presented and discussed.

NER is an NLP technique that aims to extract NEs from unstructured text documents [30]. A NE is a word or phrase that refers to a specific entity such as a person, place, or organization. NER is a crucial technique used in various applications, including IR [31], question answering systems [32], machine translation [33], and social media analysis [34]. In the medical domain, NER plays a pivotal role in **Clinical Decision Support Systems (CDSSs)** and enables clinical information mining from **Electronic Health Records (EHRs)** [35]. In recent years, NER has seen significant progress due to the development of new techniques and models, including deep learning [16]. These advancements have led to substantial improvements in the performance of NER systems, making NER one of the most extensively researched NLP tasks today [16]. In NER, there are different techniques available, including traditional, ML-based, and hybrid approaches [15]. Traditional NER approaches rely on methods that use manually created rules or are dictionary-based. Although these systems are often efficient and accurate, they are also limited by fixed rules or dictionaries and do not generalize well across different domains and languages [15]. ML-based approaches to NER have gained popularity in recent years, mainly due to the availability of large annotated datasets and advancements in deep learning techniques [16]. These approaches are capable of efficiently processing unstructured and large datasets and achieve superior results. Instead of relying on fixed rules or dictionaries, ML-based NER uses statistical models that learn to detect NEs from annotated data through a process of training and testing. ML techniques are divided into supervised, unsupervised, and semi-supervised learning [17]. Supervised learning [17] relies on manually annotated data to train a model, where the model learns to predict the labels of unseen data. Unsupervised learning [17], on the other hand, relies only on statistical algorithms to detect patterns from unlabeled data. Semi-supervised learning [17] combines these two approaches by training a model with a small set of annotated data and using it to label a larger set of unlabeled data, thus improving the accuracy of the model. In recent years,

pre-training large language models such as BERT [36], GPT-2 [37], and RoBERTA [38] on large corpora have shown remarkable improvements in NER performance. These models are capable of achieving state-of-the-art performance on NER tasks and can efficiently fine-tune on smaller datasets for domain-specific tasks. Although further improvements can be made, AI advancements have already made significant progress in addressing complex NER challenges. The research field of NER continues to evolve rapidly, with new and innovative tools being developed to address different challenges and use cases. Therefore, it is crucial for ML experts to compare and evaluate the performance of available NER tools and to select the one that best fits their specific task, such as training and fine-tuning ML models on custom datasets. This study aims to gain insight into how such comparisons are conducted in practice and identify the challenges and factors that influence the decision-making process of ML experts (RC1).

Amazon Web Service (AWS) launched in the early 2000s, pioneering the concept of cloud computing by offering scalable computing resources on demand as a service [39]. This groundbreaking technology has since evolved into a widely available solution that offers vast amounts of computing resources at any given time. The availability of scalable and cost-effective cloud computing has revolutionized the field of AI by providing a scalable and cost-effective platform for creating, training, and deploying AI models. ML-based NER is one of the many AI applications that have benefited from the cloud's capabilities [23]. The unprecedented growth of data has made it challenging to manage and analyze large amounts of information using local compute resources [16]. To tackle this issue, leading providers such as AWS, Microsoft Azure, and Google Cloud Platform offer cloud-based platforms at various levels of abstraction, including **Infrastructure as a Service (IaaS)**, **Platform as a Service (PaaS)**, and **Software as a Service (SaaS)** [39]. These platforms provide the necessary computing resources and tools to store, process, and analyze massive amounts of data efficiently and cost-effectively. Cloud-based ML platforms not only provide computing power (IaaS), but also offer a comprehensive suite of tools and services for data processing, model training, and deployment (PaaS). These platforms make it easy to scale performance up or down as needed, even for demanding applications with real-time requirements. Cloud providers offer not only cloud-based ML platforms but also NLP and NER services. These services include pre-built models and **Application Programming Interfaces (APIs)** that enable users to easily incorporate AI functionality into their applications without requiring extensive expertise in the AI domain [23]. To leverage cloud-based resources effectively, ML experts must carefully evaluate which level of abstraction and which cloud-based services from which provider to use. Furthermore, utilizing cloud-based resources requires familiarity with the relevant technologies, including understanding their strengths, limitations, and best practices [40]. Although cloud technology offers many benefits, including scalability and cost effectiveness, there are legitimate concerns about privacy, security, and ethical implications, particularly in the medical field [41]. As a result, ML experts must carefully consider these factors and evaluate whether cloud-based resources can be used while still meeting regulatory and ethical standards. In summary, cloud technology has rapidly evolved into a powerful platform for creating, training, and deploying AI models. However, ML experts face the challenge of determining whether and which cloud-based resources to deploy, requiring careful evaluation of factors such as scalability, cost, security, privacy, and ethical implications. This study aims to conduct a survey of ML professionals to uncover the key factors they consider when deciding whether to use cloud-based resources (RC2).

In recent years, several scientific papers have compared and evaluated NER tools for various application domains, such as formal and social media texts [42], software documen-

tation [43], historical texts [44], news sources [26, 25], and specific languages [45]. Pinto et al. [42] conducted a study to compare and analyze the performance of multiple NLP tools, including their effectiveness on formal and social media texts in four commonly used NLP tasks, which include NER. Their findings suggest that it is a challenge *"to select which one to use, out of the range of available tools"*, and *"this choice may depend on several aspects, including the kind and source of text"* [42]. Al Omran et al. [43] conducted a comprehensive systematic review and experiments in 2017 to analyze the appropriate selection of an NLP library for the analysis of software documentation. Their study focused on tokenization and part-of-speech tagging, which are essential tasks in the process of performing NER. Their findings underscored the criticality of selecting the right library, yet revealed that a small proportion of papers in the literature provide justification for their NLP library choices. Based on their results, the authors strongly recommend that researchers carefully consider their options when comparing and selecting NLP libraries and make informed decisions. During a comparison of NER tools for use with historical texts in 2018, Won et al. [44] discovered that the *"individual performance of each NER system was corpus dependent"*. By combining various tools, they were able to achieve superior results without the need to translate historical texts into modern English. In 2019, Weiying et al. [26] conducted a benchmarking study of NLP tools for enterprise applications, which included standard NLP tasks such as NER. Their research highlights the importance of carefully selecting an appropriate NLP library as a crucial step in the development of NLP applications. Schmitt et al. [25] identified the challenges associated with selecting a NER tools. They found that objectively comparing NER tools is challenging due to the lack of replicable existing comparisons, and that research surveying NER tools users about the difficulties of selecting and comparing these tools is rare. Aldumaykhi et al. [45] recently conducted a comparative study of three NER tools for analyzing Arabic texts. Through experimentation, they also found that combining these tools resulted in improved performance. Jehangir et al. [30] examined the most relevant datasets, tools, and deep learning approaches currently used for NER problem solving. Among other things, they discussed five different available tools utilized for NER, such as spaCy, NLTK, and OpenNLP [30]. They found that each model or approach has its methodological advantages and disadvantages. For instance, Deep Learning offers benefits in terms of feature engineering and implementation complexity, while rule-based methods require significant manual effort for rule generation and are complex to implement [30]. Additionally, they noted that combining various models or approaches could potentially yield superior results [30]. This underscores the necessity of comparing NER approaches and selecting the most suitable technology for each specific application. Drawing on the academic papers presented, it is clear that it is essential to conduct a thorough comparison of NER tools and select the optimal tool to meet specific requirements. Apart from research studies that compare NER tools, there is a noticeable lack of work that directly explores the perspectives of ML experts on the strategies and challenges involved in selecting and comparing NER tools for specific use cases. Amershi et al. [46] conducted a noteworthy survey at Microsoft, where they collected feedback from over 500 software engineers working on AI and ML. The survey's primary finding was that automation is crucial to facilitate efficient data aggregation, feature extraction, and label synthesis, thus accelerating the pace of experimentation. Secondly, the survey found that *"it is necessary to blend data management tools with their ML frameworks to avoid the fragmentation of data and model management activities"* [46]. Finally, the survey [46] highlighted the importance of training and education for users with limited AI experience, implying that a system that supports users in utilizing AI could potentially reduce the need for such training. The absence of research investigating the challenges faced by

ML experts in comparing NER tools to select the appropriate solution for their project motivates the present study (RC3).

The authors identified three key RCs for conducting a survey among ML experts regarding the challenges in comparing and selecting NER tools. The first RC addresses how ML experts approach the decision-making process to compare NER tools and choose the most appropriate ones. The second RC focuses on the use of cloud resources for ML model training for NER. What are the key factors considered when deciding whether or which cloud-based resources to use? The third RC highlights the lack of current research on the challenges that ML experts face when comparing and selecting suitable NER tools for their projects. After discussing the current state of the art in science and technology, identifying RCs in the areas of NER and cloud, and reviewing related studies, the following chapter describes the modeling and design of the survey.

3   Expert Survey Modeling

This chapter focuses on the theory-building phase by addressing the RCs presented in Chapter 2, grounded in the current state of the art in science and technology. To address the RQs defined in Chapter 1, it is necessary to systematically collect and assess knowledge from experienced ML experts in the field of NER. Expert surveys are a widely used and effective method for obtaining opinions and insights from experts in a particular field [47]. By gathering input from these experts, a deeper understanding of the topic can be gained and inform this research accordingly. This work aims to address the defined RQs by surveying ML experts in the field of NER using the Kasunic model, an established framework for expert surveys [48]. The model outlines a comprehensive process consisting of seven stages to be followed when conducting a survey (see Figure 1). The stages outlined in the Kasunic model provide a structured approach to conducting expert surveys effectively, thereby ensuring that the resulting data is accurate and informative. This chapter will describe how to apply the guidelines and defined stages of the Kasunic model to survey ML experts in the field of NER. By following these guidelines and stages, meaningful experts' insights can be gathered and the RQs can be addressed more effectively.
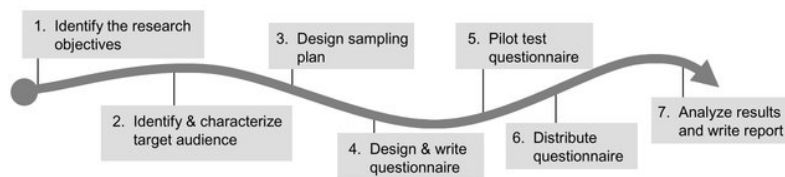

Fig. 1. Survey Research Process by Kasunic [48]

Objectives. The first stage of the Kasunic model is to establish the **Research Objectives (ROs)** for the survey. As Chapter 1 has already defined the RQs for this work, they can be used as a basis for establishing the ROs. The following objectives have been derived: *(RO1) Identify the criteria that ML experts use to evaluate NER tools and determine the importance of each criterion. (RO2) Investigate the primary challenges that ML experts encounter when selecting a suitable NER tool.*

Target Audience. Identifying and characterizing the target audience is the next stage in the survey process. The survey aims to gather insights from ML experts with demonstrated experience in NER. It is assumed that suitable experts should have more than

three years of experience in ML, NLP, and NER, and individuals who have written a Ph.D. thesis in any of these areas will also be considered.

Sampling Plan. In the third stage, the focus is on defining the sampling plan. This involves determining whether it is necessary to include a representative cross section of the target group in the survey and, if so, outlining how this can be accomplished. The objective of this study is to gather information and challenges from experts using NER, without the need to generalize the findings. Therefore, there is no requirement to cover a representative cross section of the target population or to develop a detailed sampling plan. Participants were invited to participate on a voluntary basis in the survey. Initially, the survey included questions to collect information about the background and experiences of the participants. This approach will help ensure that the responses received come from individuals who meet the criteria to be considered expert in the field of NER.

Questionnaire. Following the design of the sampling plan, the next step was to create the questionnaire. It is crucial to carefully formulate the questionnaire items in a way that translates the ROs, as this facilitates the analysis and interpretation of the survey data. The questionnaire is structured into three sections: *1. General Information*, *2. Experience in selecting and using NER Tools* and *3. Final Questions*. Section 1 is designed to gather essential demographic and background information about participants. It includes details such as their academic background, age, general experience with NER, their specific roles in NER projects, and the knowledge domains where NER techniques have been applied. At the end of the section, participants were asked about their experience with various NER tools and frameworks. The options included locally installable NER tools and frameworks such as spaCy, Stanford Named Entity Recognizer, Hugging Face Transformers, Natural Language Toolkit, Flair, AllenNLP, OpenNLP, and GATE. Furthermore, frequently used cloud-based services such as IBM Watson Natural Language Understanding, Amazon Comprehend, Google Cloud Natural Language API, Microsoft Azure Cognitive Services, and OpenAI GPT-4 were considered to address the requirements of RC2. To compile the list of NER tools and frameworks, a comprehensive review was carried out to identify commonly used NER solutions. In addition, participants could specify an additional tool in a free text field. Based on this input, further questions were posed in Section 2 regarding each selected tool or framework. These questions address the RCs 1-3 and ROs 1-2, including the comparison and evaluation of ML-based NER tools and challenges in tool selection. The questions for each framework in Section 2 are listed in Table 1. Finally, Section 3 invites participants to rate their survey experience and includes an open-ended question, providing an opportunity for them to share additional thoughts or insights. The complete questionnaire is available online[1].

Pilot Test Questionnaire. To eliminate errors and improve the questionnaire, it is important to test it with members of the target group. The test runs were conducted with a group of three ML experts who have at least five years of experience in ML, NLP, and NER. The survey was revised and improved based on the errors and problems identified during the testing process, which included the following: In an early draft of the questionnaire, respondents were asked which NER tools they had experience with and which key factors were important to them when selecting NER tools. During pretests, this approach was found to not allow for a specific identification of key factors that were important for the selection of a particular NER tool. This is especially true when respondents had experience with multiple NER tools. For example, performance might be a crucial factor when selecting a cloud-based tool, whereas privacy concerns might lead to the selection of locally installed tools. As a result, the questionnaire was adjusted so that respondents could select

---

[1] `https://umfrage.fernuni-hagen.de/v3/671211?lang=en`

Table 1. Questions per selected Tool

Question	Answer Options	Selection Type
Please indicate your level of experience using $< selectedTool >$:	1 (Very Poor), 2 (Poor), 3 (Average), 4 (Good), 5 (Excellent)	Single
How important were the following criteria in your evaluation of $< selectedTool >$ compared to other existing NER frameworks or tools, such $< examples >$?	Performance; Customization; Integration; Documentation and support; Licensing and cost; Accessibility; User interface and ease of use; Knowledge Domain Requirements; Privacy	Matrix, 5-point scale per option: 1 (Not Important), 2 (Slightly Important), 3 (Moderately Important), 4 (Important), 5 (Very Important)
In addition to the previously mentioned factors, were there any other criteria you considered important in your evaluation of $< selectedTool >$ compared to other existing NER frameworks or tools, such as $< examples >$?	$< openText >$	Text
How important were the other criteria in your evaluation of $< selectedTool >$ compared to other existing NER frameworks or tools, such as $< examples >$?	1 (Not Important), 2 (Slightly Important), 3 (Moderately Important), 4 (Important), 5 (Very Important)	Single
How hindering have the following challenges or limitations with $< selectedTool >$ been in the past?	Time and effort to learn the new framework; Lack of documentation; Challenges with integration into existing applications; Performance issues; Cost; Lack of support, such as documentation, community resources and paid support options	Matrix, 5-point scale per option 1 (Not Hindering), 2 (Slightly Hindering), 3 (Moderately Hindering), 4 (Hindering), 5 (Very Hindering)
Have you encountered any other challenges or limitations with $< selectedTool >$ in the past? If yes, please describe them briefly.	$< openText >$	Text
How hindering have the other challenges or limitations with $< selectedTool >$ been in the past?	1 (Not Hindering), 2 (Slightly Hindering), 3 (Moderately Hindering), 4 (Hindering), 5 (Very Hindering)	Single

from a list of NER tools they had previously used. Then, specific questions were posed for each of the selected tools. However, further tests revealed that this approach increased the number of questions that needed to be answered in proportion to the number of NER tools selected, making the questionnaire too lengthy overall. Consequently, the questions asked for each NER tool were critically reviewed and reduced from 13 to 7. Furthermore, it was found that the clarity of some questions could be improved by providing examples, which was subsequently implemented.

Questionnaire Distribution. After completing quality assurance, the next step is to distribute the questionnaire to the appropriate target group. Chapter 4 provides a comprehensive description of the survey implementation.

Analysis. Finally, the collected results were analyzed and presented using appropriate graphical representations to facilitate the understanding of the findings. The detailed report containing these diagrams is provided in the following chapter 5.

This chapter presented the modeling of the expert survey using a survey based on Kasunic's model [48], encompassing all seven stages of the framework. The next chapter describes how the survey was conducted.

4   Implementation

This chapter addresses the experimentation phase and provides a detailed description of how the survey was conducted, based on the modeling developed in the previous chapter. The finalized questionnaire was distributed to selected members of the target group by email on August 19, 2024. Invitations were sent to a variety of individuals, including authors of relevant research papers, as well as employees of universities, research institutes, and industrial companies working in the field of ML and NLP. A total of 27 invitations were sent to individuals. In addition, an invitation was sent to an email distribution list of the Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V., which reaches approximately 60 researchers in the NLP field. The survey was open for participation from August 18, 2024, to October 13, 2024, and received 23 responses, resulting in a response rate of 26%. This response rate is better than that of other expert surveys such as [47] (8%), [49] (11%) or [50] (13%). After describing the conduct of the survey, the next chapter presents a summary and interpretation of the results of the expert survey.

5   Evaluation

In the previous chapter, the conduct of the expert survey was explained. This chapter focuses on the experimentation phase and presents a summary and interpretation of the results of the expert survey. It begins with a presentation of the background and demographics of the survey participants. The subsequent sections investigate the results of the expert survey based on the defined ROs 1-2. The sections cover a comparative analysis and evaluation of ML-based NER tools, and challenges encountered in the selection of tools. In addition, potential threats to the validity of the results are discussed.
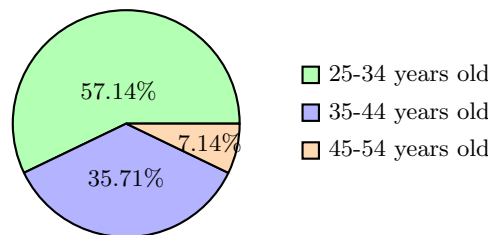
5.1   Demographics

Fig. 2. Age Distribution

First, the demographic structure of the survey participants is analyzed. The participant group consisted predominantly of academics aged 25 to 34 years (57.14%) and 35 to 44 years (35.71%) (see Figure 2). A smaller proportion of the participants was between 45 and 54 years old. As shown in Figure 3, most of the respondents had a doctorate (42. 86%), while more than a third (35. 71%) had a master's degree, and 21. 43% had a bachelor's degree. This suggests that the survey was conducted with a highly qualified participant group. Most of the participants (85. 71%) obtained their academic degrees in the field of Computer Science (Figure 4). However, there were also participants from other disciplines, such as History and Economics. The aim of this work is to gain insight to support domain
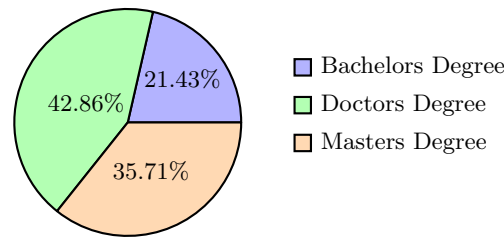
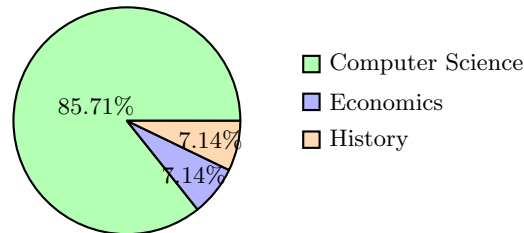**Fig. 3.** Distribution of Academic Degrees

Fig. 4. Field of Study Distribution

experts. Nevertheless, due to the predominant Computer Science backgrounds of the participants, caution is warranted. The results of this survey cannot be directly generalized to domain experts and must be interpreted and evaluated accordingly.
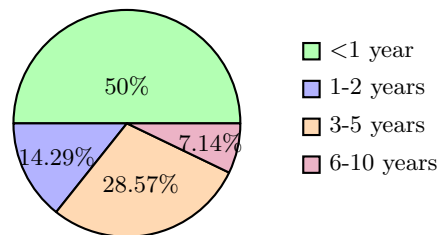

Fig. 5. Distribution of NER Experience by Years

Now, the experiences of the participants in the field of NER was examined. Figure 5 illustrates that 50% of the participants have been involved with NER for less than a year, while 14.20% have been involved for up to two years. In contrast, 28.57% of the participants have accumulated over three years of experience in this area. Only 35.71% of the respondents reported having minimal experience with NER (Figure 6). The majority rate their NER experience as average (42.86%) or good (21.43%). In summary, given the high level of education of the participants, this group can be considered experienced in NER, even though many have only recently entered this field.

Although most of the participants come from the field of Computer Science, they apply NER in their projects across a variety of roles and domains. The question *"In your current or most recent Named Entity Recognition project, what was your primary responsibility or role?"* allowed multiple selections. This led to a diverse representation of roles in which participants were involved in their NER projects, such as Software Developer (32%), Data Scientist (24%) or Machine Learning Engineer (16%), as well as Domain Expert or Project Manager (each 4%) (Figure 7). As indicated in Figure 8 for the question *"In which domains have you previously applied Named Entity Recognition techniques?"*, NER was primarily
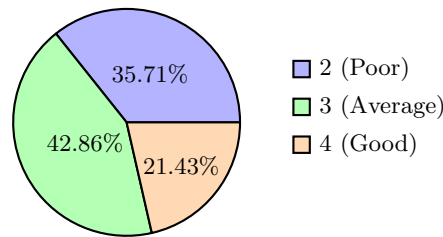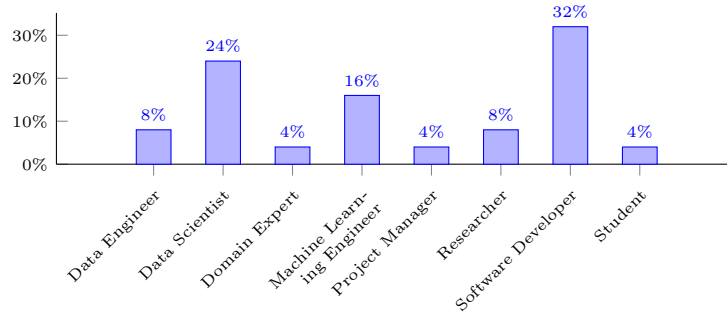
Fig. 6. Distribution of NER Experience Level

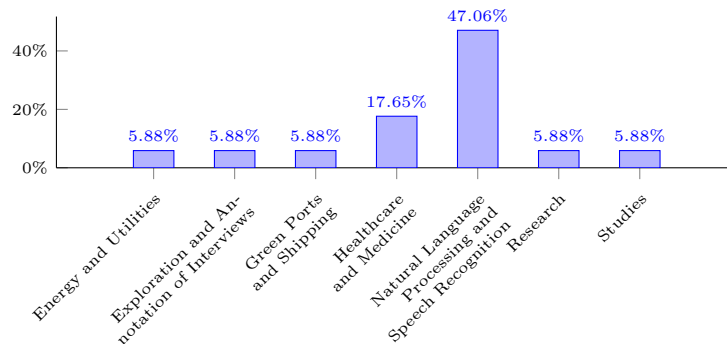
Fig. 7. Distribution of Roles in NER Projects

Fig. 8. Distribution of NER Project Domains

utilized in the field of NLP and speech recognition (47.06%). In addition, a manifold picture emerges: NER was also applied in domains such as Healthcare and Medicine (17.65%), as well as Green Ports and Shipping and E-Commerce and Retail (each 5.88%). Again, multiple selections were allowed.

5.2 Experience with Selecting and using NER Tools

After initially analyzing the demographic structure of the participants, the second step provides a detailed examination of the NER frameworks employed. For this purpose, participants were asked which NER tools and frameworks they have experience with, allowing multiple selections. The results are presented in Figure 9. The most frequently mentioned NER tools and frameworks were spaCy (25%), Hugging Face Transformers (18.18%), OpenAI GPT-4 (15.91%), Natural Language Toolkit (15.91%), and Stanford Named Entity Recognizer (13.64%). All other systems remained below 5% or were not mentioned at all.
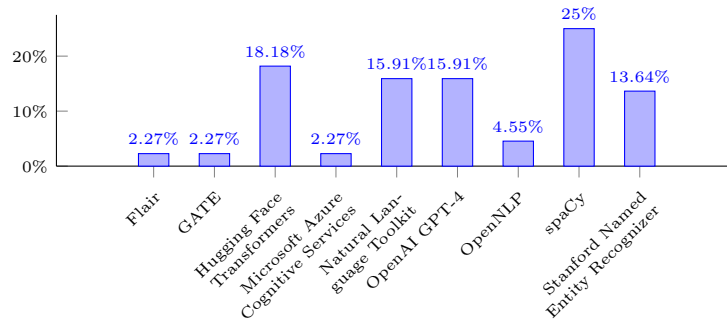
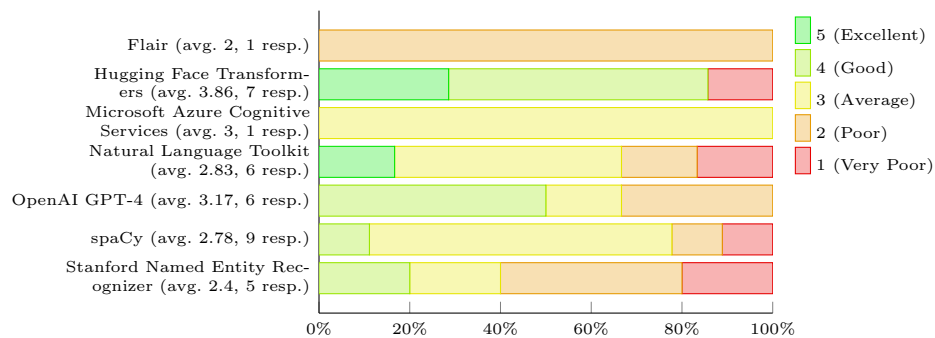**Fig. 9.** Distribution of Experience with NER Framework

Fig. 10. Distribution of Experience Levels per NER Framework

The levels of experience of participants with various NER frameworks (Figure 10) reveal that Hugging Face Transformers (average 3.86 from 7 responses) and OpenAI GPT-4 (average 3.17 from 6 responses) achieved the highest ratings. These tools were rated favorably for ease of use and effectiveness. The Hugging Face Transformers library, widely regarded for its intuitive application of **Large Language Models (LLMs)** [51], appears to facilitate rapid learning and adoption, making it a preferred choice for both novices and experienced developers alike. OpenAI GPT-4, while designed primarily for text generation, is highly adaptable for NER tasks due to its user-friendly natural language prompting mechanism [52]. Other frameworks, such as Microsoft Azure Cognitive Services (average 3.1, 1 response) and Natural Language Toolkit (average 2.83, 6 responses), showed more modest ratings. Flair, spaCy, and Stanford Named Entity Recognizer scored lower in terms of perceived usability (averages ranging between 2.4 and 2.78), likely reflecting their steeper learning curves or limitations in broader applicability.

Figure 11 details the evaluations of participants of various selection criteria across all NER tools and frameworks, revealing performance (average 4.57) as the most critical factor. The consistently high prioritization of performance underscores the need for NER tools to deliver accurate and reliable results in different operational contexts. As described in Chapter 3, questions were also posed regarding cloud-based tools, such as Microsoft Azure Cognitive Services. Especially, for these cloud-based tools, licensing and cost (average 3.89), as well as user interface and ease of use (average 3.59), also rank highly. This reflects the growing importance of affordability and accessibility in encouraging adoption among diverse user groups, such as newbies and ML experts. In contrast, factors such as integration (average 2.95) and customization (average 3.27) were moderately important, suggesting that participants found these areas less critical when evaluating NER tools.
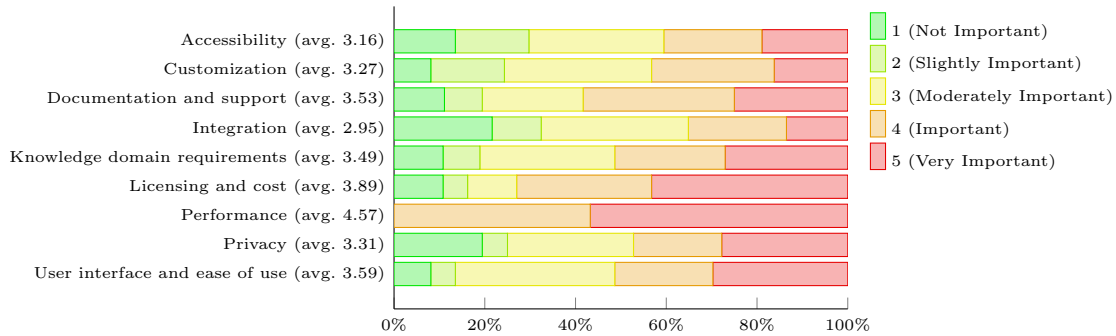
Fig. 11. Priority Distribution per Selection Criteria

Privacy (average 3.31) received mixed scores, signaling varying levels of concern depending on the application and deployment model. Participants also had the opportunity to specify and evaluate additional criteria in a free text field. For spaCy "GPU compatibility for Transformers" was mentioned and rated as "3 (Moderately Important)". For OpenAI GPT-4, "Response Time" was noted and rated as "5 (Very Important)", which can be classified as performance. In general, it can be concluded that all the criteria were considered relevant. This suggests that the importance of criteria is highly dependent on the specific project and that there are no criteria that can be universally deemed unimportant. This is also illustrated by the fact that there are different specific challenges for NER, depending on which knowledge domain NER is to be applied [53]. However, the performance criterion remains significant in all cases. The significance of performance in NER tools and frameworks is underscored by numerous studies that compare their effectiveness [25, 18, 26]. When analyzing the results of cloud-based and locally installable NER tools and frameworks, as outlined in Chapter 3, and calculating the average performance values, notable insights emerge. Table 2 illustrates the average results for both types of tools, highlighting the differences (Delta) between them. For locally installable systems, documentation and support play a critical role, as evidenced by a Delta of -1.14. This observation aligns with findings from related studies, where documentation is frequently emphasized as a key factor in evaluating NER tools. For instance, Schmitt et al. argue that criteria such as documentation should be carefully assessed before selecting and deploying an NER solution [25]. However, for cloud-based systems, the user interface and ease of use are particularly relevant (Delta 1.09). Tamla et al. have already pointed out that managing cloud-based resources is a challenge for newbies and ML experts [23], thus increasing the relevance of the user interface and usability for the use of cloud-based resources. Kurdi et al. also recognized that a good user interface is important for the usability of cloud-based services [54]. Thus, it is clear that users of cloud-based NER services could benefit from systems that simplify their use.

The responses to the question *"How hindering have the following challenges or limitations with < selectedTool > been in the past?"* for all NER tools and frameworks are presented in Figure 12. The responses are surprising in that very few challenges were classified by participants as very hindering. The most frequently cited issue was "Time and effort to learn the new framework" (average 2.84). This was followed by "Performance Issues" (average 2.57), which is consistent with the findings in Figure 11. "Challenges with integration into existing applications" were mentioned the least as a hindrance (average 2.05). Other challenges, such as "Cost" (average 2.36), "Lack of support" (average 2.32), and "Lack of documentation" (average 2.30), were ranked mid-range but low. Each chal-

Table 2. Comparison of Average Priority per Selection Criteria

Criteria	Cloud	Local	Delta
Accessibility	3.57	3.0	0.57
Customization	3.43	3.31	0.12
Documentation and support	2.57	3.71	-1.14
Integration	2.57	3.1	-0.53
Knowledge Domain Requirements	3.43	3.45	-0.02
Licensing and cost	3.71	3.9	-0.19
Performance	4.57	4.55	0.02
Privacy	3.14	3.29	-0.15
User interface and ease of use	4.43	3.34	1.09
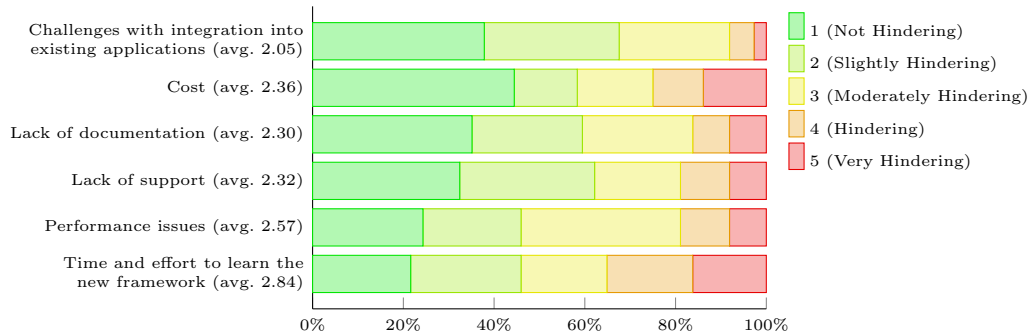
Fig. 12. Priority Distribution per Hindrance Criteria

lenge was mentioned at least once as hindering or very hindering, supporting the assertion that the requirements for NER tools and frameworks are project specific. For spaCy, it was also noted that there are difficulties with supported file formats: *"limited compatibility with common annotation formats; must be converted into spaCy's own format"*. This was rated as "3 (Moderately Hindering". For OpenAI GPT-4, the following challenges were mentioned in the free text field and rated as "4 (Hindering)": *"licensing, privacy considerations, and closed source"*. In general, it can be concluded that reducing the time and effort required to learn new frameworks is essential. Here, also, the results of this question were grouped with respect to cloud-based and locally installable NER tools and frameworks and the average value was calculated. As shown in Table 3, the time and effort required to learn the new framework are particularly restrictive for locally installable systems (Delta -0.57). Therefore, a system that supports the comparison and selection of NER tools and frameworks must also assist users to work quickly and easily with the various systems. For cloud-based services, costs are, as expected, a particularly significant barrier to their adoption (Delta 1.79). Therefore, when using cloud-based resources, it is important to control costs and pay attention to cost efficiency [28].

In addition to the specified NER tools and frameworks, the participants were able to indicate experience with other NER tools or frameworks. One participant noted the use of *"Self-hosted open source LLM (many)"*, suggesting that he downloaded freely available LLMs from the Internet, such as those from Hugging Face[2], for local NER applications. This participant rated their expertise with this technology as "5 (Very High)". When asked, *"How important were the following criteria in your evaluation of your selected NER tool or framework compared to other existing NER frameworks or tools, such as spaCy or Amazon Comprehend?"* the participant rated the criteria: Documentation and support, licensing

---

[2] https://huggingface.co/models

Table 3. Comparison of Average Priority per Hindrance Criteria

Hindrance	Cloud	Local	Delta
Challenges with integration into existing applications	2.29	1.90	0.39
Cost	3.71	1.93	1.79
Lack of documentation	2.43	2.31	0.12
Lack of support	3.00	2.21	0.79
Performance issues	2.29	2.62	-0.33
Time and effort to learn the new framework	2.43	3.00	-0.57

and cost, accessibility, user interface and ease of use, knowledge domain requirements, and privacy as "5 (Very Important)". Customization and integration were rated only "1 (Not Important)". Regarding the challenges or limitations of this technology, the challenges with integration into existing applications and cost were rated "5 (Very Hindering)", while performance issues received a "3 (Moderately Hindering)". Other factors, such as time and effort to learn the new framework, lack of documentation, and lack of support, such as documentation, community resources, and paid support options, were rated "1 (Not Hindering)". This suggests that the respondent has high expectations for an NER tool or framework, leading to the use of local open-source LLMs for NER, while considering him an expert in this area. However, it remains unclear which tools and frameworks were used alongside the local LLMs. It is likely that a system from the specified selection was employed, such as Hugging Face Transformers. The rationale behind the high ratings for integration challenges and costs remains ambiguous. It is possible that the question was misinterpreted and answered in the context of other solutions, given that open-source LLMs typically offer low costs and high flexibility [55]. Should the responses have indeed pertained to open-source LLMs, the effort involved in utilizing local open-source LLMs may have resulted in increased costs and challenges related to integration. Ultimately, a significant finding of this study is that locally operated open-source LLMs represent a relevant technology for NER and are already being used in other projects [55]. Open-source LLMs must be considered in the comparison and selection of suitable NER tools and frameworks.

5.3 Final Questions

Finally, participants were asked to assess their experience with the survey and to provide any additional comments. The survey received overwhelmingly positive feedback, as illustrated in Figure 13. In the final remarks, it was noted that some questions were occasionally too detailed. This feedback can be incorporated into future surveys, although careful consideration must be given to whether such detailed questions are necessary for achieving the survey's objectives.
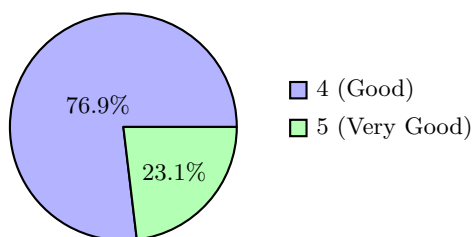
Fig. 13. Distribution of Survey Experience

5.4   Threats to Validity

When conducting this survey, several potential threats to the validity of the results should be considered. First, the 23-part participant count can be deemed too low, which could limit the generalizability of the findings. However, the survey primarily aims to provide information on the relevant criteria to compare and select NER tools and frameworks. This potentially limited generalizability thus poses a minor risk. Second, most of the participants come from the field of computer science. This one-sided composition may lead to biases, as the perspectives and experiences of this group do not necessarily reflect the views of other relevant disciplines. While the responses from computer scientists are significant for domain experts, they should be carefully translated to meet the specific needs of those experts. Future work may, therefore, consider surveying additional domain experts. Third, a participant noted that some questions were too detailed. This may have resulted in some participants not completing the survey in full or having difficulty answering the questions appropriately. However, it was important to ask specific questions to obtain precise answers. In future surveys, the level of detail in the questions should be critically reassessed. The mentioned threats to validity have been adequately considered in interpreting the results. Consequently, the remaining risk to the findings of this survey is assessed as low.

6   Conclusion

In this article, an expert-based survey was presented to gain insights into how NER experts compare and select NER tools and frameworks, as well as the challenges they face. The structured research methodology proposed by Nunamaker was applied. Chapter 1 introduces the research area and motivates the work through related research projects. Chapter 2 analyzes the current state of science and technology, reviews and compares similar works, and presents the RCs. In Chapter 3, the ROs were defined on the basis of Kasunic's Survey Research Process, followed by the design and quality assurance of the survey. The survey publication is described in Chapter 4. Finally, Chapter 5 provides a detailed analysis and interpretation of the results. Various results were obtained related to RO1, which focuses on identifying relevant selection criteria for NER tools and frameworks. The survey highlighted that performance is a particularly important criterion. Furthermore, expert opinions varied significantly. All specified selection criteria were regarded important or very important at least once. This indicates that the relevance of the criteria may vary depending on the project. Therefore, a supportive system should be flexible enough to accommodate various criteria along with performance. In the context of RO2, a distinction is made between cloud-based services and locally installed tools and frameworks. For cloud-based services, cost and user-friendliness are particularly significant. Both aspects represent important requirements for a system that uses cloud-based services for NER. In the case of locally installed systems, the effort required for users to adopt a new system should be minimized, which can be facilitated by an appropriate software solution. Additionally, there were indications of the relevance of locally installed open-source large language models, which should also be integrated into future software systems.

In summary, the defined research objectives have been successfully achieved, and the RCs identified in Chapter 2 have been addressed. The results provide valuable information for future systems designed to support the selection and comparison of NER tools and frameworks. Future work should focus on adapting these results to the needs of domain experts, enabling them to utilize NER in the development of CPGs.

Bibliography

[1] Hamish Cunningham, Valentin Tablan, Angus Roberts, and Kalina Bontcheva. Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLoS computational biology*, 9(2):e1002854, 2013. Publisher: Public Library of Science San Francisco, USA.

[2] Venkat N. Gudivada and Kamyar Arbabifard. Open-Source Libraries, Application Frameworks, and Workflow Systems for NLP. In *Handbook of Statistics*, volume 38, pages 31–50. Elsevier, 2018.

[3] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, September 2010.

[4] Denis Gavrilov, Alexander Gusev, Igor Korsakov, Roman Novitsky, and Larisa Serova. Feature extraction method from electronic health records in Russia. In *Conference of open innovations association, FRUCT*, pages 497–500, 2020. Number: 26 tex.organization: FRUCT Oy.

[5] Ning Liu, Qian Hu, Huayun Xu, Xing Xu, and Mengxin Chen. Med-BERT: A Pre-Training Framework for Medical Records Named Entity Recognition. *IEEE Transactions on Industrial Informatics*, pages 1–1, 2021.

[6] Christian Nawroth, Felix Engel, and Matthias Hemmje. Emerging Named Entity Recognition in a Medical Knowledge Management Ecosystem:. In *KEOD*, pages 29–41, Budapest, Hungary, 2020. SCITEPRESS - Science and Technology Publications.

[7] David Bawden and Lyn Robinson. The dark side of information: overload, anxiety and other paradoxes and pathologies. *Journal of information science*, 35(2):180–191, 2009. Publisher: Sage Publications Sage UK: London, England.

[8] Earl Steinberg, Sheldon Greenfield, Dianne Miller Wolman, Michelle Mancher, Robin Graham, and others. *Clinical practice guidelines we can trust*. national academies press, 2011.

[9] Taher S. Valika, Sarah E. Maurrasse, and Lara Reichert. A Second Pandemic? Perspective on Information Overload in the COVID-19 Era. *Otolaryngology–Head and Neck Surgery*, 163(5):931–933, November 2020.

[10] Florian Freund, Philippe Tamla, and Matthias Hemmje. Towards improving clinical practice guidelines through named entity recognition: Model development and evaluation. In *2023 31st irish conference on artificial intelligence and cognitive science (AICS)*, pages 1–8, 2023.

[11] Hai Leong Chieu and Hwee Tou Ng. Named entity recognition with a maximum entropy approach. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 160–163, Edmonton, Canada, May 2003. Association for Computational Linguistics.

[12] Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. Named entity recognition through classifier combination. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 168–171, Edmonton, Canada, May 2003. Association for Computational Linguistics.

[13] Andrew McCallum and Wei Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 -*,

volume 4, pages 188–191, Edmonton, Canada, 2003. Association for Computational Linguistics.

[14] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv:1802.05365 [cs]*, March 2018. arXiv: 1802.05365.

[15] Ing Michal Konkol. *Named entity recognition*. PhD Thesis, PhD thesis, University of West Bohemia, 2015.

[16] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70, January 2022.

[17] Ariruna Dasgupta and Asoke Nath. Classification of Machine Learning Algorithms. *Figshare*, 2016.

[18] Arlene Casey, Emma Davidson, Michael Poon, Hang Dong, Daniel Duma, Andreas Grivas, Claire Grover, Víctor Suárez-Paniagua, Richard Tobin, William Whiteley, Honghan Wu, and Beatrice Alex. A systematic review of natural language processing applied to radiology reports. *BMC Medical Informatics and Decision Making*, 21(1):179, December 2021.

[19] Bielefeld University. RATIO: Rationalizing Recommendations (RecomRatio), 2017.

[20] Christian Nawroth. *Supporting Information Retrieval of Emerging Knowledge and Argumentation*. PhD thesis, FernUniversität in Hagen, Hagen, November 2020.

[21] Irene Li, Jessica Pan, Jeremy Goldwasser, Neha Verma, Wai Pan Wong, Muhammed Yavuz Nuzumlalı, Benjamin Rosand, Yixin Li, Matthew Zhang, David Chang, R. Andrew Taylor, Harlan M. Krumholz, and Dragomir Radev. Neural Natural Language Processing for Unstructured Data in Electronic Health Records: a Review, July 2021. arXiv:2107.02975 [cs].

[22] FTK. Artificial Intelligence for Hospitals, Healthcare & Humanity (AI4H3). R&D White Paper, FTK e.V. Research Institute for Telecommunications and Cooperation, Dortmund, Germany, April 2020.

[23] Philippe Tamla, Benedict Hartmann, Nhan Nguyen, Calvin Kramer, Florian Freund, and Matthias Hemmje. Cie: A cloud-based information extraction system for named entity recognition in aws, azure, and medical domain. In Frans Coenen, Ana Fred, David Aveiro, Jan Dietz, Jorge Bernardino, Elio Masciari, and Joaquim Filipe, editors, *Knowledge Discovery, Knowledge Engineering and Knowledge Management*, pages 127–148, Cham, 2023. Springer Nature Switzerland.

[24] Florian Freund, Philippe Tamla, Thoralf Reis, Matthias Hemmje, and Paul Mc Kevitt. FIT4NER - Towards a Framework-Independent Toolkit for Named Entity Recognition. 8th Collaborative European Research Conference (CERC 2023), Barcelona, Spain, June 9-10, 2023, 2023.

[25] Xavier Schmitt, Sylvain Kubler, Jeremy Robert, Mike Papadakis, and Yves LeTraon. A Replicable Comparison Study of NER Software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 338–343, Granada, Spain, October 2019. IEEE.

[26] Kok Weiying, Duc Nghia Pham, Yasaman Eftekharypour, and Ang Jia Pheng. Benchmarking NLP Toolkits for Enterprise Application. In Abhaya C. Nayak and Alok Sharma, editors, *PRICAI 2019: Trends in Artificial Intelligence*, Lecture Notes in Computer Science, pages 289–294, Cham, 2019. Springer International Publishing.

[27] Mladen A Vouk. Cloud Computing – Issues, Research and Implementations. *Cloud Computing*, page 12, 2008.

[28] Ryan Chard, Kyle Chard, Kris Bubendorfer, Lukasz Lacinski, Ravi Madduri, and Ian Foster. Cost-Aware Elastic Cloud Provisioning for Scientific Workloads. In *2015 IEEE 8th International Conference on Cloud Computing*, pages 971–974, New York City, NY, USA, June 2015. IEEE.

[29] Jay F. Nunamaker Jr, Minder Chen, and Titus D. M. Purdin. Systems Development in Information Systems Research. *Journal of Management Information Systems*, 7(3):89–106, December 1990. Publisher: Routledge _eprint: https://doi.org/10.1080/07421222.1990.11517898.

[30] Basra Jehangir, Saravanan Radhakrishnan, and Rahul Agarwal. A survey on Named Entity Recognition — datasets, tools, and methodologies. *Natural Language Processing Journal*, 3:100017, June 2023.

[31] Desislava Petkova and W. Bruce Croft. Proximity-based document representation for named entity retrieval. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 731–740, Lisbon Portugal, November 2007. ACM.

[32] Raju Barskar, Gulfishan Firdose Ahmed, and Nepal Barskar. An Approach for Extracting Exact Answers to Question Answering (QA) System for English Sentences. *Procedia Engineering*, 30:1187–1194, 2012.

[33] Yuval Marton and Imed Zitouni. Transliteration normalization for information extraction and machine translation. *Journal of King Saud University-Computer and Information Sciences*, 26(4):379–387, 2014. Publisher: Elsevier.

[34] Juae Kim, Yejin Kim, and Sangwoo Kang. Weakly labeled data augmentation for social media named entity recognition. *Expert Systems with Applications*, 209:118217, December 2022.

[35] Naveen S Pagad and N Pradeep. Clinical named entity recognition methods: an overview. In *International conference on innovative computing and communications*, pages 151–165, 2022. tex.organization: Springer.

[36] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pretraining of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[37] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and others. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[38] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[39] Sandesh Achar. Cloud-based system design. *International Journal of All Research Education and Scientific Methods (IJARESM)*, 7(8):23–30, 2019.

[40] Won Kim. Cloud Computing: Today and Tomorrow. *CLOUD COMPUTING*, 8(1):8, 2009.

[41] Matthias Blohm, Claudia Dukino, Maximilien Kintz, Monika Kochanowski, Falko Koetter, and Thomas Renner. Towards a Privacy Compliant Cloud Architecture for Natural Language Processing Platforms:. In *Proceedings of the 21st International Conference on Enterprise Information Systems*, pages 454–461, Heraklion, Crete, Greece, 2019. SCITEPRESS - Science and Technology Publications.

[42] Alexandre Pinto, Hugo Gonçalo Oliveira, and Ana Oliveira Alves. Comparing the performance of different NLP toolkits in formal and social media text. In Marjan Mernik, José Paulo Leal, and Hugo Gonçalo Oliveira, editors, *5th symposium on languages, applications and technologies (SLATE'16)*, volume 51 of *OpenAccess series in informat-*

ics (OASIcs), pages 3:1–3:16, Dagstuhl, Germany, 2016. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. ISSN: 2190-6807 tex.urn: urn:nbn:de:0030-drops-60086.

[43] Fouad Nasser A Al Omran and Christoph Treude. Choosing an NLP Library for Analyzing Software Documentation: A Systematic Literature Review and a Series of Experiments. In *2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR)*, pages 187–197, Buenos Aires, Argentina, May 2017. IEEE.

[44] Miguel Won, Patricia Murrieta-Flores, and Bruno Martins. Ensemble Named Entity Recognition (NER): Evaluating NER Tools in the Identification of Place Names in Historical Corpora. *Frontiers in Digital Humanities*, 5:2, March 2018.

[45] Abdullah Aldumaykhi, Saad Otai, and Abdulkareem Alsudais. Comparing open arabic named entity recognition tools, 2022.

[46] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. Software Engineering for Machine Learning: A Case Study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, pages 291–300, Montreal, QC, Canada, May 2019. IEEE.

[47] Florian Auer and Michael Felderer. Important Experimentation Characteristics: An Expert Survey. In *Proceedings of the 15th ACM / IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 1–6, Bari Italy, October 2021. ACM.

[48] Mark Kasunic. *Designing an Effective Survey.* Citeseer, September 2005.

[49] Fabian Scheller, Frauke Wiese, Jann Michael Weinand, Dominik Franjo Dominković, and Russell McKenna. An expert survey to assess the current status and future challenges of energy system analysis. *Smart Energy*, 4:100057, 2021.

[50] Isela Mendoza, Marcos Kalinowski, Uéverton Souza, and Michael Felderer. Relating verification and validation methods to software product quality characteristics: Results of an expert survey. In Dietmar Winkler, Stefan Biffl, and Johannes Bergsmann, editors, *Software quality: The complexity and challenges of software engineering and software quality in the cloud*, pages 33–44, Cham, 2019. Springer International Publishing.

[51] Shashank Mohan Jain. *Introduction to Transformers for NLP: With the Hugging Face Library and Models to Solve Problems.* Apress, Berkeley, CA, 2022.

[52] OpenAI et al. GPT-4 Technical Report, March 2024.

[53] Kalyani Pakhale. Comprehensive Overview of Named Entity Recognition: Models, Domain-Specific Applications and Challenges, September 2023. arXiv:2309.14084 [cs].

[54] Heba A. Kurdi, Safwat Hamad, and Amal Khalifa. Towards a Friendly User Interface on the Cloud. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Alfred Kobsa, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Demetri Terzopoulos, Doug Tygar, Gerhard Weikum, and Aaron Marcus, editors, *Design, User Experience, and Usability. User Experience Design for Diverse Interaction Platforms and Environments*, volume 8518, pages 148–157. Springer International Publishing, Cham, 2014. Series Title: Lecture Notes in Computer Science.

[55] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. FinGPT: Open-Source Financial Large Language Models, June 2023. arXiv:2306.06031 [q-fin].