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 a pivotal methodology in natural language processing (NLP) that addresses challenges posed by out-of-distribution examples, which are common across various applications [1]. This approach enhances the adaptability and generalization of models in diverse fields, including finance, healthcare, and law. In the insurance sector, for instance, incorporating domain-specific knowledge into Large Language Models (LLMs) is essential for tailoring these models to practical business scenarios, thereby enhancing their performance [2].

In healthcare, LLMs are increasingly utilized to analyze Electronic Health Records (EHRs), filling a critical gap in comprehensive data review and application [3]. The transformative potential of LLMs in Biomedical and Health Informatics (BHI) is notable, although it presents ethical and practical challenges that require careful consideration [4]. Additionally, DA is crucial for improving machine translation systems, addressing issues such as language style variations and out-of-vocabulary words to enhance translation quality.

DA is also significant in scientific literature, where the sheer volume of publications demands efficient text analysis. It plays a vital role in managing complexity and improving information retrieval [5]. During critical events like the COVID-19 pandemic, DA is essential for adapting models to new domains, overcoming label and conditional shifts in misinformation detection [6].

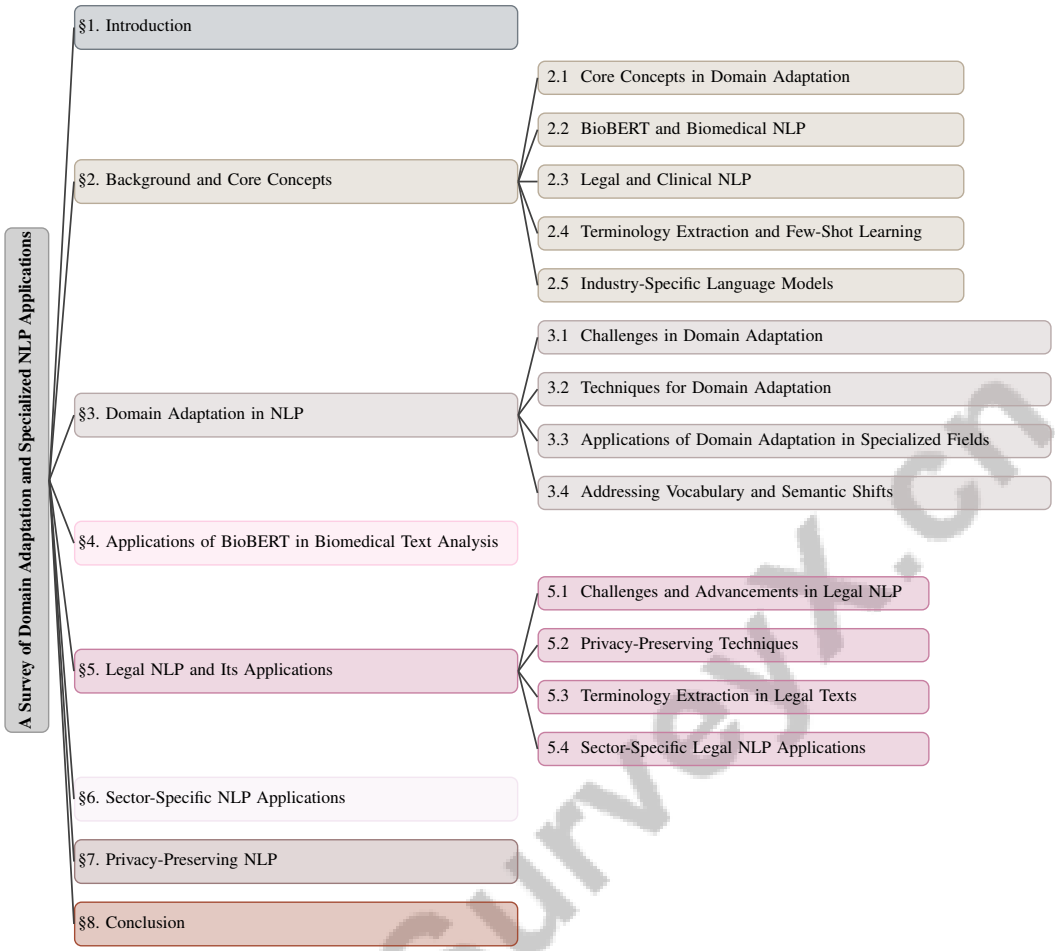


Figure 1: chapter structure

Moreover, DA supports essential news element extraction using frameworks like 5W1H, which are crucial for event extraction and text summarization [7]. In legal NLP applications, DA facilitates the analysis of complex documents, 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 the source domain distribution, effectively addressing distribution shift problems [9]. This is particularly important in scenarios lacking labeled target domain data, enhancing predictive performance across industries [10].

DA is integral to contemporary NLP, effectively tackling the challenge of data distribution mismatch that arises when training and test datasets originate from different contexts. This discrepancy can impair model performance in real-world applications. By aligning features of disparate domains, DA enhances the generalization capabilities of NLP models, enabling improved performance across sectors, including sentiment analysis, machine translation, and quality estimation. This advancement not only addresses specific industry needs but also contributes to solving complex real-world challenges in a privacy-preserving manner [11, 12, 1, 13, 14].

1.2 Importance of Specialized NLP Applications

Specialized Natural Language Processing (NLP) applications are critical for enhancing the efficiency and effectiveness of industry-specific processes by customizing language models to meet the unique demands of different sectors. In finance, the deployment of LLMs has revolutionized decision-making and operational efficiency through the precise processing of vast amounts of financial data [15], which

is 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 [16]. These applications are vital for extracting adverse drug reactions from unstructured text, employing a unified NLP pipeline that integrates document classification, named entity recognition, and relation extraction [17]. Furthermore, the development of multilingual models aims to serve as low-resource medical assistants, providing crucial medical information in indigenous languages and improving healthcare accessibility [18].

In the legal sector, the rising demand for effective NLP techniques to process and understand legal documents is driven by the exponential growth of pending legal cases in populous countries like India [19]. Although the adoption of NLP tools in legal contexts has been historically slow, their potential to address the access to justice crisis by automating and streamlining legal processes is increasingly recognized [20]. Efficient legal NLP models are essential for maintaining performance while reducing resource consumption, thereby enhancing the accessibility and cost-effectiveness of legal services [21]. These models aid in regulatory compliance and improve document clarity, facilitating better investment decisions [8].

The integration of specialized NLP applications across industries not only boosts performance and efficiency but also addresses sector-specific challenges, underscoring the vital role of tailored NLP solutions in advancing industry processes. The adoption of novel parallel processing frameworks leveraging distributed computing further enhances the efficiency of machine learning algorithms, highlighting the ongoing evolution and sophistication of NLP methods [22].

1.3 Impact Across Industries

Domain adaptation and specialized NLP applications significantly impact various industries by addressing the unique challenges and requirements of each sector. In finance, deploying LLMs encounters distinct challenges, including reliance on professional expertise, managing confidential data, and strict regulatory compliance [15]. These factors necessitate the development of tailored NLP models capable of efficiently processing complex financial documents, thereby improving decision-making and operational efficiency.

In healthcare, LLMs are increasingly applied to EHRs, with applications encompassing named entity recognition, information extraction, text summarization, and medical question-answering. These applications enhance the accuracy and reliability of healthcare data analysis, bridging the gap between domain-specific SLMs and LLMs. Furthermore, the use of NLP in healthcare extends to dialogue summarization and EHR generation, contributing to the sector's ability to manage and utilize vast amounts of data effectively [23].

The legal industry also benefits from domain adaptation and specialized NLP applications, particularly in legal research, reasoning, contract review, and document analysis [24]. NLP in the legal domain addresses the unique characteristics of legal texts, often lengthy, unstructured, and containing specialized lexicons [19]. The analysis of crypto asset white papers under regulatory frameworks like the European Union's Markets in Crypto-Assets Regulation (MiCAR) further emphasizes the need for effective NLP methods to navigate the evolving legal landscape [8].

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

1.4 Structure of the Survey

This survey is organized into several key sections, each addressing distinct aspects of domain adaptation and specialized NLP applications. The introductory section outlines the scope and significance of domain adaptation, emphasizing its impact across 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 then provides a comprehensive analysis of domain adaptation in NLP, emphasizing its role in facilitating knowledge transfer across fields. It addresses challenges associated with adapting models to new domains, particularly focusing on techniques like few-shot domain transfer, which leverages limited labeled data to enhance model performance. Additionally, it explores innovative methods such as generating synthetic document collections from brief textual descriptions of target domains and leveraging large language models for schema-constrained data annotation, showcasing their effectiveness in improving relation extraction in specialized domains like architecture and engineering [25, 26]. The applications of BioBERT in biomedical text analysis are scrutinized, focusing on its contributions to clinical text analysis and healthcare data understanding.

Subsequent sections explore NLP's use in processing legal documents, discussing challenges, advancements, and privacy-preserving techniques. The survey also investigates sector-specific NLP applications, providing examples from finance and healthcare. The importance of protecting sensitive information is addressed in the privacy-preserving NLP section, which examines 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 is pivotal in NLP for transferring knowledge from source to target domains, especially when target domain data is limited or unlabeled. This process ensures models adjust to the distinct vocabulary, syntax, and semantics of various domains, often absent in original training datasets [1]. A major challenge is maintaining model performance across domains amid label and conditional shifts [6].

Unsupervised domain adaptation (UDA) aligns source and target domain distributions without labeled target data, employing machine learning techniques to minimize distributional discrepancies and enhance generalization [10]. In neural machine translation (NMT), domain adaptation tailors models for effective translation of domain-specific content, crucial when in-domain parallel corpora are scarce [27].

Adapting Named Entity Recognition (NER) models and event extraction to different domains highlights the limitations of traditional models, necessitating domain-specific strategies [28]. In healthcare, domain adaptation in processing Electronic Health Records (EHRs) exemplifies its importance, where LLMs enhance data analysis accuracy [29]. Developing comprehensive biomedical vocabularies underscores the complexity of integrating diverse terminologies, necessitating models adept at managing domain-specific language [30].

Relation extraction in data-scarce domains, such as the AECO sector, underscores the necessity for robust domain adaptation techniques [26]. Benchmarks like CORE evaluate few-shot relation classification models, emphasizing domain adaptation's role in ensuring model flexibility [31]. The AdaptSum benchmark addresses adapting abstractive summarization models to low-resource domains [32].

In Biomedical and Health Informatics (BHI), domain adaptation is crucial for interdisciplinary research involving clinical decision support, patient interaction, and medical document analysis [4]. Adapting pre-trained language models to temporal and domain variations is critical for improving document classification, as language use varies significantly across domains and evolves over time [33]. Domain adaptation thus provides a framework for NLP models to tackle diverse domain-specific challenges, enhancing knowledge transfer and model robustness [7].

2.2 BioBERT and Biomedical NLP

BioBERT is a domain-specific language representation model pre-trained on extensive biomedical corpora, designed to address the unique challenges of biomedical text mining [34]. It captures specialized terminologies and semantic nuances, excelling in tasks like Named Entity Recognition (NER), relation extraction, and question answering [35]. Its ability to extract and normalize biomedical information is crucial for research and clinical applications.

BioBERT's fine-tuning capabilities for specific challenges, such as recognizing rare diseases, demonstrate its utility [36]. Performance in benchmarks like PubMedQA underscores its effectiveness in biomedical question answering [37]. Integrating BioBERT with models like GAN-BioBERT showcases its application in sentiment classification of clinical trial abstracts [38].

BioBERT's role in relation extraction tasks, such as identifying chemical-gene interactions, highlights its broad applicability [39]. Optimization through knowledge distillation and continual learning has led to lightweight models maintaining performance while reducing computational demands [40]. Its application in zero-shot and few-shot learning frameworks for classifying biomedical articles using MeSH terms demonstrates its versatility in low-resource settings [41].

Advancements in optimizing transformer-based models for biomedical tasks include compact architectures like Bioformer, which reduce parameter counts while retaining performance [40]. BioBERT's application in medical entity linking within Spanish clinical texts highlights the limitations of existing multilingual models, necessitating tailored approaches for enhanced accuracy and efficiency in clinical text analysis [42].

2.3 Legal and Clinical NLP

NLP integration in legal and clinical domains automates complex data processing tasks and enhances decision-making, though it requires specialized solutions. In the legal domain, the complexity of legal language poses significant barriers to scaling NLP solutions. Legal texts, characterized by intricate structures and specialized vocabularies, complicate the adaptation of general NLP models. Despite BERT models' potential to enhance legal NLP tasks, their adaptation is under-explored, necessitating tailored models for legal research and computational law [43].

A key challenge in legal NLP is the lack of explainability in legal outcome prediction models, limiting usability for legal professionals who require transparent reasoning [44]. Additionally, mining, classifying, and analyzing legal arguments, particularly in court decisions, reveals a gap between current NLP methodologies and the nuanced understanding demanded by legal experts [45]. The lengthy and complex nature of legal sentences, along with specialized terminology, complicates the adaptation of generic NLP models [46].

In the clinical domain, NLP is vital for processing EHRs and other medical texts, particularly in resource-limited environments where large pre-trained models may be inefficient [47]. The focus is on extracting structured information from unstructured clinical texts, identifying clinical entities, and understanding temporal relationships [48]. This extraction is essential for improving clinical decision-making and patient outcomes. Predictive analytics in NLP identifies diseases, advancing healthcare diagnostics and treatment planning [3].

Accurate tagging of pharmaceutical organizations and drugs presents another challenge in clinical NLP, exacerbated by variations in naming conventions and a lack of reliable labeled datasets [49]. Addressing these challenges necessitates developing specialized models and datasets accommodating the unique linguistic and structural characteristics of clinical texts.

2.4 Terminology Extraction and Few-Shot Learning

Terminology extraction and few-shot learning are crucial for enhancing domain adaptation, especially in specialized fields characterized by unique vocabularies and limited annotated data. Few-shot learning (FSL) addresses the challenge of training models with minimal data, enabling generalization from a few examples, essential in clinical and biomedical NLP where data scarcity is common [49]. This approach is vital for adapting NLP applications to niche domains, facilitating the extraction and understanding of domain-specific language.

In the biomedical domain, terminology extraction is critical for identifying key concepts from specialized texts, essential for accurate information retrieval and analysis. Techniques like the Clinical Concept Extraction Method (CCEM) emphasize aligning extracted clinical concepts with standardized medical terminologies, improving interoperability and enhancing practical utility in clinical settings. This alignment facilitates accurate identification and semantic relationships among clinical entities, supporting better clinical decision-making and knowledge management [50, 51, 52]. Integrating instance weighting techniques, adversarial domain adaptation, and domain-adaptive fine-tuning further enhances model performance in specialized tasks.

Few-shot learning benchmarks in BioNLP tasks highlight the need for models capable of performing with limited training samples, addressing the intricacies of biomedical language. The AdaptKeyBERT model utilizes regularized attention to enhance keyword extraction in low-resource domains, demonstrating FSL's effectiveness in improving domain adaptation capabilities. Advanced techniques like the AHAM methodology facilitate improved scientific text analysis through tailored topic modeling adaptations, while evaluations of domain-adapted sentence embeddings show significant improvements in document retrieval accuracy within specialized vocabularies. Ongoing advancements in Legal NLP suggest a growing alignment with both methodological sophistication and professional standards observed in broader scientific practices [5, 53, 20].

In the legal domain, terminology extraction is vital for bridging the gap between NLP research and practical applications. Accurate mapping of legal terminologies across classification systems enhances alignment with legal professionals' needs, addressing the scarcity of useful NLP applications in the legal field. Task-adaptive pre-training with word embedding regularization (TAPTER) represents a significant advancement in fine-tuning pre-trained language models (PTLMs) specifically for legal terminology. By aligning static word embeddings with domain-specific meanings from target legal texts, TAPTER enhances models' adaptability and relevance in legal contexts without requiring additional pre-training corpora. This method outperforms standard fine-tuning techniques in legal NLP tasks, particularly where initial pre-training data lacks sufficient domain representation [54, 21, 55, 56].

Terminology extraction and few-shot learning enhance domain adaptation in NLP by enabling the automatic identification of relevant domain-specific terms from text corpora and facilitating efficient training of models on limited labeled data. Specifically, terminology extraction automates identifying key terms through unsupervised methods leveraging statistical measures, while few-shot learning harnesses large language models' capabilities to adapt quickly to new domains with minimal expert input, improving performance in relation extraction and information retrieval tasks across specialized fields [57, 7, 26, 25, 5]. These techniques help models overcome challenges posed by domain-specific language and limited data availability, enhancing the accuracy, relevance, and impact of NLP applications across diverse industries.

2.5 Industry-Specific Language Models

Industry-specific language models represent a significant evolution in NLP, offering customized solutions that effectively address the distinct linguistic features and challenges across various sectors. The AHAM methodology, for instance, utilizes the LLaMa2 generative language model to enhance scientific text analysis through domain-specific adaptation, improving topic modeling in specialized fields like literature-based discovery. Similarly, advancements in NLP within the legal domain reflect increasing methodological sophistication and a growing body of research that collectively enhance the applicability and effectiveness of NLP technologies in sector-specific contexts [5, 20]. These models are tailored to improve the accuracy and relevance of NLP applications by incorporating domain-specific knowledge and datasets, essential for managing specialized vocabulary and contextual nuances.

In the legal domain, pre-trained language models (PLMs) developed specifically for Indian legal tasks demonstrate substantial performance improvements over existing models. These PLMs leverage extensive legal corpora for fine-tuning, addressing the complexities of legal language and facilitating more effective legal text processing [58]. Such targeted models are crucial for automating legal research, contract analysis, and case law retrieval, ultimately enhancing the efficiency and accessibility of legal services.

The financial sector also benefits from industry-specific language models, exemplified by BioFinBERT, a fine-tuned large language model designed to analyze and predict sentiment in financial texts related to biotech stocks. By focusing on financial narratives, BioFinBERT provides valuable insights into market trends and investor sentiment, aiding financial analysts and decision-makers in navigating complex financial data [59].

In the fashion industry, creating a diverse and balanced dataset from English and Russian fashion magazines supports developing terminology-building capabilities within language models. This dataset enables models to capture the dynamic and culturally nuanced language of fashion, facilitating improved information retrieval and analysis in fashion-related NLP tasks [60].

Additionally, detecting multiword expressions (MWEs) in specialized datasets illustrates the potential of transformer-based models to address challenges in understanding complex linguistic structures within specific domains. Tailoring transformer models to detect MWEs enhances the precision of NLP applications across various fields, including those with highly specialized vocabularies [61].

The development and implementation of industry-specific language models underscore the critical need for tailored NLP solutions that address the unique challenges and requirements of various sectors. Methodologies like AHAM demonstrate the effectiveness of adapting large language models (LLMs) for scientific literature mining by leveraging domain expertise to refine topic modeling. Advancements in legal NLP reveal a trend toward increased methodological sophistication, aligning with broader scientific standards, while empirical studies in insurance highlight the necessity of incorporating domain-specific knowledge to enhance reasoning capabilities. These examples underscore the importance of customizing NLP tools to improve precision and relevance across diverse applications [2, 7, 5, 20]. Such models not only enhance NLP application performance but also drive innovation by enabling more accurate and contextually aware language processing in specialized domains.

In recent years, the field of Natural Language Processing (NLP) has witnessed significant advancements driven by domain adaptation techniques. These techniques are crucial for enhancing the performance of NLP models across various specialized fields, such as biomedical, finance, and legal domains. As illustrated in Figure 2, the hierarchical structure of domain adaptation encompasses several critical components, including challenges, techniques, applications, and strategies aimed at addressing vocabulary and semantic shifts.

The challenges section elucidates key issues such as data discrepancies, domain complexity, and ethical concerns that researchers must navigate. To tackle these challenges, various techniques have emerged, notably deep learning and cross-lingual methods, which serve to improve model robustness. Furthermore, the applications of these techniques in specialized fields underscore the necessity of tailored approaches to meet domain-specific requirements. Finally, the strategies for addressing shifts in vocabulary and semantics emphasize the importance of effective semantic representation and the utilization of synthetic data, which are essential for enhancing the adaptability of NLP systems across diverse contexts.

3 Domain Adaptation in NLP

3.1 Challenges in Domain Adaptation

Domain adaptation in NLP is fraught with challenges due to data distribution discrepancies, limited annotated datasets, and the intricate nature of domain-specific languages. The scarcity of task-specific datasets, such as those for causality extraction from Clinical Practice Guidelines, impedes model development [62]. Financial constraints of domain-specific pre-training further limit the scalability of specialized models [35].

In legal NLP, computational methods often oversimplify legal arguments, ignoring the complex structures and terminology vital for thorough analysis [45]. Benchmarks frequently overlook temporal language shifts, affecting the performance of models trained on outdated data [33]. The biomedical field faces the challenge of adapting to dynamic medical data and evolving knowledge, which traditional systems struggle to manage [63]. The complexity of medical nomenclature, characterized by multiple synonyms and modifiers, complicates adaptation efforts [36]. Integrating LLMs into healthcare systems raises ethical concerns, necessitating rigorous validation and alignment with healthcare standards [4].

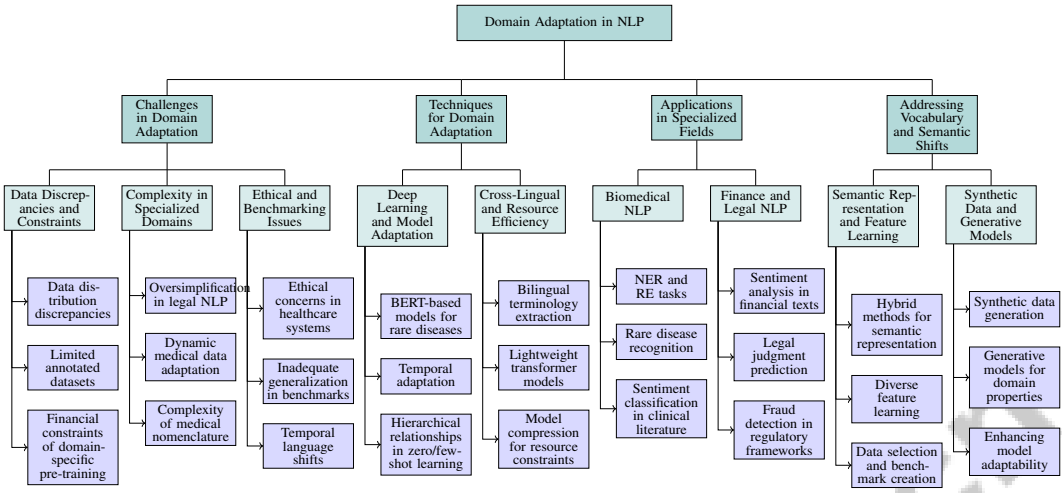


Figure 2: This figure illustrates the hierarchical structure of domain adaptation in NLP, highlighting the challenges, techniques, applications, and strategies for addressing vocabulary and semantic shifts. The challenges section outlines data discrepancies, domain complexity, and ethical concerns. Techniques include deep learning and cross-lingual methods. Applications in specialized fields focus on biomedical, finance, and legal NLP. Strategies for addressing shifts emphasize semantic representation and synthetic data utilization.

Bilingual terminology extraction is hindered by disparate corpus distributions that current methods inadequately address [64]. Adapting general-purpose LLMs for specialized biomedical tasks involves navigating complex instructions, a gap that current benchmarks often overlook [65].

In clinical NLP, the complexity of clinical language and ethical considerations surrounding sensitive patient data restrict the availability of annotated datasets essential for effective domain adaptation [48]. The labor-intensive nature of manual labeling further complicates data acquisition [49]. Existing benchmarks inadequately generalize to mathematical domains, complicating the processing of mathematical texts [66].

Innovative methodologies, such as the AHAM framework, can enhance domain-specific adaptation through collaboration with domain experts and advanced generative models like LLaMa2. This approach not only improves the adaptability of NLP models across various domains but also aids in extracting insights from complex datasets, contributing to methodological sophistication and reproducibility in the scientific community [5, 20].

As illustrated in Figure 3, the primary challenges in domain adaptation within NLP can be categorized into three main areas: data discrepancies, domain complexities, and technical limitations. Each category elucidates specific issues such as distribution discrepancies, legal complexities, and bilingual extraction difficulties, thereby providing a comprehensive overview of the multifaceted obstacles researchers face in this field.

3.2 Techniques for Domain Adaptation

Domain adaptation in NLP employs diverse techniques to facilitate knowledge transfer from a source to a target domain, particularly in contexts with limited labeled data and domain-specific language features. Deep learning models, especially those based on BERT, have shown efficacy in recognizing rare diseases from unstructured text through domain adaptation [36]. These models leverage pre-trained language representations to adapt to specialized vocabulary and semantic nuances.

Temporal adaptation, when combined with domain adaptation, aligns models with shifts in language usage, significantly influencing performance [33]. This underscores the importance of considering temporal factors to maintain model relevance.

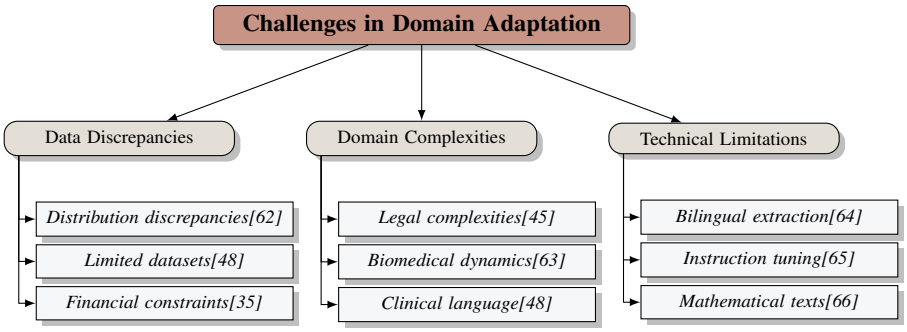


Figure 3: This figure illustrates the primary challenges in domain adaptation within NLP, categorized into data discrepancies, domain complexities, and technical limitations. Each category highlights specific issues such as distribution discrepancies, legal complexities, and bilingual extraction difficulties.

In zero-shot and few-shot learning contexts, leveraging hierarchical relationships encoded in MeSH descriptors has proven effective for classification, especially in biomedical settings with data scarcity [41]. This method enhances model adaptability by utilizing structured knowledge.

Bilingual terminology extraction techniques, such as BTE-MLT, enhance adaptation by calculating termhood through corpus comparison [64]. This is crucial for applications requiring cross-lingual capabilities.

Adapting lightweight transformer models for domain-specific tasks addresses the high computational and memory demands of existing models, making them more accessible for researchers [40]. By compressing large models into efficient versions, these techniques facilitate NLP solutions in resource-constrained settings.

The diverse techniques employed in domain adaptation highlight the need for tailoring NLP models to the linguistic and semantic characteristics of different domains. Incorporating advanced methodologies like deep learning, temporal adaptation, hierarchical knowledge leveraging, and model compression enhances the robustness and versatility of NLP models. These methods enable effective responses to domain-specific challenges, improving retrieval accuracy and isotropy in sentence embeddings, allowing models to adapt incrementally to new domains without complete retraining, and facilitating performance in sensitive areas through innovative learning setups [1, 14, 11, 53].

3.3 Applications of Domain Adaptation in Specialized Fields

Method Name	Task Type	Specialized Fields	Model Integration
BioBERT[34]	Ner, Re, QA	Biomedical	Biobert
DL-RDR[36]	Named Entity Recognition	Rare Disease Recognition	Biobert And Clinicalbert
BioBERT-Merged[39]	Relation Extraction	Biomedical Texts	Biobert Model
GBB[38]	Sentiment Classification	Clinical Trial Abstracts	Gan-BioBERT Algorithm
BF[59]	Sentiment Analysis	Biotech Sector	Biofinbert

Table 1: Overview of various domain adaptation methods applied in specialized fields, detailing their task types, specialized fields of application, and model integration strategies. The table highlights the diversity in approaches and models, such as BioBERT and GAN-BioBERT, utilized for tasks ranging from named entity recognition to sentiment analysis within biomedical and financial domains.

Domain adaptation significantly enhances NLP applications in specialized fields by addressing unique linguistic challenges and data scarcity. Table 1 provides a comprehensive overview of domain adaptation methods applied in specialized fields, showcasing the integration of various models and their applications in tasks like named entity recognition and sentiment analysis. In biomedical NLP, techniques are crucial for tasks like NER and RE, essential for identifying complex interactions within medical texts. The application of BioBERT exemplifies this, effectively recognizing biomedical entities and relationships, thereby improving performance in text mining tasks [34]. Instruction-tuned models have also shown substantial gains in biomedical NER and RE tasks, reflecting the efficacy of domain adaptation in refining model capabilities [65].

In rare disease recognition, domain adaptation using BERT-based models has improved diagnostic accuracy, showcasing the potential of these techniques to enhance outcomes in specialized healthcare fields [36]. The integration of BioBERT with deep learning approaches highlights successful applications of domain adaptation in merging diverse datasets for improved biomedical NLP tasks [39]. Additionally, the GAN-BioBERT algorithm has demonstrated effectiveness in sentiment classification within clinical trial literature, outperforming previous methodologies [38].

In finance, domain adaptation is exemplified by BioFinBERT, a model fine-tuned for sentiment analysis in financial texts related to biotech companies, critical during clinical and regulatory events where accurate sentiment analysis informs strategic decisions [59].

The legal domain also benefits from domain adaptation, with models like LegalRoBERTa pre-trained on legal corpora to navigate the complexities of legal texts. This approach employs adapters to enhance efficiency, facilitating tasks such as legal judgment prediction and cross-lingual transfer learning [56]. Moreover, domain adaptation aids in extracting insights from legal documents, supporting investor decision-making and fraud detection through the analysis of complex regulatory frameworks [8].

3.4 Addressing Vocabulary and Semantic Shifts

Addressing vocabulary and semantic shifts in domain adaptation is crucial for a model's ability to generalize across domains. Hybrid methods that enhance the semantic representation of domain-specific concepts have proven effective, particularly in the biomedical domain, where combining contextual embeddings with structured knowledge from ontologies improves the representation of biomedical concepts [67].

Advancements in aligning domain distributions to tackle vocabulary and semantic shifts have emerged through techniques that utilize domain adaptation and multi-task learning. These methods extract relevant features across diverse domains and languages using domain-specific adaptation, unsupervised term extraction, and statistical significance testing for term burstiness. This approach mitigates challenges posed by vocabulary variations and semantic differences, ensuring accurate topic modeling and terminology identification in scientific literature [57, 51, 5, 68]. By leveraging shared features, these techniques facilitate effective knowledge transfer between domains.

Promoting diverse feature learning is another promising strategy to address domain-specific challenges. Encouraging models to learn various features reduces reliance on domain-specific characteristics, enhancing generalization to new domains [69].

Data selection is vital for improving domain adaptation effectiveness. Creating benchmarks that emphasize careful data selection significantly impacts model performance, as aligning data with the target domain's vocabulary and semantic structures enhances the model's ability to navigate domain shifts. Research indicates that domain adaptation improves retrieval accuracy and tightens confidence intervals in performance metrics. Techniques such as fine-tuning and synthetic data generation optimize model performance across diverse domains, leading to better generalization and accuracy in information retrieval tasks [12, 26, 53, 25, 5].

Leveraging generative models to create synthetic data that captures target domain properties presents a promising avenue. This enables retrieval models to learn from generated data reflecting the target domain's characteristics, enhancing adaptability to new vocabulary and semantic contexts [25].

Future research could focus on enhancing model performance with less labeled data and exploring applications across other NLP tasks [70]. Additionally, understanding the impact of causal direction on NLP model performance may inform further research and applications [71]. The primary challenge in domain adaptation remains the extensive retraining required by existing methods, often rendering them impractical in dynamic translation environments [11].

Addressing vocabulary and semantic shifts necessitates a multifaceted approach combining advanced representation techniques, diverse feature learning, strategic data selection, and synthetic data utilization. These methodologies collectively enhance the adaptability and robustness of NLP models, enabling effective navigation of complexities associated with domain adaptation. The AHAM methodology exemplifies this by customizing topic modeling frameworks for scientific analysis through domain expertise and generative language models, demonstrating that fine-tuning and domain adaptation substantially improve retrieval accuracy and isotropy [5, 53].

4 Applications of BioBERT in Biomedical Text Analysis

4.1 BioBERT in Biomedical Text Mining and Question Answering

BioBERT is pivotal in biomedical text mining and question answering, leveraging extensive pre-training on biomedical corpora to adeptly handle complex medical terminologies and relationships. It excels in Named Entity Recognition (NER), Relation Extraction (RE), and Question Answering (QA), effectively integrating domain-specific linguistic features [34]. In biomedical QA, BioBERT significantly improves accuracy on datasets like PubMedQA, showcasing its utility [39].

The model's integration with frameworks that combine classification and generation enhances its capability to process biomedical information, excelling in tasks such as extracting chemical-gene relationships [39]. Its proficiency in multi-hop QA is supported by datasets designed for knowledge graph QA systems, highlighting its effectiveness in complex information retrieval [34].

BioBERT's role extends to literature-based discovery frameworks, demonstrating robustness in extracting insights from vast biomedical texts. Its zero-shot performance enhancements through domain-specific task methods solidify its standing in biomedical NLP [4]. The efficiency of Bioformer models, achieving high accuracy with fewer parameters, underscores BioBERT's potential to optimize processing times and service delivery [33]. BioBERT provides a robust framework for extracting and interpreting complex biomedical information.

4.2 Applications in Clinical Text Analysis

BioBERT is crucial in clinical text analysis, utilizing pre-training on extensive biomedical literature to enhance processing of complex clinical data [34]. It significantly improves clinical text classification and NER, extracting insights from unstructured clinical data. Models like Clinical BioBERT achieve high performance on datasets such as EBM-COMET, surpassing previous benchmarks [72].

BioBERT's integration into frameworks like PharmKE illustrates its role in comprehensive entity recognition and knowledge extraction, employing multi-stage deep learning to enhance recognition in clinical settings [49]. Its adaptability is highlighted by its application in medical entity normalization, as shown by systems like ClinLinker, using advanced pipelines for candidate retrieval and re-ranking [42].

In sentiment analysis, BioBERT is vital, exemplified by the GAN-BioBERT algorithm, which classifies clinical trial abstracts into sentiments, enhancing healthcare data analysis granularity [38]. It also contributes to causality extraction from medical guidelines, supported by specific benchmarks validating its potential [62].

BioBERT enhances recognition of rare diseases, improving healthcare data understanding and decision-making [36]. Its effectiveness in classifying biomedical articles, especially during the COVID-19 pandemic, highlights its versatility in low-resource settings [41]. Furthermore, BioBERT optimizes treatment strategies based on patient data, advancing healthcare informatics [3]. Its integration into clinical text analysis underscores its significant contributions to healthcare, supporting the evolution of biomedical research [39].

4.3 Enhancements and Future Directions

Enhancements to BioBERT in biomedical text analysis can be pursued through refining training procedures and model architectures to optimize performance across a broader range of tasks and datasets [40]. Expanding datasets and integrating advanced language models could enhance BioBERT's adaptability and effectiveness [65].

Future research should integrate multi-modal data sources—such as genomic, clinical, and environmental data—to enhance extraction of complex relationships, bolstering BioBERT's applications in disease prediction and patient management. This integration aims to leverage BioBERT's capabilities in biomedical text mining, including improved NER and RE, facilitating effective clinical decision support and knowledge discovery [34, 3, 73]. Advanced entity recognition techniques could further enhance BioBERT's ability to identify complex biomedical entities and relationships.

Expanding test sets and experimenting with various LLMs for data generation could improve BioBERT's adaptability and performance in diverse tasks [37]. Incorporating new datasets or

refining models to better handle incomplete knowledge graphs may enhance BioBERT's utility in biomedical QA [37].

In sentiment analysis, expanding datasets with more expert raters and exploring finer-grained sentiment classification could enhance healthcare data analysis granularity [38]. Further tuning of models like BioFinBERT and exploring additional datasets could improve predictive capabilities in financial texts related to biotech companies [59].

Integrating conversational contexts into relation extraction frameworks and developing metrics that capture human dialogue nuances represent additional enhancement areas. Such improvements could bolster BioBERT's utility in clinical decision support systems, particularly in patient-provider interactions. Enhancing the model's robustness against ambiguous terminologies and exploring applications beyond its current scope, such as e-commerce, could broaden BioBERT's applicability [64].

These enhancements and future research directions are poised to advance BioBERT's applications in biomedical text analysis significantly. By leveraging its superior performance in NER and RE, particularly in precision oncology, researchers can extract critical insights from the growing body of biomedical literature. This ongoing development will ensure BioBERT remains a vital tool for addressing complex biomedical challenges, facilitating informed decision-making in healthcare [34, 73]. Focusing on refining methodologies and exploring new application avenues will help maintain BioBERT's position as a leading model in biomedical NLP.

5 Legal NLP and Its Applications

Exploring the implications and potential of Natural Language Processing (NLP) within the legal domain reveals both challenges and advancements that characterize this evolving field. Legal NLP must navigate the complexities of legal language and meet the operational needs of legal professionals who rely on precise interpretations of legal texts. The following subsection examines the specific obstacles faced by practitioners and researchers in legal NLP and the significant strides made to address these challenges, ultimately facilitating more effective applications in the legal landscape.

5.1 Challenges and Advancements in Legal NLP

Legal NLP is shaped by the intricacies of legal language and the specific needs of legal professionals, presenting unique challenges and opportunities for advancement. A primary challenge is replicating the nuanced understanding and moral reasoning required for judicial decision-making, which current NLP models struggle with [74]. Limited datasets further exacerbate this issue, leading to overfitting and poor generalization to real-world legal texts [46]. Existing methods often fail to analyze specialized biomedical and financial language in biotech press releases, resulting in inaccurate sentiment predictions [59].

The scarcity of comprehensive datasets, particularly for rare argument types, negatively impacts model performance [45]. This highlights the need for more extensive and diverse datasets to enhance model training and evaluation. Ethical considerations also pose challenges, requiring a nuanced understanding of NLP technologies' capabilities and limitations [75].

Privacy concerns and the demand for robust model efficiency further complicate legal NLP system development. Many studies inadequately address data privacy issues and the need for clinical validation, which are essential for deploying NLP technologies in sensitive legal contexts [63]. Integrating ethical considerations into legal NLP systems underscores the importance of academic freedom and interdisciplinary collaboration in developing responsible technologies [75].

Despite these challenges, significant advancements have emerged. The creation of specialized benchmarks and models like LEGAL-BERT has improved accuracy over baseline models, demonstrating the potential of domain-specific adaptations [43]. Multilingual multitask benchmarks such as LEXTREME enhance fairness and comparability among models, improving legal NLP system evaluations across diverse languages and datasets [76]. Frameworks like PharmKE, with modular design and result visualization, offer advantages in user understanding and model integration [49].

Ongoing research and development continue to drive progress in legal NLP. Addressing issues related to dataset diversity, ethical considerations, and model efficiency can unlock NLP technologies'

full potential in the legal sector. This approach promises to enhance legal NLP methodologies' sophistication, aligning them with broader scientific standards and addressing the access to justice crisis by improving legal services' effectiveness and accessibility. Bridging the gap between legal practitioners' needs and NLP researchers' focus will facilitate the development of tools better aligned with real-world legal applications, fostering greater NLP integration in legal practice [77, 45, 24, 20].

5.2 Privacy-Preserving Techniques

Privacy-preserving techniques are crucial in legal NLP applications, where sensitive data is often involved. Integrating differential privacy into transformer models' pre-training enhances model performance while safeguarding data privacy, ensuring sensitive legal information remains uncompromised during the learning process [78]. Differential privacy provides a robust framework for maintaining data confidentiality, especially in unsupervised domain adaptation scenarios where source data protection is crucial [9].

Evaluating AI-generated legal responses' factual correctness is essential, as accuracy in legal contexts is paramount [79]. This requires metrics that assess performance while maintaining privacy throughout the process. Integrating external knowledge sources can enhance language model responses' accuracy and relevance in specialized legal applications, improving legal NLP systems' overall reliability [2].

Privacy-preserving techniques also augment legal decision support systems, designed to assist legal professionals while emphasizing caution to prevent biases and uphold ethical considerations [80]. Balancing academic freedom and privacy emerges as a recurring theme in legal NLP applications' development, underscoring the importance of ethical considerations in these technologies' design and deployment [75].

Adopting privacy-preserving techniques is essential for maintaining sensitive data confidentiality while enhancing language models' performance and applicability in legal contexts. By integrating differential privacy and advanced methodologies, legal NLP systems can effectively balance high performance with robust privacy protections. This ensures legal professionals can confidently use these tools to analyze sensitive legal texts without compromising ethical standards, addressing the critical need for effective legal insight while adhering to diverse legal and ethical norms. Such advancements enhance predictive capabilities and align with broader goals of improving access to justice and meeting the legal community's specific needs [75, 24, 78, 20].

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 essential for developing NLP models capable of accurately navigating legal texts' complex structures and specialized vocabularies [46]. Extracting legal terminology enables accurate identification of legal entities, such as court names, dates, and statutes, critical for tasks like legal document classification and summarization [45].

Advanced methodologies, such as parameter-efficient legal domain adaptation, utilize unsupervised legal data for pre-training, allowing models to adapt effectively with minimal parameter tuning. This approach is valuable in legal contexts where privacy and data sensitivity are paramount, enabling secure processing of sensitive legal information [75]. Annotated datasets from publicly available legal corpora, including Patent Litigations and the Caselaw Access Project, support legal terminology extraction by providing models with the necessary context to understand and process legal documents accurately [46].

Developing new annotation schemes for legal arguments, particularly those aligned with legal theory and practice for the European Court of Human Rights (ECHR), emphasizes context-aware terminology extraction's significance in legal applications [45]. These schemes enable models to learn legal documents' specific language and context, enhancing their capacity to perform complex legal NLP tasks.

Terminology extraction in legal texts improves legal NLP systems' accuracy and efficiency, facilitating advanced applications like legal document summarization and argument mining. By addressing legal language's complexities and unique characteristics, researchers can enhance legal analysis

and decision-making quality [77, 45, 20]. Utilizing annotated datasets and advanced methodologies enhances legal terminology extraction's accuracy and reliability, ultimately contributing to more effective and accessible legal services.

5.4 Sector-Specific Legal NLP Applications

Sector-specific NLP applications within the legal domain are increasingly tailored to meet various legal contexts' unique requirements, improving legal processes' efficiency and precision. Developing specialized benchmarks, such as those for Indian legal systems, underscores the need to address jurisdiction-specific linguistic and legal complexities, providing structured frameworks for evaluating legal NLP systems [45]. These benchmarks facilitate creating models sensitive to specific legal systems' nuances, enhancing NLP solutions' applicability and effectiveness in these contexts [46].

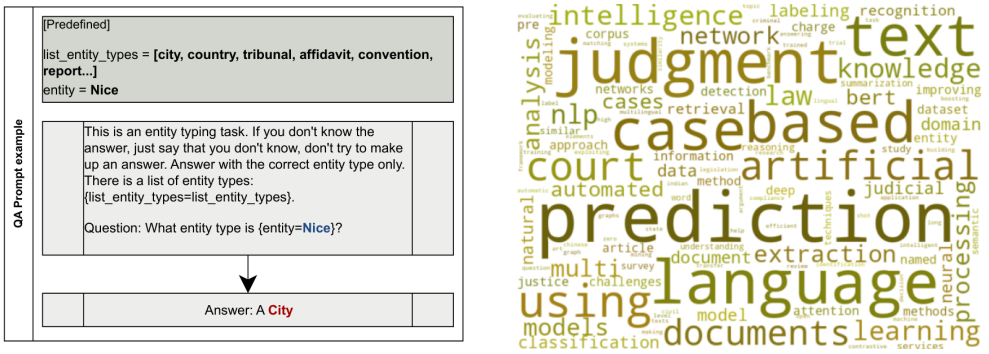
In legal entity recognition, precise entity typing is crucial for downstream tasks like Named Entity Recognition (NER) and question answering, enhancing legal NLP applications' performance, especially in specialized sectors where accurate legal entity identification is critical for compliance monitoring and contract analysis. Integrating privacy-preserving techniques, such as differential privacy in BERT models' pre-training, enhances downstream performance while safeguarding sensitive legal data, ensuring confidentiality in legal proceedings [75].

Developing expert-validated benchmarks for legal question-answering introduces 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 [44]. Creating datasets from public legal texts highlights the importance of utilizing real-world legal documents to enhance legal NLP models' relevance and applicability [46].

Simplifying legal texts using unsupervised methods illustrates NLP's potential to make legal language more accessible while maintaining semantic coherence. This approach is particularly beneficial in sectors where simplifying complex legal documents can improve understanding for non-experts [74]. However, this method has limitations in handling complex multi-label classification tasks, indicating its suitability may vary across different legal NLP applications [46].

Future research in sector-specific legal NLP applications should explore additional data augmentation techniques and investigate cross-jurisdictional transfer to enhance model adaptability and robustness [75]. Adapting methods to other legal domains and languages, alongside integrating additional data sources, could further improve model robustness [44]. Expanding benchmarks to include multilingual datasets and exploring additional legal generative tasks is also a promising direction for future research [45].

Sector-specific legal NLP applications are advancing through developing specialized benchmarks, privacy-preserving techniques, and utilizing real-world legal datasets. These advancements enhance legal processes' accuracy, reliability, and accessibility across various sectors, ultimately contributing to more efficient and effective legal services. Future research should continue exploring hybrid systems that leverage expert knowledge and data-driven insights, as well as emerging trends and evaluation metrics in legal NLP [74].



(a) Entity Typing Task with City as the Answer[81] (b) Word Cloud of Legal and Linguistic Terms[24]

Figure 4: Examples of Sector-Specific Legal NLP Applications

AI-generated, for reference only.

As illustrated in Figure 4, sector-specific applications in Legal NLP are gaining traction, providing tailored solutions to complex legal challenges. The first example, "Entity Typing Task with City as the Answer," presents a flowchart designed to categorize entities within legal texts, posing the question, "What entity type is entity=Nice?" and identifying "City" as the correct answer. This flowchart is part of a broader system that classifies entities such as cities, countries, tribunals, affidavits, conventions, and reports, showcasing the precision required in legal contexts. The second example, a "Word Cloud of Legal and Linguistic Terms," visually represents a collection of relevant terms, arranging words like "intelligence," "judgment," "text," and "law" in a dynamic manner. Together, these examples underscore the potential of Legal NLP to enhance the efficiency and accuracy of legal processes by leveraging technology to interpret and organize complex legal information [81, 24].

6 Sector-Specific NLP Applications

6.1 Frameworks and Taxonomies for Sector-Specific NLP

Frameworks and taxonomies are essential for structuring sector-specific NLP applications, enabling the customization of language models to address unique industry challenges. Integrating unsupervised Neural Machine Translation (NMT) with continual learning allows scalable model adaptation across domains, ensuring relevance in dynamic environments [82]. Developing sector-specific taxonomies involves systematically identifying and organizing industry-specific linguistic characteristics and terminologies. Automated terminology extraction techniques, utilizing statistical methods, ensure taxonomies accurately reflect domain-specific nuances, aiding in tasks like sentiment analysis and risk assessment in finance, and legal research and contract analysis in law [57, 5, 20].

Frameworks leveraging distributed computing and parallel processing enhance NLP applications' efficiency, particularly in sectors demanding large-scale data processing. These frameworks support deploying advanced machine learning algorithms tailored to computational demands, as seen in legal analysis and scientific literature mining [54, 5, 20]. Robust frameworks and taxonomies are crucial for advancing sector-specific NLP, enhancing tasks like machine translation, document summarization, and information retrieval while maintaining professional standards of data availability and reproducibility [57, 20].

6.2 Biomedical and Healthcare NLP Applications

NLP is transforming biomedical and healthcare sectors by enhancing diagnostic capabilities and patient care through advanced language models [4]. Models like BioBERT capture clinical terminology intricacies, crucial for tasks like mining adverse drug reactions, where NLP pipelines integrate document classification, named entity recognition, and relation extraction [83, 17]. Benchmarks for synonymy prediction and datasets like PubMedQA drive the development of sophisticated biomedical NLP applications [84, 85].

NLP applications, such as chatbots, effectively disseminate COVID-19 information, demonstrating potential in public health communication [86]. Comparisons of tools like ChatGPT and SketchEngine highlight NLP's role in developing comprehensive biomedical vocabularies [60]. NLP supports medical research by leveraging large language models for analyzing complex biomedical literature, aiding evidence-based decision-making and precision oncology [73, 87, 50, 51, 52].

6.3 Legal and Financial NLP Applications

In legal and financial industries, NLP enhances text analysis and decision-making. Legal NLP automates processing complex texts, aiding tasks like legal research and contract analysis. The LegalEval benchmark emphasizes tailored datasets' importance for effective legal NLP [19]. In finance, NLP processes vast data for decision-making and risk assessment, with domain-specific models improving sentiment analysis accuracy, crucial for predicting market trends [2, 81, 54, 88, 59].

Recent studies highlight legal NLP's potential in automating industry-specific processes, addressing legal language complexities, and improving access to justice [77, 80, 46, 20, 24]. Specialized datasets and tailored models enhance text analysis accuracy, contributing to more effective legal and financial services.

6.4 NLP in Emerging Domains

NLP's application in emerging domains, such as law and e-commerce, enhances access to justice and customer service, respectively. Analyses reveal a rise in NLP's relevance in legal practice, though challenges remain in aligning researchers' priorities with practitioners' needs [24, 20]. In environmental science, NLP analyzes climate data, informing policy decisions and strategies. The AHAM framework improves topic modeling precision, facilitating actionable environmental insights [5, 20].

In education, NLP enhances curriculum development and tutoring systems, adapting content to individual needs [57, 5, 20]. In mental health, NLP analyzes text-based communications to identify mental well-being indicators, aiding early detection and intervention [50, 52]. NLP's transformative potential across industries drives innovation and enhances process effectiveness [58, 57, 5, 20].

6.5 Challenges and Innovations in Sector-Specific NLP

Sector-specific NLP applications face challenges due to diverse linguistic demands and specialized vocabularies. Adapting models to data-scarce domains remains a challenge, necessitating innovative approaches like few-shot learning [89, 90]. Task-adaptive pre-training (TAPTER) effectively adapts pre-trained models to specific domains, reducing computational costs [55].

Detecting rare phrases is crucial for processing specialized texts, with future research needed to improve methods for detecting these elements [91]. Structured benchmarks, like those in the mathematical domain, assess performance and drive innovation [66]. Challenges in legal and scientific research highlight the need for ongoing innovation, with models like BERT showing promise in improving reference extraction [92, 24, 5, 20]. Advanced techniques and robust benchmarks can overcome limitations, enhancing NLP technologies' effectiveness across industries.

7 Privacy-Preserving NLP

In Natural Language Processing (NLP), addressing privacy concerns is crucial, necessitating an understanding of ethical implications and data protection laws such as the General Data Protection Regulation (GDPR). Privacy-preserving methodologies are essential, as demonstrated by a systematic review of over sixty deep learning techniques [75, 93]. Ethical boundaries in computational law research must be considered, given the diverse legal norms and potential moral implications. This awareness is vital for developing robust, ethically grounded, and privacy-conscious NLP practices. As Large Language Models (LLMs) become more prominent, understanding data usage and ethical practices is critical, fostering discussions on privacy-preserving techniques in NLP.

7.1 Ethical Considerations and Data Privacy

Ethical considerations and data privacy are integral to NLP, especially with the widespread use of LLMs. Chen et al. [15] highlight the importance of transparency, fairness, and accountability in AI, aligning systems with societal values. Data privacy is a pressing concern, particularly under regulations like the GDPR [94]. Integrating privacy-preserving techniques ensures compliance while maintaining AI system performance. Sousa et al. [93] emphasize ongoing research to adapt to evolving privacy regulations and threats.

Privacy-preserving unsupervised domain adaptation (UDA) methods improve privacy protection without major performance loss, aligning with existing UDA frameworks [9]. The FGraDA benchmark by Zhu et al. [95] aids in evaluating privacy measures in machine translation. In multilingual medical question answering, Vinod et al. [18] underscore the importance of user privacy, highlighting the need for multilingual capabilities in privacy-conscious NLP.

Addressing ethical and data privacy challenges in NLP requires safeguarding individual rights and adhering to regulatory frameworks like the GDPR while fostering innovation. This balance is vital in the legal domain, where NLP tools analyzing legal texts raise ethical questions about academic freedom, legal norms diversity, and automating judicial functions. Categorizing privacy-preserving methodologies mitigates data breach risks and navigates the privacy-utility tradeoff effectively [75, 93, 74]. By integrating advanced privacy-preserving techniques and ethical standards, the NLP community can ensure effective and responsible technologies.

7.2 Techniques for Privacy Preservation

Privacy preservation in NLP aims to protect sensitive information while maintaining language model functionality. Sousa et al. [93] classify privacy-preserving methods into data safeguarding, trusted, and verification methods, providing a framework for understanding diverse approaches.

Data safeguarding methods protect NLP model data using techniques like encryption, differential privacy, and secure multi-party computation, preventing unauthorized access and ensuring confidentiality. Trusted methods use trusted execution environments and secure enclaves to isolate sensitive computations, enhancing data security and regulatory compliance [93, 16, 94, 8].

Verification methods validate privacy-preserving claims through audits and formal processes. This includes privacy-preserving mimic models by Bannour et al. [48], enabling Named Entity Recognition (NER) model sharing while maintaining data confidentiality. The Privacy-Oriented Entity Recognizer (POER) by Papadopoulou et al. [94] exemplifies an innovative approach, automatically detecting and labeling personally identifiable information (PII) in text documents.

Integrating advanced privacy-preserving techniques is crucial for responsible NLP technology deployment. Combining data safeguarding, trusted methodologies, and verification techniques enhances user privacy while maintaining robust language processing capabilities. This approach aligns with GDPR and addresses privacy challenges in deep learning methods for privacy-preserving NLP, such as text sanitization and differential privacy in model training, ensuring high data utility [94, 8, 78, 20, 93].

7.3 Challenges and Limitations

Implementing privacy-preserving techniques in NLP faces challenges in balancing computational efficiency with privacy-utility trade-offs. Current studies often inadequately address privacy threats, especially with limited computational resources where efficient methods are critical [93].

A limitation is the focus on English-only datasets, constraining model generalizability to multilingual contexts. The computational costs of training and deploying these models necessitate developing multilingual privacy-preserving techniques that handle data from various languages without sacrificing performance or privacy [96].

Existing methods often fail to comprehensively address all PII types, including traditional named entities and other personal attributes, posing a significant challenge to ensuring sensitive information protection [94].

These challenges highlight the need for ongoing research and innovation in data traceability, computational overhead, and balancing privacy with utility, alongside ethical implications in sensitive domains like law. Recent advancements, including a taxonomy for classifying privacy-preserving methods and applying differential privacy in transformer models, emphasize robust frameworks to safeguard personal information while preserving data utility [94, 75, 78, 20, 93]. Addressing computational efficiency, expanding multilingual capabilities, and enhancing PII coverage will enable more effective privacy-preserving solutions for real-world applications.

7.4 Sector-Specific Implications

Privacy-preserving techniques in NLP have significant implications across sectors, particularly where sensitive data handling is critical. In healthcare, safeguarding patient information is essential due to the sensitive nature of health data. Privacy-preserving NLP methods securely process this data, ensuring compliance with regulations like HIPAA. Recent advancements in LLMs have improved capabilities for named entity recognition and information extraction from Electronic Health Records (EHRs), enabling secure patient information handling while addressing ethical considerations [87, 48].

In finance, privacy-preserving NLP methods securely process sensitive financial information, including personal 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, balancing data utility and privacy protection to foster customer trust and comply with regulations like the GDPR and MiCAR [9, 94, 8, 78, 93].

The legal industry benefits from privacy-preserving NLP applications, where confidentiality of legal documents and client information is paramount. Techniques that automatically detect and label PII

in legal texts, as demonstrated by Papadopoulou et al. [94], enhance client data protection while enabling efficient legal research and analysis, effectively balancing privacy protection with data utility.

Sector-specific implications of privacy-preserving NLP underscore the importance of securely managing sensitive data across industries like healthcare, legal, and marketing. With regulations like the GDPR, adopting privacy-preserving methods, from data safeguarding to verification techniques, is essential for mitigating privacy threats while enhancing NLP application utility. This approach prevents unauthorized PII disclosures and enables organizations to leverage NLP technologies effectively while maintaining compliance [93, 94, 78]. By ensuring sensitive information protection while maintaining data utility, these methods foster trust and compliance, supporting broader NLP application adoption in sectors handling sensitive data.

8 Conclusion

8.1 Future Directions and Research Opportunities

Advancements in domain adaptation and specialized NLP applications promise to significantly enhance model performance and applicability across diverse sectors. In the biomedical field, improving information retrieval and integrating advanced knowledge graphs are crucial for enhancing LLM reasoning capabilities. Expanding benchmark datasets and exploring additional biomedical NLP tasks will bolster evaluation frameworks, fostering the development of sophisticated models. Leveraging hierarchical knowledge can strengthen zero-shot and few-shot learning, thereby enhancing model robustness and representation quality in domain adaptation.

In healthcare, the integration of ethical AI frameworks and LLMs into clinical workflows is vital, with a focus on their impact on patient outcomes. The development of explainable AI models is essential to address ethical considerations in healthcare machine learning deployments. Furthermore, combining LLMs with emerging technologies like blockchain and IoT can improve interpretability and ensure ethical compliance, enhancing their effectiveness in clinical settings.

The legal sector presents opportunities for refining NLP technologies for legal practice, exploring generative models, and fostering interdisciplinary collaboration between legal and computational experts. Evaluating models like LEGAL-BERT across various legal datasets and tasks, and examining their performance in specific legal sub-domains, offers promising research avenues.

In terminology extraction, optimizing termhood computation methods and integrating additional features to enhance extraction precision are critical. Exploring broader corpora for improved term alignment can further refine NLP model precision across sectors. Additionally, developing efficient models that minimize environmental impact, enhance data privacy, and increase annotated dataset availability for non-English languages are essential priorities.

Future research should also explore the effectiveness of temporal adaptation over extended periods and with varied pre-training objectives. Enhancing model training with annotated corpora and expanding benchmarks to include additional domains, such as mathematics, can further improve NLP application effectiveness.

Moreover, efforts should focus on optimizing platforms like PharmKE for knowledge extraction, improving knowledge graph maintenance, and testing methodologies across diverse text contexts to ensure robustness and adaptability.

Research in domain adaptation and specialized NLP applications should prioritize innovative techniques that enhance model adaptability, semantic understanding, and cross-domain applicability. By focusing on these areas, researchers can drive advancements in both theoretical and practical aspects of NLP technology, 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] Christophe Servan, Josep Crego, and Jean Senellart. Domain specialization: a post-training domain adaptation for neural machine translation, 2016.

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

[13] Javad Pourmostafa Roshan Sharami, Dimitar Shterionov, Frédéric Blain, Eva Vanmassenhove, Mirella De Sisto, Chris Emmery, and Pieter Spronck. Tailoring domain adaptation for machine translation quality estimation, 2023.

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

[15] 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.

[16] 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.

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

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

[19] 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.

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

[21] 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.

[22] 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.

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

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

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

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

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

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

[29] 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.

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

[31] 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.

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

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

[34] 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.

[35] 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.

[36] 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.

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

[38] 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.

[39] 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.

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

[41] 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.

[42] 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.

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

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

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

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

[47] 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.

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

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

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

[51] 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.

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

[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] 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.

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

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

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

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

[59] 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.

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

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

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

[63] 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.

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

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

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

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

[68] Samuel Sarria Hurtado, Todd Mullen, Taku Onodera, and Paul Sheridan. A statistical significance testing approach for measuring term burstiness with applications to domain-specific terminology extraction, 2024.

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

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

[71] 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.

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

[73] 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.

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

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

[76] 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.

[77] 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.

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

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

[80] 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.

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

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

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

[84] 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.

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

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

[87] 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.

[88] Marina Sedinkina, Nikolas Breitkopf, and Hinrich Schütze. Automatic domain adaptation outperforms manual domain adaptation for predicting financial outcomes, 2020.

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

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

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

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

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

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

[95] 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.

[96] 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.