A Survey of Domain Adaptation and Specialized NLP Applications

www.surveyx.cn

Abstract

Natural Language Processing (NLP) has seen transformative advancements through domain adaptation and specialized applications across various sectors, including healthcare, finance, and law. Domain adaptation addresses challenges posed by out-of-distribution data, enhancing model performance by transferring knowledge across fields. In healthcare, models like BioBERT are pivotal for biomedical text analysis, improving tasks such as Named Entity Recognition and information retrieval. The legal sector benefits from NLP by automating document analysis and supporting legal research, although challenges persist in adapting models to complex legal texts. Privacy-preserving techniques are crucial in these domains, ensuring data confidentiality through methods like differential privacy. Sector-specific language models further tailor NLP applications to unique industry needs, driving innovation in financial sentiment analysis and healthcare diagnostics. Future research opportunities lie in refining domain adaptation techniques, expanding multilingual capabilities, and integrating ethical AI frameworks, particularly in healthcare and legal contexts. By advancing these areas, NLP can continue to enhance industry-specific processes, improve data privacy, and foster more robust and adaptable language models. These developments underscore the critical role of tailored NLP solutions in addressing real-world challenges and driving technological progress across diverse sectors.

1 Introduction

1.1 Scope and Significance of Domain Adaptation

Domain adaptation (DA) is pivotal in natural language processing (NLP) for addressing challenges associated with out-of-distribution examples across various applications [1]. It enhances model adaptability and generalization, particularly in sectors like finance, healthcare, and law. For instance, in the insurance industry, integrating domain-specific knowledge into Large Language Models (LLMs) is crucial for adapting these models to practical business scenarios, thereby improving performance [2].

In healthcare, LLMs analyze Electronic Health Records (EHRs), bridging significant gaps in comprehensive data review and application [3]. Despite the transformative potential of LLMs in Biomedical and Health Informatics (BHI), ethical and practical challenges persist [4]. Additionally, DA is essential for enhancing machine translation systems by addressing language style variations and out-of-vocabulary issues, ultimately improving translation quality.

DA's significance extends to scientific literature, where the vast volume of publications necessitates efficient text analysis, playing a critical role in managing complexity and improving information retrieval [5]. It is particularly vital for misinformation detection during events like the COVID-19 pandemic, adapting models to new domains and overcoming label and conditional shifts [6].

Moreover, DA facilitates essential news element extraction using frameworks like 5W1H, crucial for event extraction and text summarization [7]. In legal NLP, DA aids in analyzing complex documents,

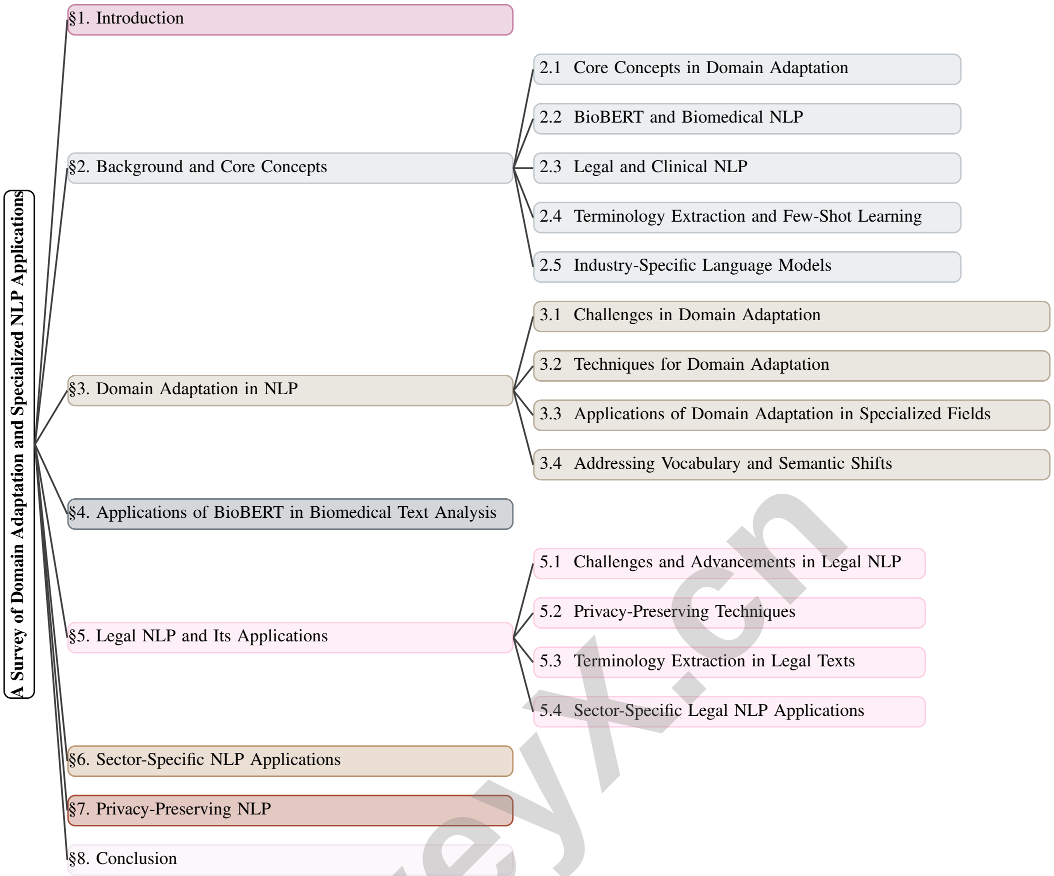

A Survey of Domain Adaptation and Specialized NLP Applications
§1. Introduction
§2. Background and Core Concepts
2.1 Core Concepts in Domain Adaptation
2.2 BioBERT and Biomedical NLP
2.3 Legal and Clinical NLP
2.4 Terminology Extraction and Few-Shot Learning
2.5 Industry-Specific Language Models
§3. Domain Adaptation in NLP
3.1 Challenges in Domain Adaptation
3.2 Techniques for Domain Adaptation
3.3 Applications of Domain Adaptation in Specialized Fields
3.4 Addressing Vocabulary and Semantic Shifts
§4. Applications of BioBERT in Biomedical Text Analysis
§5. Legal NLP and Its Applications
5.1 Challenges and Advancements in Legal NLP
5.2 Privacy-Preserving Techniques
5.3 Terminology Extraction in Legal Texts
5.4 Sector-Specific Legal NLP Applications
§6. Sector-Specific NLP Applications
§7. Privacy-Preserving NLP
§8. Conclusion

Figure 1: chapter structure

such as crypto asset white papers, under regulatory frameworks like the European Union's Markets in Crypto-Assets Regulation (MiCAR) [8].

Unsupervised domain adaptation (UDA) aligns unlabeled target domain data with source domain distributions, effectively mitigating distribution shift issues [9]. This is especially important in scenarios lacking labeled target domain data, enhancing predictive performance across various industries [10].

DA is fundamental in modern NLP, addressing data distribution mismatches that arise when training and test datasets originate from different sources or become outdated. By enabling models trained in one domain to adapt for use in another, DA enhances generalizability and supports the application of NLP solutions to real-world problems, particularly in sensitive domains where data privacy is critical. Recent studies highlight strategies, including unsupervised domain adaptation and innovative learning setups, that bridge the domain gap, leading to more robust and effective NLP applications [11, 1].

1.2 Importance of Specialized NLP Applications

Specialized Natural Language Processing (NLP) applications are vital in enhancing the efficiency and effectiveness of industry-specific processes by customizing language models to meet the unique requirements of different sectors. In finance, the deployment of LLMs has transformed decision-making and operational efficiency through precise processing of extensive financial data [12], essential for interpreting complex financial documents and generating insights for strategic business decisions.

In healthcare, specialized NLP applications bridge the performance gap between domain-specific Small Language Models (SLMs) and LLMs, improving the accuracy and reliability of healthcare data analysis [13]. These applications are crucial for extracting adverse drug reactions from unstructured text through a unified NLP pipeline that incorporates document classification, named entity recognition, and relation extraction [14]. Furthermore, the development of multilingual models aims to serve as low-resource medical assistants, providing essential medical information in indigenous languages and enhancing healthcare accessibility [15].

In the legal sector, the increasing demand for NLP techniques to process and understand legal documents arises from the exponential growth of pending legal cases in populous countries like India [16]. Although the adoption of NLP tools in legal contexts has historically been slow, their potential to address the access to justice crisis by automating and streamlining legal processes is increasingly recognized [17]. Efficient legal NLP models are critical for maintaining performance while minimizing resource consumption, making legal services more accessible and cost-effective [18]. These models assist in regulatory compliance and enhance document clarity, ultimately facilitating better investment decisions [8].

The integration of specialized NLP applications across various industries not only enhances performance and efficiency but also addresses sector-specific challenges. This underscores the indispensable role of tailored NLP solutions in advancing industry-specific processes and ensuring that language models evolve in alignment with sectoral needs. The adoption of novel parallel processing frameworks that leverage distributed computing further enhances the efficiency of machine learning algorithms, highlighting the ongoing evolution and sophistication of NLP methods [19].

1.3 Impact Across Industries

Domain adaptation and specialized NLP applications significantly impact various industries by addressing the unique challenges and requirements inherent to each sector. In finance, deploying LLMs faces distinct challenges, including reliance on professional expertise, managing confidential data, and adhering to strict regulatory compliance [12]. These challenges necessitate developing tailored NLP models capable of efficiently processing complex financial documents, enhancing decision-making and operational efficiency.

In healthcare, LLMs are increasingly applied to EHRs, with applications including named entity recognition, information extraction, text summarization, and medical question-answering. These applications bridge the gap between domain-specific SLMs and LLMs, improving the accuracy and reliability of healthcare data analysis. Additionally, NLP in healthcare extends to dialogue summarization and EHR generation, further enhancing the sector's ability to manage and utilize vast amounts of data effectively [20].

The legal industry also benefits significantly from domain adaptation and specialized NLP applications, particularly in tasks such as legal research, reasoning, contract review, and document analysis [21]. Applying NLP in the legal domain addresses the unique characteristics of legal texts, which are often lengthy, unstructured, and contain specialized lexicons [16]. Furthermore, analyzing crypto asset white papers under regulatory frameworks like the European Union's Markets in Crypto-Assets Regulation (MiCAR) highlights the need for effective NLP methods to navigate the evolving legal landscape [8].

Domain adaptation and specialized NLP applications are crucial in overcoming challenges posed by sensitive domains, underscoring their importance across various industries [1]. By tailoring NLP models to the specific needs of different sectors, these applications enhance industries' capability to manage data, comply with regulations, and improve operational efficiency, thereby driving technological advancement and innovation.

1.4 Structure of the Survey

This survey is organized into several key sections, each focusing on distinct aspects of domain adaptation and specialized NLP applications. The introductory section outlines the scope and significance of domain adaptation, emphasizing its impact across various industries such as finance, healthcare, and law. Following this, the background and core concepts section delves into fundamental ideas, including domain adaptation, BioBERT, legal NLP, clinical text analysis, and privacy-preserving NLP, while highlighting the relevance of terminology extraction and few-shot learning.

The survey transitions into a detailed examination of domain adaptation in NLP, discussing its role in knowledge transfer across fields, addressing challenges, and exploring techniques such as few-shot domain transfer. The applications of BioBERT in biomedical text analysis are comprehensively examined, emphasizing its significant role in enhancing clinical text analysis and improving the understanding of healthcare data. This includes a systematic exploration of its capabilities in tasks such as biomedical named entity recognition, relation extraction, and clinical document classification, where it demonstrates superior performance compared to other models. The analysis also addresses methodologies for fine-tuning BioBERT to meet the specific needs of the healthcare domain, discusses ethical considerations like patient privacy and data security, and highlights its potential to facilitate clinical decision support and efficient information retrieval within the rapidly growing field of biomedical literature [22, 3, 23, 24].

Subsequent sections explore the use of NLP in processing legal documents, discussing challenges, advancements, and privacy-preserving techniques. The survey also investigates sector-specific NLP applications, providing examples from industries like finance and healthcare. The importance of protecting sensitive information is addressed in the privacy-preserving NLP section, which explores various techniques and their implications.

The survey concludes with a synthesis of key findings and insights, highlighting future directions and potential research areas in domain adaptation and specialized NLP applications. Throughout the survey, references to recent advancements, such as the use of Gaussian Mixture Models under a differential privacy setting [9], are integrated to provide a comprehensive overview of the current state and future prospects of the field.The following sections are organized as shown in Figure 1.

2 Background and Core Concepts

2.1 Core Concepts in Domain Adaptation

Domain adaptation (DA) is crucial in NLP for transferring knowledge from a source to a target domain, particularly when target domain data is scarce. This adaptation allows models to accommodate distinct vocabularies, syntax, and semantics not present in original training datasets [1]. Maintaining performance across diverse domains amidst label and conditional shifts is a key challenge [6]. Unsupervised domain adaptation (UDA) aligns distributions without labeled target data, employing shallow and deep learning models to minimize discrepancies and enhance generalization [10]. In neural machine translation, DA is vital for translating domain-specific content with limited parallel corpora [25].

Adaptation in Named Entity Recognition (NER) models highlights the need for domain-specific strategies beyond traditional unified models, extending to event extraction where aligning source and target domain distributions is essential [26]. In healthcare, DA processes Electronic Health Records (EHRs), leveraging Large Language Models (LLMs) for improved data analysis [27]. The integration of comprehensive biomedical vocabularies underscores the complexity of managing diverse terminologies [28].

Relation extraction in data-scarce domains, such as AECO, necessitates robust DA techniques [29]. Benchmarks like CORE evaluate few-shot relation classification models, emphasizing DA's role in model adaptability [30]. The AdaptSum benchmark simulates low-resource conditions for abstractive summarization model adaptation [31].

In Biomedical and Health Informatics (BHI), DA supports interdisciplinary research in clinical decision support and medical document analysis [4]. Adapting pre-trained models to temporal and domain variations enhances document classification, addressing significant language usage shifts over time [32]. DA is foundational in modern NLP, addressing vocabulary, structural, and domain-specific challenges, facilitating knowledge transfer, and enhancing model robustness [7].

2.2 BioBERT and Biomedical NLP

BioBERT, a domain-specific language model, is pre-trained on extensive biomedical corpora to address the unique challenges in biomedical text mining [23]. It effectively captures specialized terminologies and semantic nuances, excelling in tasks like Named Entity Recognition (NER), relation extraction, and question answering [33]. BioBERT's fine-tuning capabilities for specific challenges, such as identifying rare diseases, highlight its utility [34]. Its performance in benchmarks like PubMedQA underscores its effectiveness in biomedical question answering [35]. Integration with models like GAN-BioBERT showcases adaptability in sentiment classification of clinical trial abstracts [36].

BioBERT's application in relation extraction, such as identifying chemical-gene interactions, exemplifies its broad applicability [37]. Techniques like knowledge distillation and continual learning optimize BioBERT for biomedical applications, maintaining performance with reduced computational demands [38]. Its use in zero-shot and few-shot learning frameworks for classifying biomedical articles using MeSH terms illustrates versatility in low-resource settings [39].

Advancements in optimizing transformer-based models for biomedical tasks include compact architectures like Bioformer, which reduce parameter counts while retaining performance [38]. BioBERT's application in medical entity linking within Spanish clinical texts highlights the limitations of existing multilingual models, emphasizing tailored approaches for accuracy and efficiency in clinical text analysis [40].

2.3 Legal and Clinical NLP

NLP integration in legal and clinical domains automates complex data processing and enhances decision-making but presents distinct challenges requiring specialized solutions. In legal NLP, the intricate structure and specialized vocabulary of legal language complicate the adaptation of general models. Despite BERT models' potential, their adaptation in legal NLP is under-explored, necessitating tailored BERT models for legal research [41]. Explainability in legal outcome prediction models is a major challenge, limiting their usability for legal professionals who need transparent reasoning [42]. Mining, classifying, and analyzing legal arguments, particularly in court decisions, reveal gaps between current methodologies and the nuanced understanding required by legal experts [43]. The complexity of legal sentences and specialized terminology further complicate model adaptation [44].

In clinical NLP, processing Electronic Health Records (EHRs) and other medical texts is crucial, especially in resource-limited environments where large pre-trained models may be inefficient [45]. Effective extraction of structured information from unstructured clinical texts, including identifying clinical entities and understanding temporal relationships, is essential for improving decision-making and patient outcomes [46]. Predictive analytics aids in disease identification, advancing healthcare diagnostics and treatment planning [3]. Accurate tagging of pharmaceutical organizations and drugs is challenged by naming variations and scarce labeled datasets [47]. Addressing these challenges requires specialized models and datasets to accommodate clinical texts' unique linguistic and structural characteristics.

2.4 Terminology Extraction and Few-Shot Learning

Terminology extraction and few-shot learning enhance domain adaptation, particularly in fields with unique vocabularies and limited annotated data. Few-shot learning (FSL) enables models to generalize from minimal examples, addressing data scarcity in clinical and biomedical NLP [47]. This approach adapts NLP applications to niche domains, facilitating domain-specific language extraction and understanding. In biomedical NLP, terminology extraction identifies key concepts, enhancing information retrieval and analysis. Techniques like the Clinical Concept Extraction Method (CCEM) map extracted concepts to standardized terminologies, improving interoperability and supporting biomedical knowledge graphs [46, 48, 49, 50]. Instance weighting, adversarial domain adaptation, and domain-adaptive fine-tuning enhance performance by aligning outputs with domain-specific expectations.

Few-shot learning benchmarks in BioNLP highlight the need for models capable of operating with limited samples, addressing biomedical language complexities. The AdaptKeyBERT model improves keyword extraction in low-resource domains through regularized attention mechanisms, showcasing FSL's potential in enhancing domain adaptation capabilities [51, 52, 53, 31, 5]. These methodologies collectively bolster NLP models' robustness in specialized fields, enabling adaptation to linguistic and semantic nuances of domain-specific texts.

In legal NLP, terminology extraction bridges the gap between research and practical applications. Accurate mapping of legal terminologies enhances alignment with legal professionals' needs, addressing the scarcity of useful NLP applications. Task-adaptive pre-training with word embedding regularization (TAPTER) advances fine-tuning pre-trained language models (PTLMs) for specialized terminologies in legal contexts, aligning static word embeddings with domain-specific meanings using fastText embeddings without additional corpora. TAPTER enhances standard fine-tuning methods, especially when pre-training data lacks in-domain representation [54, 55].

Terminology extraction and few-shot learning are pivotal in enhancing domain adaptation within NLP, enabling automatic identification of relevant terms from specialized corpora and facilitating efficient knowledge transfer across domains. Terminology extraction employs unsupervised methods to identify significant domain-specific words or phrases, while few-shot learning leverages large language models' in-context learning capabilities to improve relation extraction in knowledge graph generation, allowing adaptation to new domains with minimal labeled data [29, 56]. These techniques help models overcome challenges posed by domain-specific language and limited data availability, enhancing NLP applications' accuracy and relevance across diverse industries.

2.5 Industry-Specific Language Models

Industry-specific language models represent a significant advancement in NLP, offering tailored solutions for unique linguistic characteristics and challenges across sectors. These models integrate domain-specific knowledge and curated datasets, enhancing NLP applications' accuracy and relevance by addressing specialized vocabulary and contextual intricacies in fields such as law and scientific research. Advancements in Legal NLP reflect growing methodological sophistication aligned with scientific standards, while methodologies like AHAM enhance literature mining through tailored topic modeling. Few-shot learning in relation extraction models further underscores domain specificity's significance in improving NLP performance [17, 29, 5].

In the legal domain, pre-trained language models (PLMs) developed for Indian legal tasks show substantial performance improvements. These PLMs leverage extensive legal corpora to fine-tune capabilities, addressing legal language complexities and facilitating effective legal text processing [57]. Targeted models automate legal research, contract analysis, and case law retrieval, enhancing legal services' efficiency and accessibility.

The financial sector benefits from industry-specific models like BioFinBERT, a finetuned large language model analyzing sentiment in financial texts related to biotech stocks. Focusing on financial narratives, BioFinBERT provides insights into market trends and investor sentiment, aiding financial analysts and decision-makers in navigating complex data [58].

In fashion, a diverse dataset from English and Russian magazines and websites supports language models' terminology-building capabilities. This dataset captures fashion's dynamic and culturally nuanced language, improving information retrieval and analysis in fashion-related NLP tasks [59].

Detecting multiword expressions (MWEs) in specialized datasets illustrates transformer-based models' potential to tackle complex linguistic structures within specific domains. Tailoring transformer models for MWE detection enhances NLP applications' precision across fields with specialized vocabularies [60].

Developing and utilizing industry-specific language models underscores the importance of customizing NLP solutions to meet various sectors' distinct needs. These advanced models significantly enhance specialized domains' applications' performance, facilitating more accurate and contextually aware language processing, fostering innovation through improved methodologies, domain adaptation, and the capacity to handle complex terminologies and long sentences inherent in specialized texts [17, 44, 5].

3 Domain Adaptation in NLP

In the dynamic landscape of natural language processing (NLP), domain adaptation is crucial for optimizing model performance across varied contexts. This section explores the challenges of adapting NLP models to specific domains, focusing on the complexities of managing diverse data distributions and domain-specific language constraints. As illustrated in Figure 2, the hierarchical structure of domain adaptation encompasses various challenges, techniques, applications, and approaches designed to address vocabulary and semantic shifts. This figure highlights key challenges, such as differences in data distribution and domain-specific complexities, which complicate the adaptation

AI-generated, for reference only.

process. Techniques like deep learning, hierarchical methods, and model compression are detailed, emphasizing their role in enhancing model robustness. Additionally, applications span specialized fields like biomedical and legal NLP, showcasing domain adaptation's significant impact on performance. The subsequent subsection will delve into these challenges, outlining factors that complicate the domain adaptation process and their impact on model efficacy and reliability. Strategies for addressing vocabulary shifts, including hybrid learning and synthetic data generation, underscore the need for advanced representation techniques and targeted data selection.

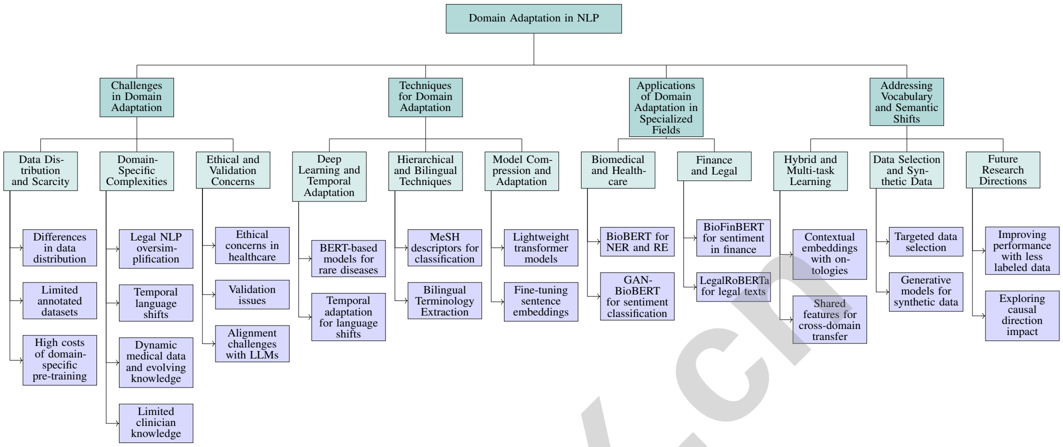

Domain Adaptation in NLP
Challenges in Domain Adaptation
Techniques for Domain Adaptation
Applications of Domain Adaptation in Specialized Fields
Addressing Vocabulary and Semantic Shifts
Data Distribution and Scarcity
Domain-Specific Complexities
Ethical and Validation Concerns
Deep Learning and Temporal Adaptation
Hierarchical and Bilingual Techniques
Model Compression and Adaptation
Biomedical and Healthcare
Finance and Legal
Hybrid and Multi-task Learning
Data Selection and Synthetic Data
Future Research Directions
Differences in data distribution
Limited annotated datasets
High costs of domain-specific pre-training
Legal NLP oversimplification
Temporal language shifts
Dynamic medical data and evolving knowledge
Limited clinician knowledge
Ethical concerns in healthcare
Validation issues
Alignment challenges with LLMs
BERT-based models for rare diseases
Temporal adaptation for language shifts
MeSH descriptors for classification
Bilingual Terminology Extraction
Lightweight transformer models
Fine-tuning sentence embeddings
BioBERT for NER and RE
GAN-BioBERT for sentiment classification
BioFinBERT for sentiment in finance
LegalRoBERTa for legal texts
Contextual embeddings with ontologies
Shared features for cross-domain transfer
Targeted data selection
Generative models for synthetic data
Improving performance with less labeled data
Exploring causal direction impact

Figure 2: This figure illustrates the hierarchical structure of domain adaptation in NLP, encompassing challenges, techniques, applications, and approaches to address vocabulary and semantic shifts. It highlights key challenges such as data distribution differences and domain-specific complexities. Techniques like deep learning, hierarchical methods, and model compression are detailed, emphasizing their role in enhancing model robustness. Applications span specialized fields like biomedical and legal NLP, showcasing domain adaptation's impact on performance. Strategies for addressing vocabulary shifts include hybrid learning and synthetic data generation, underscoring the need for advanced representation techniques and targeted data selection.

3.1 Challenges in Domain Adaptation

Domain adaptation in NLP encounters challenges due to data distribution differences, limited annotated datasets, and domain-specific language intricacies. The scarcity of task-specific datasets, such as those for causality extraction from Clinical Practice Guidelines, impedes model evaluation and development [61]. High costs associated with domain-specific pre-training further limit the scalability of tailored models [33].

In legal NLP, computational methods often oversimplify legal arguments, missing the nuanced structures needed for comprehensive analysis [43]. Benchmarks frequently overlook temporal language shifts, affecting models trained on outdated data [32]. The biomedical field faces dynamic medical data, complex language, and evolving knowledge that traditional systems struggle to accommodate [62]. Limited clinician knowledge about rare diseases and complex nomenclature complicate adaptation [34]. Integrating large language models (LLMs) into healthcare raises ethical concerns, validation issues, and alignment challenges [4].

Bilingual terminology extraction is hindered by differing corpus distributions, complicating optimization [63]. Adapting general-purpose LLMs to specialized biomedical tasks requires understanding and executing complex instructions, inadequately addressed by current benchmarks [64].

In clinical NLP, the complexity of clinical language and ethical considerations surrounding sensitive data limit annotated dataset availability [46]. Reliable labeled datasets are challenging to obtain due to labor-intensive manual labeling [47]. Existing benchmarks struggle to generalize to mathematical domains, complicating mathematical text processing [65].

Innovative strategies are essential to address data scarcity, distributional discrepancies, and computational efficiency needs in NLP. Domain-specific adaptations, like AHAM for literature mining with generative models such as LLaMa2, exemplify advancements in topic modeling [17, 53, 51, 5].

Insights from the legal domain highlight the sophistication of NLP methods, emphasizing data availability and reproducibility. Advancements in domain adaptation for sentence embeddings and retrieval models illustrate potential performance enhancements by leveraging domain-specific attributes, even in zero-shot scenarios without direct target data access.

3.2 Techniques for Domain Adaptation

Domain adaptation in NLP employs various techniques to transfer knowledge from a source to a target domain, particularly in contexts with limited labeled data and domain-specific language features. Deep learning techniques, notably BERT-based models, show promise in recognizing rare diseases from unstructured text through domain adaptation [34]. These models leverage pre-trained language representations to adjust to specialized vocabulary and semantic nuances.

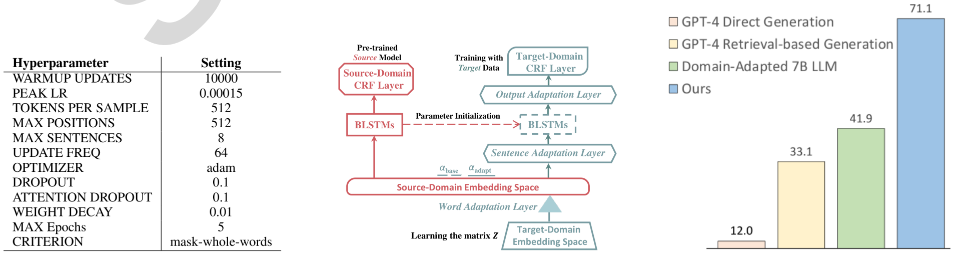
Temporal adaptation, combined with domain adaptation, offers a novel perspective on aligning models with temporal language shifts, significantly impacting performance [32]. This approach underscores the importance of considering temporal factors in maintaining model relevance.

In zero-shot and few-shot learning contexts, leveraging hierarchical relationships encoded in MeSH descriptors enhances classification performance, particularly in biomedical scenarios with data scarcity [39]. This method improves adaptability in situations with minimal training data by utilizing structured knowledge.

Bilingual terminology extraction techniques, such as Bilingual Terminology Extraction Using Multi-level Termhood (BTE-MLT), enhance domain adaptation by assessing termhood through corpus comparison [63]. This approach is crucial for cross-lingual capabilities.

Adapting lightweight transformer models for domain-specific tasks addresses the high computational and memory demands of existing models, improving accessibility for researchers and practitioners [38]. By compressing large models into efficient versions, these techniques facilitate NLP solution deployment in resource-constrained environments.

The diverse techniques in domain adaptation highlight the necessity of tailoring NLP models to specific linguistic and semantic characteristics of various domains. By integrating methodologies such as deep learning, temporal adaptation, hierarchical knowledge leveraging, and model compression, these techniques enhance NLP model robustness and applicability. They enable effective handling of domain-specific texts, particularly in scenarios where training and test data originate from different distributions. For instance, domain adaptation improves retrieval accuracies by fine-tuning sentence embeddings tailored to specific domains, crucial in document retrieval and machine translation. Incremental adaptation using newly available in-domain data allows rapid adjustments without complete retraining, enhancing efficiency in real-world applications [11, 53, 1, 66].



Hyperparameter	Setting
WARMUP UPDATES	10000
PEAK LR	0.00015
TOKENS PER SAMPLE	512
MAX POSITIONS	512
MAX SENTENCES	8
UPDATE FREQ	64
OPTIMIZER	adam
DROPOUT	0.1
ATTENTION DROPOUT	0.1
WEIGHT DECAY	0.01
MAX Epochs	5
CRITERION	mask-whole-words

(a) Hyperparameter Settings for a Machine Learning Model[67]

(b) Adaptive Transfer Learning for Cross-Domain Sentence Representation Learning[68]

(c) Comparison of Generation Performance in Text Generation Tasks[69]

Figure 3: Examples of Techniques for Domain Adaptation

As illustrated in Figure 3, domain adaptation is essential in NLP for enabling models to perform effectively across different domains by leveraging knowledge from a source domain to a target domain. The figure highlights various techniques for domain adaptation, emphasizing their significance in enhancing model performance. The first subfigure details hyperparameter settings critical for tuning machine learning models across domains, showcasing parameters like warmup updates, peak learning rate, and tokens per sample. This meticulous configuration is vital for optimizing model performance.

The second subfigure presents a flowchart of adaptive transfer learning, underscoring the importance of cross-domain sentence representation, which utilizes a pre-trained source model and a source-domain CRF layer to facilitate knowledge transfer. The third subfigure compares the generation performance of different methods in text generation tasks, including GPT-4 Direct Generation, GPT-4 Retrieval-based Generation, and a Domain-Adapted 7B LLM, illustrating the efficacy of domain adaptation techniques in enhancing text generation quality.

3.3 Applications of Domain Adaptation in Specialized Fields

Domain adaptation significantly enhances NLP applications across specialized fields by addressing unique linguistic challenges and data scarcity. In biomedical NLP, domain adaptation techniques are crucial for tasks like named entity recognition (NER) and relation extraction (RE), essential for identifying complex relationships within medical texts. The application of BioBERT, a domain-specific language model, exemplifies this by effectively recognizing biomedical entities and relationships, thereby improving performance in biomedical text mining [23]. Instruction-tuned models have also demonstrated substantial gains in biomedical NER and RE tasks, showcasing the efficacy of domain adaptation in refining model capabilities [64].

In rare disease recognition, domain adaptation with BERT-based models has improved diagnostic accuracy, highlighting the potential of these techniques to enhance outcomes in specialized healthcare fields [34]. The integration of BioBERT with deep learning approaches illustrates successful applications of domain adaptation in merging diverse datasets to enhance biomedical NLP tasks [37]. Additionally, the GAN-BioBERT algorithm has proven effective in sentiment classification within clinical trial literature, outperforming prior methodologies and manual expert evaluations [36].

In finance, domain adaptation is exemplified by BioFinBERT, a model fine-tuned to analyze sentiment in financial texts related to biotech companies. This adaptation is critical during clinical and regulatory events, where accurate sentiment analysis informs strategic decision-making [58].

The legal domain also benefits from domain adaptation, with models like LegalRoBERTa being further pre-trained on legal corpora to navigate the complexities of legal texts. This approach employs adapters to enhance model efficiency, facilitating tasks such as legal judgment prediction and cross-lingual transfer learning [54]. Domain adaptation in legal NLP extends to extracting insights from legal documents, aiding investor decision-making and fraud detection through the analysis of complex regulatory frameworks [8].

As illustrated in Figure 4, the applications of domain adaptation in specialized fields are multi-faceted, highlighting its impact on biomedical NLP, financial sentiment analysis, and legal NLP tasks. The diagram categorizes specific models and methodologies, such as BioBERT, BioFinBERT, and LegalRoBERTa, showcasing their roles in enhancing performance in their respective domains.

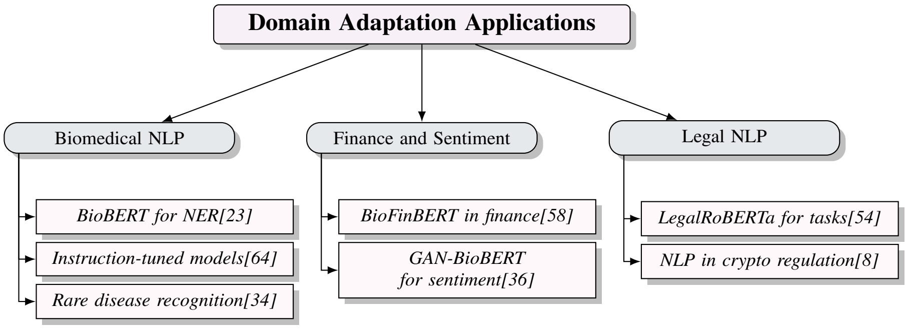

Domain Adaptation Applications
Biomedical NLP
Finance and Sentiment
Legal NLP
BioBERT for NER[23]
Instruction-tuned models[64]
Rare disease recognition[34]
BioFinBERT in finance[58]
GAN-BioBERT for sentiment[36]
LegalRoBERTa for tasks[54]
NLP in crypto regulation[8]

Figure 4: This figure illustrates the applications of domain adaptation in specialized fields, highlighting its impact on biomedical NLP, financial sentiment analysis, and legal NLP tasks. The diagram categorizes specific models and methodologies, such as BioBERT, BioFinBERT, and LegalRoBERTa, showcasing their roles in enhancing performance in their respective domains.

3.4 Addressing Vocabulary and Semantic Shifts

Addressing vocabulary and semantic shifts in domain adaptation is critical, as it directly impacts a model's generalization across domains. One effective strategy involves hybrid methods that enhance the semantic representation of domain-specific concepts. In the biomedical domain, combining contextual embeddings with structured knowledge from ontologies has proven effective in improving the representation of biomedical concepts, thereby enhancing adaptability [70].

Advancements in aligning domain distributions have emerged through techniques employing domain adaptation and multi-task learning, which successfully extract relevant features across different domains and languages. By utilizing shared features and representations, these methods enhance knowledge transfer effectiveness, particularly in document retrieval and relation extraction contexts. For instance, fine-tuning sentence embeddings improves retrieval accuracies, while leveraging large language models for few-shot learning facilitates the adaptation of models to specialized domains like Architecture, Construction, Engineering, and Operations (AECO) [53, 29].

Promoting diverse feature learning is another promising strategy, as it encourages models to learn various features, reducing reliance on domain-specific characteristics and enhancing generalization to new domains [71]. This approach mitigates overfitting risks, thereby improving model robustness across contexts.

Data selection is crucial for enhancing domain adaptation effectiveness. Developing benchmarks that emphasize meticulous data selection underscores the need to choose datasets aligned with the target domain's specific characteristics and terminology. This is particularly vital in fields like built asset management, where technical text complexity necessitates effective mapping to domain-specific classification systems. In NLP tasks, targeted data selection and fine-tuning can lead to significant model accuracy improvements [53, 48, 72, 73]. By selecting data that closely matches the vocabulary and semantic structures of the target domain, models are better prepared to handle domain shifts.

Leveraging generative models to create synthetic data that captures target domain properties is a promising approach. This method enables retrieval models to learn from generated data reflecting the target domain's characteristics, enhancing adaptability to new vocabulary and semantic contexts [51].

Future research should focus on improving model performance with less labeled data and exploring applications in other NLP tasks [74]. Additionally, understanding causal direction's impact on NLP model performance could guide further research [75]. The primary challenge in domain adaptation lies in the extensive retraining required by existing methods, often making them time-consuming and impractical in dynamic translation environments [66].

A comprehensive strategy is essential for effectively addressing vocabulary and semantic shifts across domains. This strategy should integrate advanced representation techniques, such as those employed in LLMs like LLaMa2, alongside diverse feature learning methods that leverage domain adaptation to improve topic modeling precision. Strategic data selection, including curated synthetic data generated from textual domain descriptions, is crucial for enhancing retrieval performance. Incorporating unsupervised term extraction techniques can further refine the understanding of domain-specific terminology, leading to more accurate analyses in scientific literature and other fields [51, 76, 71, 56, 5]. Collectively, these methods contribute to the development of adaptable and robust NLP models capable of effectively navigating the complexities of domain adaptation.

4 Applications of BioBERT in Biomedical Text Analysis

4.1 BioBERT in Biomedical Text Mining and Question Answering

BioBERT plays a pivotal role in biomedical text mining and question answering, leveraging pre-training on vast biomedical corpora to adeptly handle complex medical terminologies and relationships. Its efficacy is particularly pronounced in Named Entity Recognition (NER), Relation Extraction (RE), and Question Answering (QA), where it incorporates domain-specific linguistic features [23]. In biomedical QA, BioBERT has demonstrated significant accuracy improvements on datasets like PubMedQA, underscoring its practical utility [37].

BioBERT's integration with frameworks combining classification and generation components enhances its ability to process and synthesize biomedical information, as seen in its effective extraction of chemical-gene relationships [37]. Its proficiency in multi-hop QA is supported by datasets for training knowledge graph QA systems, highlighting its capability in complex information retrieval [23].

In literature-based discovery frameworks, BioBERT's robustness is evident in extracting insights from extensive biomedical texts. Methods enhancing zero-shot performance further establish its status in biomedical NLP applications [4]. The efficiency of models like Bioformer, achieving high accuracy with fewer parameters, underscores BioBERT's potential to optimize processing times and enhance service delivery [32]. BioBERT's integration into various applications highlights its ongoing development and promise for future advancements in extracting and interpreting complex biomedical information.

4.2 Applications in Clinical Text Analysis

In clinical text analysis, BioBERT is indispensable, utilizing pre-training on extensive biomedical literature to enhance the processing and interpretation of complex clinical data [23]. Its application in clinical text classification and NER significantly improves insights extraction from unstructured clinical data, with models like Clinical BioBERT achieving substantial performance improvements on datasets such as EBM-COMET [77].

BioBERT's role in frameworks like PharmKE exemplifies its comprehensive entity recognition and knowledge extraction capabilities, employing multi-stage deep learning for enhanced clinical entity recognition [47]. Its adaptability is further demonstrated in medical entity normalization systems like ClinLinker, which utilize advanced pipelines for candidate retrieval and re-ranking [40].

Beyond entity recognition, BioBERT contributes to sentiment analysis in clinical texts. The GAN-BioBERT algorithm employs a semi-supervised approach to classify clinical trial abstracts by sentiment, enhancing analysis granularity [36]. BioBERT also aids in causality extraction from medical guidelines, supported by specific benchmarks validating its potential [61].

BioBERT enhances understanding and decision-making by recognizing rare diseases and their manifestations [34]. Its effectiveness in classifying biomedical articles during the COVID-19 pandemic showcases its versatility in low-resource settings [39]. Additionally, BioBERT optimizes treatment strategies based on patient data, advancing healthcare informatics [3]. Its integration into clinical text analysis underscores its significant contributions to healthcare by enhancing clinical decision-making processes and supporting biomedical research evolution [37].

4.3 Enhancements and Future Directions

Enhancing BioBERT's applications in biomedical text analysis involves refining training procedures and model architectures to optimize performance across diverse biomedical tasks [38]. Expanding datasets and integrating advanced language models could significantly enhance BioBERT's adaptability and effectiveness [64].

Future research should focus on integrating multi-modal data sources to improve relationship extraction within biomedical texts, enhancing robustness. Leveraging diverse healthcare data, including clinical and molecular profiles, can improve interpretability and effectiveness in tasks like NER and relation extraction, leading to precise clinical decision support and personalized treatment strategies. Addressing data privacy and model customization challenges is crucial for BioBERT's successful deployment in healthcare contexts [22, 3, 23, 24]. Advanced entity recognition techniques could further enhance BioBERT's capabilities in identifying complex biomedical entities and relationships.

Expanding test sets and experimenting with different LLMs for data generation could improve BioBERT's adaptability and performance in diverse biomedical tasks [35]. Incorporating additional datasets or refining models to handle incomplete knowledge graphs could enhance BioBERT's utility in biomedical QA [35].

In sentiment analysis, expanding datasets with more expert raters and exploring finer-grained sentiment classification could enhance data analysis granularity [36]. Further tuning of models like BioFinBERT and investigating additional datasets could improve predictive capabilities in financial texts related to biotech companies [58].

Integrating conversational contexts into relation extraction frameworks and developing metrics that capture human dialogue intricacies could enhance BioBERT's utility in clinical decision support, particularly in patient-provider interactions. Enhancing robustness against ambiguous terminologies and exploring applications beyond the current scope, such as e-commerce, could broaden BioBERT's applicability [63].

These enhancements and future research directions will advance BioBERT applications in biomedical text analysis, ensuring its relevance and effectiveness in addressing complex biomedical challenges. By refining methodologies and exploring innovative applications, BioBERT can solidify its status as a premier model in biomedical NLP. This includes improving pre-training on extensive biomedical corpora to better address challenges in biomedical text mining like NER, relation extraction, and QA. Ongoing efforts to fine-tune BioBERT for specific healthcare tasks—coupled with systematic data gathering, annotation, and preprocessing—will ensure its continued effectiveness in enhancing clinical decision support and information retrieval while addressing ethical considerations like patient privacy and data security [3, 23].

As shown in Figure 5, this figure illustrates the enhancements and future directions for BioBERT, focusing on model refinement, data integration, and application expansion. It highlights the importance of refining training procedures and integrating advanced models to enhance BioBERT's performance. The integration of multi-modal and healthcare data is crucial for improving interpretability and effectiveness, while expanding applications to areas like sentiment analysis and e-commerce broadens BioBERT's utility. The figure captures BioBERT's applications and future potential, juxtaposing it with other significant NLP advancements. The depiction of the BioBERT/BERT Base Model highlights its sophisticated architecture, where tokenized inputs are processed using contextual word embedding vectors, demonstrating its proficiency in interpreting complex biomedical texts. This is complemented by illustrations of Legal NLP Benchmarks and in-context learning settings, providing a broader perspective on how similar NLP technologies are adapted across various domains. The exploration of zero-shot, one-shot, and few-shot learning settings underscores the adaptability and potential for enhancement in BioBERT's application, paving the way for more nuanced and accurate biomedical text analysis. These advancements signify current capabilities while pointing towards future directions for refining and integrating BioBERT into more complex biomedical informatics systems [78, 79, 80].

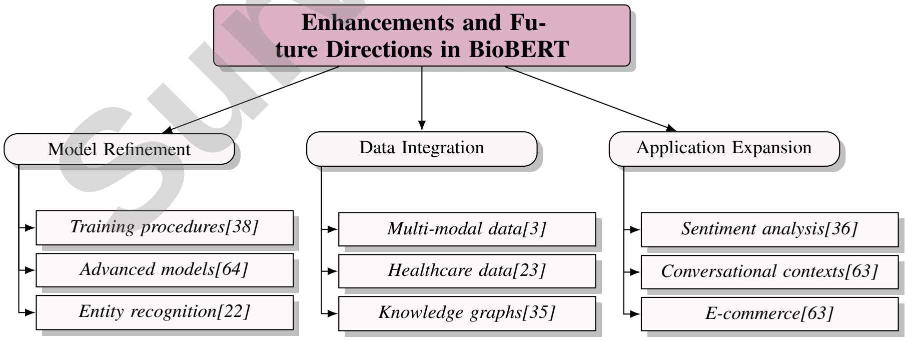


Figure 5: This figure illustrates the enhancements and future directions for BioBERT, focusing on model refinement, data integration, and application expansion. It highlights the importance of refining training procedures and integrating advanced models to enhance BioBERT's performance. The integration of multi-modal and healthcare data is crucial for improving interpretability and effectiveness, while expanding applications to areas like sentiment analysis and e-commerce broadens BioBERT's utility.

5 Legal NLP and Its Applications

The exploration of Natural Language Processing (NLP) in the legal domain involves navigating the complexities of legal language and addressing the operational needs of legal professionals who require precise interpretations of legal texts. This section delves into the challenges faced by practitioners and researchers, alongside advancements that have enhanced the effectiveness of legal applications.

5.1 Challenges and Advancements in Legal NLP

Legal NLP is characterized by the complexities of legal language and the specific requirements of legal practitioners. A significant challenge is replicating the nuanced understanding and moral reasoning essential for judicial decision-making, which current NLP models struggle to achieve [81]. Limited datasets exacerbate this issue, leading to overfitting and poor generalization to real-world legal texts [44]. Additionally, existing methods often fail to accurately analyze the specialized language in biotech press releases, affecting sentiment prediction [58].

The scarcity of comprehensive datasets, especially for rare argument types, further limits model performance [43]. This highlights the need for diverse datasets to improve model training and evaluation. Ethical considerations also present challenges, requiring a nuanced understanding of NLP systems' capabilities and limitations [82].

Privacy concerns and the demand for model efficiency complicate legal NLP system development. Current studies often inadequately address data privacy and clinical validation, which are crucial for deploying NLP technologies in sensitive legal contexts [62]. Integrating ethical considerations emphasizes the importance of academic freedom and interdisciplinary collaboration to create responsible technologies [82].

Despite these challenges, advancements like specialized benchmarks and models such as LEGAL-BERT have shown significant accuracy improvements, demonstrating the potential of domain-specific adaptations [41]. Multilingual multitask benchmarks like LEXTREME enhance model evaluation across diverse languages and datasets [83]. Frameworks like PharmKE, with their modular design and visualization capabilities, provide advantages in user understanding and model integration [47].

Ongoing research and development are driving advancements in legal NLP. Addressing dataset diversity, ethical considerations, and model efficiency is essential for harnessing NLP technologies' potential in the legal domain. Enhancing dataset diversity improves model applicability across contexts, while prioritizing ethical considerations mitigates biases and promotes fair representation in legal outcomes. Optimizing model efficiency is crucial given the complexities of legal documents, which pose unique summarization challenges. Tackling these issues can significantly enhance legal services' effectiveness and accessibility, contributing to a more equitable legal system [17, 84].

5.2 Privacy-Preserving Techniques

In legal NLP applications, privacy-preserving techniques are crucial due to the frequent involvement of sensitive data. Integrating differential privacy into transformer models' pre-training enhances performance while safeguarding data privacy, ensuring sensitive legal information remains secure during the learning process [85]. Differential privacy provides a robust framework for data confidentiality, particularly in unsupervised domain adaptation scenarios [9].

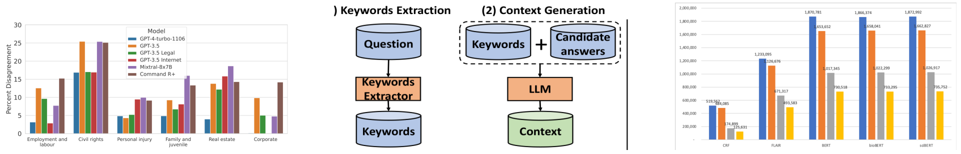
The implementation of privacy-preserving methodologies is further supported by the need to evaluate AI-generated legal responses' factual correctness, where accuracy is paramount [86]. Metrics that assess performance while ensuring privacy is upheld throughout the process are necessary. Integrating external knowledge sources can enhance language model responses' accuracy and relevance in specialized legal applications, improving legal NLP systems' reliability [2].

Privacy-preserving techniques augment legal decision support systems designed to assist legal professionals while emphasizing caution to prevent biases and uphold ethical considerations [87]. Balancing academic freedom and privacy is a recurring theme in legal NLP applications' development, underscoring ethical considerations' importance in designing and deploying these technologies [82].

Adopting privacy-preserving techniques is crucial for maintaining sensitive data confidentiality while enhancing language models' performance and applicability in legal contexts. By integrating differential privacy with advanced methodologies, legal NLP systems can optimize performance while protecting sensitive information, enabling legal professionals to utilize these tools effectively while adhering to ethical standards and addressing diverse legal norms. Ongoing research emphasizes balancing academic freedom with moral considerations in deploying these technologies, fostering a more responsible and effective legal NLP landscape [17, 82, 85].

Figure 6 illustrates the critical role of privacy-preserving techniques in legal NLP through three visual examples. The bar chart highlights percent disagreement among models across legal categories,

AI-generated, for reference only.



(a) The image depicts a bar chart comparing the percent disagreement of different models across various legal categories.[86]

(b) Context Generation from Keywords and Candidate Answers[13]

(c) Comparison of CRF, FLAIR, BERT, bioBERT, and sciBERT models in terms of accuracy[88]

Figure 6: Examples of Privacy-Preserving Techniques

emphasizing consensus challenges in legal AI solutions. The flowchart details context generation from keywords and candidate answers, underscoring precise context creation's importance in legal inquiries. The third image compares machine learning models' accuracy, demonstrating ongoing advancements in legal text processing model performance [86, 13, 88].

5.3 Terminology Extraction in Legal Texts

Terminology extraction in legal texts is a fundamental aspect of legal NLP, focusing on identifying and interpreting domain-specific language within legal documents. This process is vital for developing NLP models capable of navigating the complex structures and specialized vocabularies typical of legal texts [44]. Extracting legal terminology facilitates precise identification of legal entities, essential for tasks like legal document classification and summarization [43].

Advanced methodologies, such as parameter-efficient legal domain adaptation, leverage unsupervised legal data for pre-training, enabling models to adapt effectively with minimal parameter tuning. This approach is valuable in legal contexts where privacy and data sensitivity are paramount, allowing secure processing of sensitive legal information [82]. Utilizing annotated datasets from publicly available legal corpora supports legal terminology extraction by providing models with the necessary context to accurately process legal documents [44].

New annotation schemes for legal arguments, aligned with legal theory and practice, underscore context-aware terminology extraction's importance in legal applications [43]. These schemes enable models to learn legal documents' specific language and context, enhancing their ability to perform complex legal NLP tasks.

Terminology extraction underpins advanced legal NLP systems' development, facilitating effective summarization, argument mining, and comprehension of complex legal documents. This foundational work is essential for improving legal NLP applications' accuracy and sophistication, as evidenced by advancements in summarization methods and argument mining techniques aligning with legal reasoning practices [17, 84, 43]. Employing annotated datasets and advanced methodologies enhances legal terminology extraction's accuracy and reliability, contributing to more effective and accessible legal services.

5.4 Sector-Specific Legal NLP Applications

Sector-specific NLP applications in the legal domain are tailored to meet various legal contexts' unique requirements, improving legal processes' efficiency and precision. Specialized benchmarks, such as those for Indian legal systems, address linguistic and legal complexities specific to jurisdictions, providing structured frameworks for evaluating legal NLP systems [43]. These benchmarks facilitate models sensitive to specific legal systems' nuances, enhancing NLP solutions' applicability and effectiveness [44].

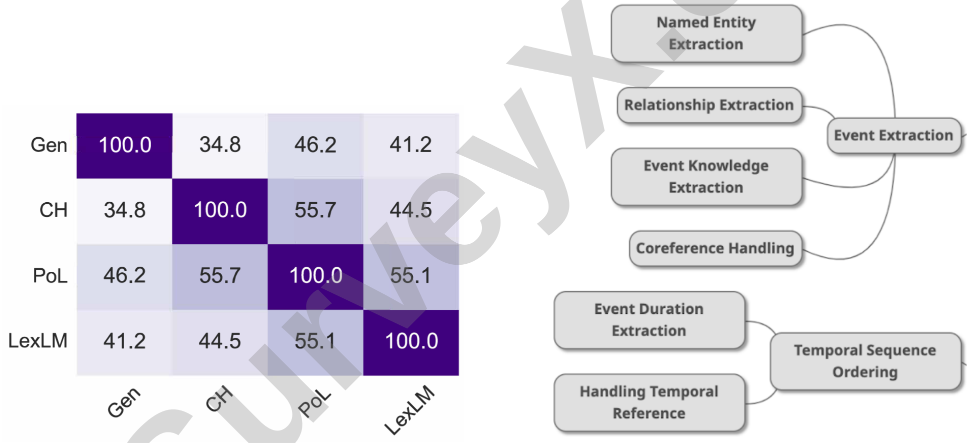
In legal entity recognition, precise entity typing is crucial for tasks like NER and question answering, enhancing legal NLP applications' performance, particularly in sectors where accurate legal entity identification is vital for compliance monitoring and contract analysis. Integrating privacy-preserving techniques, such as differential privacy in pre-training BERT models, enhances performance and safeguards sensitive legal data, ensuring confidentiality in legal proceedings [82].

Expert-validated benchmarks for legal question-answering provide a novel dataset of expert-validated legal question-answer pairs and an automatic evaluation protocol based on factuality, ensuring legal AI solutions' accuracy and reliability [42]. Utilizing public legal texts enhances legal NLP models' relevance and applicability [44].

Simplifying legal texts using unsupervised methods showcases NLP's potential to make legal language more accessible while maintaining semantic coherence. This approach benefits sectors where simplifying complex legal documents improves understanding and accessibility for non-experts [81]. However, the method shows limitations in handling complex multi-label classification tasks, indicating its suitability may vary across legal NLP applications [44].

Future research in sector-specific legal NLP applications should explore additional data augmentation techniques and cross-jurisdictional transfer to enhance model adaptability and robustness [82]. Adapting methods to other legal domains and languages and integrating additional data sources could further improve model robustness [42]. Expanding benchmarks to include multilingual datasets and exploring additional legal generative tasks is a promising direction for future research [43].

Sector-specific legal NLP applications are evolving through specialized benchmarks, privacy-preserving techniques, and real-world legal datasets. These advancements enhance legal processes' accuracy, reliability, and accessibility, contributing to more efficient and effective legal services. Future research should continue investigating hybrid systems that leverage expert knowledge and data-driven insights, alongside emerging trends and evaluation metrics in legal NLP [81].



(a) The image is a confusion matrix showing the performance of different models on a classification task.[89]

(b) Event Extraction[78]

Figure 7: Examples of Sector-Specific Legal NLP Applications

Figure 7 illustrates how NLP can be tailored to meet the legal sector's specific demands. The confusion matrix evaluates various models' performance on a classification task, highlighting NLP models' potential in enhancing legal document classification. The event extraction flowchart maps the process of extracting relevant events from legal texts, including named entity extraction, relationship extraction, and temporal sequence ordering. These visualizations provide a comprehensive overview of how sector-specific NLP applications streamline legal processes, improve accuracy, and enhance legal practitioners' efficiency [89, 78].

6 Sector-Specific NLP Applications

6.1 Frameworks and Taxonomies for Sector-Specific NLP

Developing frameworks and taxonomies for sector-specific NLP is vital for structuring methodologies across industries, enhancing language models' precision by addressing unique sector challenges. Integrating unsupervised Neural Machine Translation (NMT) with continual learning enables seamless domain adaptation, maintaining model relevance across diverse environments [90]. Sector-specific taxonomies organize linguistic features and terminologies, facilitating the creation of specialized models for processing industry-specific texts. In finance, these taxonomies handle the intricate language of financial documents, aiding sentiment analysis and risk assessment. In legal contexts, they improve research efficiency and contract analysis by structuring legal language and concepts, supporting advanced summarization and argument mining techniques [17, 89, 84, 43].

Frameworks leveraging distributed computing and parallel processing enhance NLP efficiency in data-intensive sectors, supporting machine learning algorithms tailored to specific computational needs. In legal NLP, methodologies align with data availability and reproducibility standards, as shown by the analysis of over six hundred papers. The AHAM methodology exemplifies effective adaptation of topic modeling frameworks for scientific text analysis, refining topics through domain expertise and generative language models [17, 5].

Robust frameworks and taxonomies are crucial for advancing sector-specific NLP applications, empowering models to navigate linguistic and contextual intricacies in industries like law and healthcare. This fosters innovation and enhances language processing technologies' effectiveness, with Legal NLP addressing methodological standards and accessibility challenges, and clinical text analysis frameworks improving information extraction for better public health outcomes [17, 50, 21, 5].

6.2 Biomedical and Healthcare NLP Applications

NLP's application in biomedical and healthcare sectors is transformative, enhancing diagnostics and patient care through advanced language models and structured data processing [4]. Models like BioBERT, with contextual embeddings, capture clinical terminology intricacies, aiding in accurate classification of clinical outcomes [91]. This precision is critical for tasks like mining adverse drug reactions, where NLP pipelines in libraries like Spark NLP integrate document classification, named entity recognition, and relation extraction to improve healthcare data analysis [14].

Benchmarks for synonymy prediction provide evaluation frameworks for model performance, essential for refining biomedical NLP [92]. Datasets like PubMedQA, with diverse biomedical research questions, drive sophisticated NLP application development in healthcare [93]. NLP applications, like chatbots, effectively disseminate COVID-19 information, showcasing NLP's potential in public health communication [94].

NLP's precision in term extraction is exemplified by comparisons among tools like ChatGPT, SketchEngine, and TBXTools, where ChatGPT shows higher precision despite lower recall [59]. This capability is vital for developing comprehensive biomedical vocabularies and enhancing information retrieval in healthcare settings.

Integrating NLP in biomedical and healthcare applications exemplifies advancements in language processing technologies, enabling precise information extraction from clinical texts and electronic health records (EHRs), enhancing diagnostics and patient care. These advancements streamline communication among healthcare professionals, facilitate biomedical knowledge graph development, and improve clinical coding accuracy. NLP advances medical research by extracting relevant information from vast biomedical literature, supporting decision-making in precision oncology and infectious disease management, highlighting its essential contribution to transforming healthcare practices and improving patient outcomes [95, 49, 22, 50, 48].

6.3 Legal and Financial NLP Applications

NLP enhances text analysis efficiency and decision-making in legal and financial industries. In the legal domain, NLP automates complex legal text processing, aiding research, contract analysis, and compliance monitoring. The LegalEval benchmark, with diverse legal text datasets, is crucial for evaluating NLP models' performance, ensuring accurate legal language interpretation [16].

In finance, NLP processes vast financial data, enabling informed decision-making and risk assessment. Domain-specific models, fine-tuned for financial sentiment analysis, predict market trends and investor sentiment, supporting strategic business decisions. These models analyze financial narratives, extracting actionable insights from complex documents, improving financial institutions' agility. Advanced NLP techniques summarize lengthy texts, identify key information, and facilitate regulatory compliance, enabling informed decisions in a dynamic environment [48, 8, 5, 84].

AI-generated, for reference only.

NLP's integration in legal and financial industries showcases its transformative potential by automating and optimizing complex processes. Recent advancements in Legal NLP, including methodological sophistication and specialized summarization models, highlight these technologies' growing capacity to handle lengthy legal documents. Ongoing research emphasizes improving code reproducibility and data availability, essential for enhancing NLP applications' effectiveness. These trends indicate a promising future for NLP in automating legal document summarization and processing, despite challenges in adapting general models to specialized legal language and structure [17, 44, 84]. By leveraging specialized datasets and tailored models, NLP enhances text analysis accuracy and efficiency, contributing to more effective legal and financial services.

6.4 NLP in Emerging Domains

NLP's application in emerging domains showcases its versatility and innovation potential across industries. In e-commerce, NLP improves customer service through advanced chatbots and virtual assistants, enhancing user experience and operational efficiency in online retail [87, 17]. These applications streamline operations by automating routine interactions, allowing human agents to focus on complex tasks.

In environmental science, NLP analyzes climate-related data, facilitating insights for policy decisions and environmental strategies. By examining textual data from sources like scientific publications and social media, NLP models detect trends and patterns essential for addressing environmental challenges. Advanced methodologies, like topic modeling and domain adaptation, enhance analytical precision, supporting informed decision-making in environmental research and policy development [17, 84, 5].

In education, NLP contributes through intelligent tutoring systems delivering tailored educational experiences. These systems leverage linguistic knowledge from pretraining, using domain-specific adaptations and syntactic information integration to enhance learning outcomes by providing personalized feedback based on individual student needs [96, 89, 5].

In mental health, NLP applications analyze text-based communications for mental well-being indicators, employing techniques like Named Entity Recognition (NER) and Relation Extraction (RE) to identify key entities in clinical data, improving diagnostic accuracy. Pre-trained transformer models extract relevant information, providing insights into mental health statuses and aiding in intervention development [49, 48, 50].

NLP's increasing integration across diverse sectors highlights its transformative potential. Recent advancements in Legal NLP, including sophisticated pre-trained models tailored to specific contexts, improve tasks like statute identification and judgment prediction. Methodologies like AHAM enhance scientific literature mining by adapting models to domain-specific requirements, while unsupervised term extraction techniques revolutionize information retrieval processes. These trends indicate NLP's potential to reshape operational efficiencies and analytical capabilities in various industries [17, 57, 5, 56]. Leveraging advanced models and processing techniques, NLP drives innovation and enhances effectiveness in new fields.

6.5 Challenges and Innovations in Sector-Specific NLP

Sector-specific NLP applications face challenges due to diverse linguistic demands and specialized vocabularies across industries. Adapting models to domains with limited data is a significant challenge, necessitating innovative approaches like few-shot learning to enhance adaptability and performance in data-scarce environments [97, 98].

Task-adaptive pre-training (TAPTER) offers a promising solution, adapting pre-trained models to specific domains without extensive corpora, reducing computational costs and enhancing efficiency [55]. TAPTER streamlines adaptation, valuable in sector-specific applications with limited computational resources.

Addressing data scarcity, sector-specific NLP must improve rare phrase detection methods, crucial for processing specialized texts. Future research could explore rare linguistic elements' applicability to sequence data like genomic sequences or speech [99].

Structured benchmarks, like those for evaluating models in the mathematical domain, provide resources for assessing performance and driving innovation in sector-specific applications [65]. These benchmarks ensure models meet different sectors' unique demands.

Challenges in sector-specific NLP underscore the need for innovation and research. Integrating methodologies like few-shot learning and task-adaptive pre-training, alongside comprehensive benchmarks, enhances model performance in specialized domains like legal and clinical NLP, where annotated datasets are scarce. Domain-specific pre-training yields significant performance improvements in legal tasks, while few-shot learning techniques like Siamese Neural Networks are effective in clinical settings with limited data. Focusing on these strategies tailors NLP solutions to meet various sectors' demands, driving advancements in the field [100, 44, 84, 101].

7 Privacy-Preserving NLP

7.1 Ethical Considerations and Data Privacy

Ethical considerations and data privacy are fundamental in NLP, especially with the proliferation of Large Language Models (LLMs). Chen et al. [12] advocate for ethical AI practices, emphasizing transparency, fairness, and accountability to maintain public trust and align AI systems with societal values. Data privacy is essential, given regulations like the GDPR, which protect individuals' privacy rights [102]. Integrating privacy-preserving techniques in NLP is crucial for compliance while ensuring AI systems' usability and performance. Sousa et al. [103] highlight the need for ongoing research to adapt to evolving privacy regulations and threats, ensuring NLP technologies' security and effectiveness.

Privacy-preserving unsupervised domain adaptation (UDA) methods, as discussed by An et al. [9], enhance privacy protection without significantly compromising performance. These methods integrate seamlessly into existing UDA frameworks, offering robust solutions for safeguarding sensitive information in NLP applications. The FGraDA benchmark proposed by Zhu et al. [104] provides insights into fine-grained domain adaptation, facilitating privacy measures in machine translation.

In multilingual medical question answering, Vinod et al. [15] emphasize maintaining user privacy while delivering accurate medical information, highlighting multilingual capabilities and low-resource requirements for privacy-conscious NLP solutions. The ethical and data privacy challenges in NLP necessitate a balanced approach that prioritizes individual rights and regulatory compliance while fostering innovation. By systematically integrating advanced privacy-preserving techniques, the NLP community can enhance technology effectiveness while responsibly managing sensitive data, particularly in compliance with regulations like the GDPR. This focus addresses privacy threats and human biases while navigating the complex ethical landscape of NLP applications in sensitive domains, including legal text analysis [82, 103].

7.2 Techniques for Privacy Preservation

Privacy preservation in NLP focuses on protecting sensitive information while maintaining language models' functionality. Sousa et al. [103] categorize privacy-preserving methods into data safeguarding, trusted, and verification methods, offering a structured framework for understanding diverse approaches to ensuring data privacy in NLP applications. Data safeguarding methods employ encryption, differential privacy, and secure multi-party computation to protect personal data, preventing unauthorized access and ensuring data confidentiality, particularly in deep learning models. A systematic review of over sixty privacy-preserving NLP methods underscores the urgent need for effective data protection strategies as reliance on private data grows in sectors like healthcare and marketing. Regulations such as the GDPR highlight the importance of safeguarding personal information, necessitating advanced solutions like text sanitization to mask personally identifiable information (PII) while preserving data utility [103, 102].

Trusted methods utilize trusted execution environments and secure enclaves to provide isolated environments for executing sensitive computations, enhancing data security. Verification methods ensure the validity and reliability of privacy-preserving claims made by NLP models, often involving audits and formal verification processes. This includes the development of privacy-preserving mimic

AI-generated, for reference only.

models introduced by Bannour et al. [46], enabling the sharing of Named Entity Recognition (NER) models while maintaining data confidentiality.

The Privacy-Oriented Entity Recognizer (POER) proposed by Papadopoulou et al. [102] exemplifies an innovative approach to privacy preservation by automatically detecting and labeling various types of PII spans in text documents. This enhances NLP systems' ability to manage sensitive data responsibly, ensuring compliance with privacy regulations like the GDPR. Integrating advanced privacy-preserving techniques, including differential privacy and text sanitization, is crucial for the ethical deployment of NLP technologies, particularly in sensitive domains like healthcare, legal, and marketing. This integration safeguards personal information from breaches while enhancing data utility, allowing effective analysis without compromising individual privacy [85, 103, 102]. By employing a combination of data safeguarding, trusted, and verification methods, the NLP community can ensure that its applications protect user privacy while delivering high-quality language processing capabilities.

7.3 Challenges and Limitations

Implementing privacy-preserving techniques in NLP presents challenges and limitations that must be addressed to ensure effective data protection while maintaining model performance. A significant challenge is balancing computational efficiency and privacy-utility trade-offs in real-world applications. Current studies often inadequately address the full spectrum of privacy threats, particularly in scenarios with limited computational resources where efficient privacy-preserving methods are critical [103]. Another limitation is the predominant focus on English-only datasets, which restricts the generalizability of privacy-preserving models to multilingual contexts. The computational expense of training and deploying these models further exacerbates this issue, limiting their applicability across diverse linguistic settings [105]. This highlights the need for developing multilingual privacy-preserving techniques capable of handling data from various languages without compromising performance or privacy.

Moreover, existing privacy-preserving methods often fail to comprehensively address all types of Personally Identifiable Information (PII), including traditional named entities and other personal attributes. This gap poses a significant challenge in ensuring adequate protection of sensitive information, necessitating the development of more robust methods to identify and safeguard a broader range of PII [102]. The challenges and limitations associated with privacy-preserving NLP underscore the urgent need for ongoing research and innovation, particularly in addressing data traceability, computation overhead, dataset size, and the balance between privacy and utility, as identified in recent systematic reviews and studies. These efforts are essential for compliance with regulations like the GDPR and for developing effective methods of text sanitization and differential privacy to safeguard sensitive information across various applications, including legal contexts [102, 82, 103, 85]. By enhancing computational efficiency, expanding multilingual capabilities, and improving PII coverage, the NLP community can develop more effective privacy-preserving solutions that meet real-world application demands.

7.4 Sector-Specific Implications

The implementation of privacy-preserving techniques in NLP has significant implications across sectors where sensitive data handling is critical. In healthcare, protecting patient information is paramount; privacy-preserving NLP methods ensure that sensitive health data is processed securely, complying with regulations like the Health Insurance Portability and Accountability Act (HIPAA). Advanced text mining techniques facilitate the extraction and semantic analysis of clinical data from unstructured sources, such as electronic health records and radiology reports, while implementing robust privacy measures. By leveraging methods like Named Entity Recognition (NER) and Relation Extraction (RE), these approaches enhance healthcare applications' functionality, improving disease classification and knowledge graph construction without compromising patient confidentiality [48, 49, 46, 9].

In the financial sector, privacy-preserving NLP methods enable secure processing of sensitive financial information, such as personal financial records and transaction details. Techniques like differential privacy and secure multi-party computation allow financial institutions to analyze customer data for personalized services without exposing sensitive information to unauthorized access. Balancing data utility and robust privacy protection is essential for fostering customer trust, adhering to stringent financial regulations like the GDPR, and navigating complexities introduced by frameworks like the Markets in Crypto-Assets Regulation (MiCAR). This balance safeguards sensitive information while enhancing compliance through advanced NLP techniques that facilitate regulatory document analysis, addressing privacy threats and supporting the financial sector's evolving needs [8, 103].

The legal industry also benefits from privacy-preserving NLP applications, where confidentiality of legal documents and client information is critical. Techniques that automatically detect and label PII in legal texts, as demonstrated by Papadopoulou et al. [102], significantly enhance the ability to protect sensitive client data while enabling efficient legal research and analysis. These methods show substantial improvement in PII detection compared to existing approaches, effectively balancing privacy protection with data utility.

The sector-specific implications of privacy-preserving NLP underscore the essential role these techniques play in enabling secure and effective NLP technology use across various industries. By implementing advanced privacy-preserving techniques that protect sensitive information while ensuring data utility, these methods foster trust and compliance with regulations like the GDPR. This ultimately facilitates the wider adoption of NLP applications across sectors, including healthcare, legal, and marketing, where sensitive data handling is critical. A systematic review of over sixty deep learning methods categorizes these techniques into data safeguarding, trusted, and verification methods, highlighting the importance of text sanitization in mitigating privacy risks associated with PII. Such efforts are crucial in addressing privacy concerns while maintaining the functional integrity of the data [103, 102].

8 Conclusion

8.1 Future Directions and Research Opportunities

The field of domain adaptation and specialized NLP applications is poised for significant growth, offering numerous avenues for enhancing model performance and expanding their applicability across diverse sectors. In the biomedical domain, there is a pressing need to refine information retrieval systems and integrate sophisticated knowledge graphs to boost the reasoning capabilities of LLMs. Expanding benchmark datasets and exploring a broader range of biomedical NLP tasks will be crucial for developing robust evaluation frameworks and fostering the advancement of more sophisticated models. Leveraging hierarchical knowledge structures can bolster zero-shot and few-shot learning, thereby enhancing model robustness and representation quality.

In healthcare, the integration of ethical AI frameworks and the incorporation of LLMs into clinical workflows are pivotal areas of focus. Assessing the impact of these models on patient outcomes and developing explainable AI systems are essential for addressing ethical concerns associated with machine learning in healthcare settings. Additionally, the fusion of LLMs with emerging technologies such as blockchain and IoT holds potential for enhancing interpretability and ensuring ethical compliance, thereby increasing effectiveness in clinical environments.

The legal sector presents opportunities to refine NLP technologies for legal practice, explore the capabilities of generative models, and foster interdisciplinary collaboration between legal and computational experts. Evaluating models like LEGAL-BERT across a variety of legal datasets and tasks, and investigating their performance in specific legal sub-domains, offers promising directions for future research.

In the realm of terminology extraction, optimizing methods for computing termhood and integrating additional features to improve extraction precision are critical research areas. Analyzing broader corpora for enhanced term alignment can refine the precision of NLP models across different sectors. Furthermore, developing efficient models that minimize environmental impact, safeguard data privacy, and expand the availability of annotated datasets for non-English languages are essential priorities.

Future research should also focus on the effectiveness of temporal adaptation over extended time periods and varying pre-training objectives. Enhancing model training with annotated corpora and expanding benchmarks to include additional domains, such as mathematics, can further augment the effectiveness of NLP applications. Moreover, initiatives like PharmKE should aim to optimize the knowledge extraction process, improve knowledge graph maintenance, and test methodologies across diverse text contexts to ensure robustness and adaptability.

In advancing domain adaptation and specialized NLP applications, researchers should prioritize innovative techniques that enhance model adaptability, semantic understanding, and cross-domain applicability. By targeting these areas, the field can achieve significant progress in both theoretical and practical dimensions, ultimately enhancing the effectiveness and impact of NLP applications across various industries.

References

[1] Eyal Ben-David, Yftah Ziser, and Roi Reichart. Domain adaptation from scratch, 2022.

[2] Minh-Tien Nguyen, Duy-Hung Nguyen, Shahab Sabahi, Hung Le, Jeff Yang, and Hajime Hotta. When giant language brains just aren't enough! domain pizzazz with knowledge sparkle dust, 2023.

[3] Shyni Sharaf and V. S. Anoop. An analysis on large language models in healthcare: A case study of biobert, 2023.

[4] Huizi Yu, Lizhou Fan, Lingyao Li, Jiayan Zhou, Zihui Ma, Lu Xian, Wenyue Hua, Sijia He, Mingyu Jin, Yongfeng Zhang, et al. Large language models in biomedical and health informatics: A bibliometric review. *arXiv preprint arXiv:2403.16303*, 2024.

[5] Boshko Koloski, Nada Lavrač, Bojan Cestnik, Senja Pollak, Blaž Škrlj, and Andrej Kastrin. Aham: Adapt, help, ask, model – harvesting llms for literature mining, 2023.

[6] Zhenrui Yue, Huimin Zeng, Ziyi Kou, Lanyu Shang, and Dong Wang. Contrastive domain adaptation for early misinformation detection: A case study on covid-19, 2022.

[7] Yang Cao, Yangsong Lan, Feiyan Zhai, and Piji Li. 5w1h extraction with large language models, 2024.

[8] Carolina Camassa. Legal nlp meets micar: Advancing the analysis of crypto white papers, 2023.

[9] Qiyuan An, Ruijiang Li, Lin Gu, Hao Zhang, Qingyu Chen, Zhiyong Lu, Fei Wang, and Yingying Zhu. A privacy-preserving unsupervised domain adaptation framework for clinical text analysis, 2022.

[10] Cuong D. Tran, Ognjen Rudovic, and Vladimir Pavlovic. Unsupervised domain adaptation with copula models, 2017.

[11] Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R. Arabnia. A brief review of domain adaptation, 2020.

[12] Zhiyu Zoey Chen, Jing Ma, Xinlu Zhang, Nan Hao, An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian McAuley, Linda Petzold, and William Yang Wang. A survey on large language models for critical societal domains: Finance, healthcare, and law. *arXiv preprint arXiv:2405.01769*, 2024.

[13] Xinlu Zhang, Shiyang Li, Xianjun Yang, Chenxin Tian, Yao Qin, and Linda Ruth Petzold. Enhancing small medical learners with privacy-preserving contextual prompting. *arXiv preprint arXiv:2305.12723*, 2023.

[14] Hasham Ul Haq, Veysel Kocaman, and David Talby. Mining adverse drug reactions from unstructured mediums at scale, 2022.

[15] Vishal Vinod, Susmit Agrawal, Vipul Gaurav, Pallavi R, and Savita Choudhary. Multilingual medical question answering and information retrieval for rural health intelligence access, 2021.

[16] Ashutosh Modi, Prathamesh Kalamkar, Saurabh Karn, Aman Tiwari, Abhinav Joshi, Sai Kiran Tanikella, Shouvik Kumar Guha, Sachin Malhan, and Vivek Raghavan. Semeval 2023 task 6: Legaleval - understanding legal texts, 2023.

[17] Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael J. Bommarito II au2. Natural language processing in the legal domain, 2023.

[18] Benjamin Clavié, Akshita Gheewala, Paul Briton, Marc Alphonsus, Rym Laabiyad, and Francesco Piccoli. Legalmfit: Efficient short legal text classification with lstm language model pre-training, 2021.

[19] Cheng Qian, Xianglong Shi, Shanshan Yao, Yichen Liu, Fengming Zhou, Zishu Zhang, Junaid Akram, Ali Braytee, and Ali Anaissi. Optimized biomedical question-answering services with llm and multi-bert integration, 2024.

[20] Dandan Wang and Shiqing Zhang. Large language models in medical and healthcare fields: applications, advances, and challenges. *Artificial Intelligence Review*, 57(11):299, 2024.

[21] Robert Mahari, Dominik Stammbach, Elliott Ash, and Alex 'Sandy' Pentland. The law and nlp: Bridging disciplinary disconnects, 2023.

[22] Ting He, Kory Kreimeyer, Mimi Najjar, Jonathan Spiker, Maria Fatteh, Valsamo Anagnostou, and Taxiarchis Botsis. Ai-assisted knowledge discovery in biomedical literature to support decision-making in precision oncology, 2024.

[23] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining, 2019.

[24] Jiacheng Hu, Runyuan Bao, Yang Lin, Hanchao Zhang, and Yanlin Xiang. Accurate medical named entity recognition through specialized nlp models, 2024.

[25] Shoetsu Sato, Jin Sakuma, Naoki Yoshinaga, Masashi Toyoda, and Masaru Kitsuregawa. Vocabulary adaptation for distant domain adaptation in neural machine translation, 2020.

[26] Li Fang, Qingyu Chen, Chih-Hsuan Wei, Zhiyong Lu, and Kai Wang. Bioformer: an efficient transformer language model for biomedical text mining, 2023.

[27] Chong Wang, Mengyao Li, Junjun He, Zhongruo Wang, Erfan Darzi, Zan Chen, Jin Ye, Tianbin Li, Yanzhou Su, Jing Ke, et al. A survey for large language models in biomedicine. *arXiv preprint arXiv:2409.00133*, 2024.

[28] Zi-Yi Dou, Antonios Anastasopoulos, and Graham Neubig. Dynamic data selection and weighting for iterative back-translation, 2020.

[29] Vanni Zavarella, Juan Carlos Gamero-Salinas, and Sergio Consoli. A few-shot approach for relation extraction domain adaptation using large language models, 2024.

[30] Philipp Borchert, Jochen De Weerdt, Kristof Coussement, Arno De Caigny, and Marie-Francine Moens. Core: A few-shot company relation classification dataset for robust domain adaptation, 2023.

[31] Tiezheng Yu, Zihan Liu, and Pascale Fung. Adaptsum: Towards low-resource domain adaptation for abstractive summarization, 2021.

[32] Paul Röttger and Janet B. Pierrehumbert. Temporal adaptation of bert and performance on downstream document classification: Insights from social media, 2021.

[33] Paul Grouchy, Shobhit Jain, Michael Liu, Kuhan Wang, Max Tian, Nidhi Arora, Hillary Ngai, Faiza Khan Khattak, Elham Dolatabadi, and Sedef Akinli Kocak. An experimental evaluation of transformer-based language models in the biomedical domain, 2020.

[34] Isabel Segura-Bedmar, David Camino-Perdonas, and Sara Guerrero-Aspizua. Exploring deep learning methods for recognizing rare diseases and their clinical manifestations from texts, 2021.

[35] Dattaraj J. Rao, Shraddha S. Mane, and Mukta A. Paliwal. Biomedical multi-hop question answering using knowledge graph embeddings and language models, 2022.

[36] Joshua J Myszewski, Emily Klossowski, Patrick Meyer, Kristin Bevil, Lisa Klesius, and Kristopher M Schroeder. Validating gan-biobert: A methodology for assessing reporting trends in clinical trials, 2021.

[37] Bridget T. McInnes, Jiawei Tang, Darshini Mahendran, and Mai H. Nguyen. Biobert-based deep learning and merged chemprot-drugprot for enhanced biomedical relation extraction, 2024.

[38] Omid Rohanian, Mohammadmahdi Nouriborji, Samaneh Kouchaki, and David A. Clifton. On the effectiveness of compact biomedical transformers, 2022.

[39] Simon Lupart, Benoit Favre, Vassilina Nikoulina, and Salah Ait-Mokhtar. Zero-shot and few-shot classification of biomedical articles in context of the covid-19 pandemic, 2022.

[40] Fernando Gallego, Guillermo López-García, Luis Gasco-Sánchez, Martin Krallinger, and Francisco J. Veredas. Clinlinker: Medical entity linking of clinical concept mentions in spanish, 2024.

[41] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. Legal-bert: The muppets straight out of law school, 2020.

[42] Josef Valvoda and Ryan Cotterell. Towards explainability in legal outcome prediction models, 2024.

[43] Ivan Habernal, Daniel Faber, Nicola Recchia, Sebastian Bretthauer, Iryna Gurevych, Indra Spiecker genannt Döhmann, and Christoph Burchard. Mining legal arguments in court decisions, 2023.

[44] Ha-Thanh Nguyen. Toward improving attentive neural networks in legal text processing, 2022.

[45] Omid Rohanian, Mohammadmahdi Nouriborji, Hannah Jauncey, Samaneh Kouchaki, ISARIC Clinical Characterisation Group, Lei Clifton, Laura Merson, and David A. Clifton. Lightweight transformers for clinical natural language processing, 2023.

[46] Nesrine Bannour. *Information Extraction from Electronic Health Records: Studies on temporal ordering, privacy and environmental impact.* PhD thesis, Université Paris-Saclay, 2023.

[47] Nasi Jofche, Kostadin Mishev, Riste Stojanov, Milos Jovanovik, and Dimitar Trajanov. Pharmke: Knowledge extraction platform for pharmaceutical texts using transfer learning, 2021.

[48] Daniel Reichenpfader, Henning Müller, and Kerstin Denecke. A scoping review of large language model based approaches for information extraction from radiology reports. *NPJ Digital Medicine*, 7(1):222, 2024.

[49] Hasham Ul Haq, Veysel Kocaman, and David Talby. Deeper clinical document understanding using relation extraction, 2021.

[50] Shaina Raza and Syed Raza Bashir. Leveraging foundation models for clinical text analysis, 2023.

[51] Helia Hashemi, Yong Zhuang, Sachith Sri Ram Kothur, Srivas Prasad, Edgar Meij, and W. Bruce Croft. Dense retrieval adaptation using target domain description, 2023.

[52] Aman Priyanshu and Supriti Vijay. Adaptkeybert: An attention-based approach towards few-shot zero-shot domain adaptation of keybert, 2022.

[53] Sujoy Roychowdhury, Sumit Soman, H. G. Ranjani, Vansh Chhabra, Neeraj Gunda, Shashank Gautam, Subhadip Bandyopadhyay, and Sai Krishna Bala. Towards understanding domain adapted sentence embeddings for document retrieval, 2024.

[54] Saibo Geng, Rémi Lebret, and Karl Aberer. Legal transformer models may not always help, 2021.

[55] Kosuke Nishida, Kyosuke Nishida, and Sen Yoshida. Task-adaptive pre-training of language models with word embedding regularization, 2021.

[56] Suman Dowlagar and Radhika Mamidi. Unsupervised technical domain terms extraction using term extractor, 2021.

[57] Shounak Paul, Arpan Mandal, Pawan Goyal, and Saptarshi Ghosh. Pre-trained language models for the legal domain: A case study on indian law, 2023.

[58] Valentina Aparicio, Daniel Gordon, Sebastian G. Huayamares, and Yuhuai Luo. Biofinbert: Finetuning large language models (llms) to analyze sentiment of press releases and financial text around inflection points of biotech stocks, 2024.

[59] Anastasiia Bezobrazova, Miriam Seghiri, and Constantin Orasan. Benchmarking terminology building capabilities of chatgpt on an english-russian fashion corpus, 2024.

[60] Damith Premasiri, Amal Haddad Haddad, Tharindu Ranasinghe, and Ruslan Mitkov. Transformer-based detection of multiword expressions in flower and plant names, 2022.

[61] Seethalakshmi Gopalakrishnan, Luciana Garbayo, and Wlodek Zadrozny. Causality extraction from medical text using large language models (llms), 2024.

[62] Yanxin Zheng, Wensheng Gan, Zefeng Chen, Zhenlian Qi, Qian Liang, and Philip S Yu. Large language models for medicine: a survey. *International Journal of Machine Learning and Cybernetics*, pages 1–26, 2024.

[63] Chengzhi Zhang and Dan Wu. Bilingual terminology extraction using multi-level termhood, 2013.

[64] Omid Rohanian, Mohammadmahdi Nouriborji, and David A. Clifton. Exploring the effectiveness of instruction tuning in biomedical language processing, 2023.

[65] Jacob Collard, Valeria de Paiva, and Eswaran Subrahmanian. Mathematical entities: Corpora and benchmarks, 2024.

[66] Christophe Servan, Josep Crego, and Jean Senellart. Domain specialization: a post-training domain adaptation for neural machine translation, 2016.

[67] Rong Zhang, Revanth Gangi Reddy, Md Arafat Sultan, Vittorio Castelli, Anthony Ferritto, Radu Florian, Efsun Sarioglu Kayi, Salim Roukos, Avirup Sil, and Todd Ward. Multi-stage pre-training for low-resource domain adaptation, 2020.

[68] Bill Yuchen Lin and Wei Lu. Neural adaptation layers for cross-domain named entity recognition, 2018.

[69] Zhen wan, Yating Zhang, Yexiang Wang, Fei Cheng, and Sadao Kurohashi. Reformulating domain adaptation of large language models as adapt-retrieve-revise: A case study on chinese legal domain, 2024.

[70] Katikapalli Subramanyam Kalyan and Sivanesan Sangeetha. A hybrid approach to measure semantic relatedness in biomedical concepts, 2021.

[71] Jitin Krishnan, Hemant Purohit, and Huzefa Rangwala. Diversity-based generalization for unsupervised text classification under domain shift, 2020.

[72] Mehrzad Shahinmoghadam and Ali Motamedi. Benchmarking pre-trained text embedding models in aligning built asset information, 2024.

[73] Dan Iter and David Grangier. On the complementarity of data selection and fine tuning for domain adaptation, 2021.

[74] Lei Yu. Tackling sequence to sequence mapping problems with neural networks, 2018.

[75] Zhijing Jin, Julius von Kügelgen, Jingwei Ni, Tejas Vaidhya, Ayush Kaushal, Mrinmaya Sachan, and Bernhard Schölkopf. Causal direction of data collection matters: Implications of causal and anticausal learning for nlp, 2021.

[76] Hina Raja, Asim Munawar, Mohammad Delsoz, Mohammad Elahi, Yeganeh Madadi, Amr Hassan, Hashem Abu Serhan, Onur Inam, Luis Hermandez, Sang Tran, Wuqas Munir, Alaa Abd-Alrazaq, Hao Chen, and SiamakYousefi. Using large language models to automate category and trend analysis of scientific articles: An application in ophthalmology, 2023.

[77] Micheal Abaho, Danushka Bollegala, Paula R Williamson, and Susanna Dodd. Assessment of contextualised representations in detecting outcome phrases in clinical trials, 2022.

[78] Prathamesh Kalamkar, Janani Venugopalan Ph. D., and Vivek Raghavan Ph. D. Indian legal nlp benchmarks : A survey, 2021.

[79] Sai Krishna Telukuntla, Aditya Kapri, and Wlodek Zadrozny. Uncc biomedical semantic question answering systems. bioasq: Task-7b, phase-b, 2020.

[80] Israt Jahan. Studying the effectiveness of large language models in benchmark biomedical tasks. 2024.

[81] Josef Valvoda, Alec Thompson, Ryan Cotterell, and Simone Teufel. The ethics of automating legal actors, 2023.

[82] Dimitrios Tsarapatsanis and Nikolaos Aletras. On the ethical limits of natural language processing on legal text, 2021.

[83] Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. Lextreme: A multi-lingual and multi-task benchmark for the legal domain, 2024.

[84] Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, and Saptarshi Ghosh. Legal case document summarization: Extractive and abstractive methods and their evaluation, 2022.

[85] Ying Yin and Ivan Habernal. Privacy-preserving models for legal natural language processing, 2022.

[86] Jonathan Li, Rohan Bhambhoria, Samuel Dahan, and Xiaodan Zhu. Experimenting with legal ai solutions: The case of question-answering for access to justice, 2024.

[87] Ram Mohan Rao Kadiyala, Siddartha Pullakhandam, Kanwal Mehreen, Subhasya Tippareddy, and Ashay Srivastava. Augmenting legal decision support systems with llm-based nli for analyzing social media evidence, 2024.

[88] Ken Voskuil and Suzan Verberne. Improving reference mining in patents with bert, 2021.

[89] Claire Barale, Michael Rovatsos, and Nehal Bhuta. Do language models learn about legal entity types during pretraining?, 2023.

[90] Mahdis Mahdieh, Mia Xu Chen, Yuan Cao, and Orhan Firat. Rapid domain adaptation for machine translation with monolingual data, 2020.

[91] Shwetha Bharadwaj and Melanie Laffin. Automating the compilation of potential core-outcomes for clinical trials, 2021.

[92] Goonmeet Bajaj, Vinh Nguyen, Thilini Wijesiriwardene, Hong Yung Yip, Vishesh Javangula, Srinivasan Parthasarathy, Amit Sheth, and Olivier Bodenreider. Evaluating biomedical bert models for vocabulary alignment at scale in the umls metathesaurus, 2021.

[93] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering, 2019.

[94] David Oniani and Yanshan Wang. A qualitative evaluation of language models on automatic question-answering for covid-19, 2020.

[95] Lingyao Li, Jiayan Zhou, Zhenxiang Gao, Wenyue Hua, Lizhou Fan, Huizi Yu, Loni Hagen, Yongfeng Zhang, Themistocles L Assimes, Libby Hemphill, et al. A scoping review of using large language models (llms) to investigate electronic health records (ehrs). *arXiv preprint arXiv:2405.03066*, 2024.

[96] Anfu Tang, Louise Deléger, Robert Bossy, Pierre Zweigenbaum, and Claire Nédellec. Does constituency analysis enhance domain-specific pre-trained bert models for relation extraction?, 2021.

[97] Parham Abed Azad and Hamid Beigy. Multi-bert: Leveraging adapters and prompt tuning for low-resource multi-domain adaptation, 2024.

[98] Aakanksha Naik, Jill Lehman, and Carolyn Rose. Adapting event extractors to medical data: Bridging the covariate shift, 2020.

[99] Stefan Gerdjikov and Klaus U. Schulz. Corpus analysis without prior linguistic knowledge - unsupervised mining of phrases and subphrase structure, 2016.

[100] Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. When does pretraining help? assessing self-supervised learning for law and the casehold dataset, 2021.

[101] David Oniani, Sonish Sivarajkumar, and Yanshan Wang. Few-shot learning for clinical natural language processing using siamese neural networks, 2022.

[102] Anthi Papadopoulou. Automated text sanitization beyond named entities: Resources, methods, evaluation. 2024.

[103] Samuel Sousa and Roman Kern. How to keep text private? a systematic review of deep learning methods for privacy-preserving natural language processing, 2022.

[104] Wenhao Zhu, Shujian Huang, Tong Pu, Pingxuan Huang, Xu Zhang, Jian Yu, Wei Chen, Yanfeng Wang, and Jiajun Chen. Fgrada: A dataset and benchmark for fine-grained domain adaptation in machine translation, 2021.

[105] Daniel Campos, Alexandre Marques, Tuan Nguyen, Mark Kurtz, and ChengXiang Zhai. Sparse*bert: Sparse models generalize to new tasks and domains, 2023.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.