Journal Title
XX(X):1–28
©The Author(s) 2025
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Explainability in Practice: A Survey of Explainable NLP Across Various Domains

Anonymous

Abstract

Natural Language Processing (NLP) has become a cornerstone in many critical sectors, including healthcare, finance, Customer Relationship Management, etc. This is particularly true with the development and use of advanced models like GPT-4o, Gemini, and BERT, which are now widely used for decision-making processes. However, the black-box nature of these advanced NLP models has created an urgent need for transparency and explainability. This review provides an exploration of **explainable NLP (XNLP)** with a focus on its practical deployment and real-world applications, examining how these can be applied and what the challenges are in domain-specific contexts. The paper underscores the importance of explainability in NLP and provides a comprehensive perspective on how XNLP can be designed to meet the unique demands of various sectors, from healthcare's need for clear insights to finance's focus on fraud detection and risk assessment. Additionally, the review aims to bridge the knowledge gap in XNLP literature by offering a domain-specific exploration and discussing underrepresented areas such as real-world applicability, metric evaluation, and the role of human interaction in model evaluation. The paper concludes by suggesting future research directions that could lead to a better understanding and broader application of XNLP.

Keywords

Natural Language Processing (NLP), Explainable Natural Language Processing (XNLP), Transparency, Interpretability, Ethical AI

Introduction

Natural Language Processing (NLP) and, more recently, Large Language Models (LLMs) have transformed machine-human interaction by enabling systems to process and generate human language more effectively. Although newer models like OpenAI's GPT-4o and Google's Gemini have pushed the boundaries of language understanding, slightly older architectures such as BERT (1) continue to influence modern NLP pipelines. Indeed, these models have found applications across diverse domains, including healthcare, finance, and Customer Relationship Management (CRM) (2), leading to reduced processing times and enhanced automation (3; 4). For instance, a study by (5) showed that using Clinical Practice Guidelines (CPGs) in conjunction with LLMs can produce more precise and contextually relevant treatment recommendations, thus improving clinical decision support (CDS) (6). By streamlining such processes, NLP systems offer significant benefits, including rapid analysis, user-friendly interfaces, and the ability to handle substantial amounts of data efficiently (7).

Despite these advances, most high-performing NLP models operate as "black boxes." The underlying challenge is not merely the large number of parameters in models like GPT-4o or Gemini, but the lack of transparent, human-interpretable decision pathways in neural architectures (8). In simpler models such as linear regression, it is relatively easier to track how each input feature contributes to a prediction. By contrast, deep neural networks capture complex, non-linear relationships that are difficult to decode, whether they contain millions or billions of parameters. Moreover, these networks are frequently trained on massive datasets where historical or societal biases may be embedded (9; 10), creating a risk of perpetuating discrimination in downstream predictions (11). In practice, such biases have been detected in various sectors, including hiring algorithms (12), medical diagnostics (13), and financial services (14). When left unexamined, these biases can lead to adverse outcomes, such as unfair treatment of job applicants, unequal access to credit, or suboptimal healthcare recommendations.

Explainable AI (XAI) initiatives provide a collection of methods and tools to make these "black-box" processes more transparent. Within XAI, XNLP specifically addresses the interpretability of language-based models, focusing on features like word embeddings, attention mechanisms, and textual rationales. As illustrated in Figure 1, a typical XNLP pipeline begins with an input text, processes it through a model, and then employs an explanation layer (e.g., highlighting key tokens or visualizing attention weights) to clarify how the final output is derived. XNLP aims to tackle distinct linguistic challenges, such as contextual details, synonyms, or domain-specific jargon, that do not always arise in other modalities like images or tabular data (15; 16).

Corresponding author:

Email:

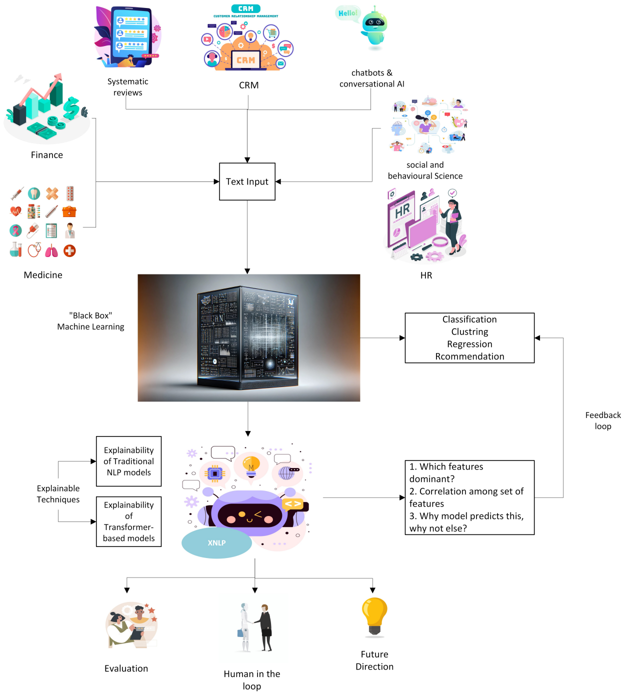

Systematic reviews
CRM
Hello!
chatbots & conversational AI
Finance
Text Input
social and behavioural Science
Medicine
HR
"Black Box" Machine Learning
Classification
Clustring
Regression
Rcommendation
Feedback loop
Explainability of Traditional NLP models
Explainable Techniques
Explainability of Transformer-based models
XNLP
1. Which features dominant?
2. Correlation among set of features
3. Why model predicts this, why not else?
Evaluation
Human in the loop
Future Direction

Figure 1. Pipeline for XNLP. The input text is processed by a model (e.g., a neural network), and an explanation mechanism highlights important tokens or latent representations. This helps stakeholders, such as clinicians, financial analysts, or end-users, interpret the model's decision-making.

Existing literature on XNLP methods, including Local Interpretable Model-Agnostic Explanations (LIME) (3), SHapley Additive exPlanations (SHAP) (17; 18), and attention-based explanations (19), has made considerable progress in unpacking model decisions. Still, much remains to be explored, especially in domain-specific contexts. In healthcare, for instance, NLP systems must provide clinically relevant insights that integrate seamlessly into physicians' workflows (20). This necessity extends to mental health applications, where language models may assist in monitoring depressive or suicidal ideation through social media posts or electronic health records (EHRs) (21). In finance, explanations must address both the complexity of specialized terminology and the high stakes of compliance and fraud detection (22). Similarly, chatbots and conversational AI systems raise new challenges regarding user trust, especially if the system's responses are critical for customer support or emergency services.

Although general XAI frameworks (23–25) provide broad guidelines, there is a gap in how these methods directly translate into *domain-tailored* XNLP solutions. Recent surveys such as (20; 26; 27) underscore the importance of interpretable NLP but tend to focus on the technical aspects rather than the nuances of real-world deployment. In response, this survey offers the following main contributions:

- **Domain-Specific Analysis:** We compile and examine research across various domains (healthcare, finance, CRM, etc.), including mental health under the medical umbrella, to highlight how XNLP can be adapted for different regulatory and practical requirements.
- **Evaluation Techniques and Metrics:** We examine the full range of evaluation approaches for XNLP from both quantitative and qualitative perspectives. For quantitative assessment, we introduce mathematical equations for key metrics such as fidelity, providing a more rigorous understanding of model performance. On the qualitative side, we discuss metrics like user trust, highlighting methods that capture human-centered insights into model explanations.
- **Critical Challenges and Trade-offs:** We address open questions regarding bias, privacy, data availability, and the balance between performance and interpretability.
- **Future Directions:** Building on the limitations found in existing literature, we propose potential research avenues, such as personalized explanations, human-in-the-loop evaluations, and mechanistic interpretability for LLMs.

The remainder of the paper is organized as follows: section Modeling Techniques for XNLP reviews foundational NLP techniques and transitions to modern transformer-based methods and LLMs, highlighting their explainability mechanisms. Section Applications and Domains of XNLP explores the application of XNLP methods in diverse fields, including *Medicine*, where we also delve into mental health applications, *Finance*, *Systematic Reviews*, *CRM*, *Chatbots and Conversational AI*, *Social and Behavioral Science*, *Human Resources (HR)*, and other emerging use cases. Section Critical Aspects of XNLP addresses the critical aspects of XNLP, including evaluation metrics, trade-offs, rationalization techniques, human evaluation, and data/code availability. Section Future Directions and Research Opportunities in XNLP outlines promising directions for further investigation, such as personalized and mechanistic explanations. Finally, section Conclusion concludes with a summary of the survey's key insights.

Modeling Techniques for XNLP

Explainability of Traditional NLP Models

Traditional NLP models, notably Bag of Words (BoW) and its variants such as term frequency (TF) and term frequency-inverse document frequency (TF-IDF), provide foundational approaches to textual representation. A BoW model encodes a document as a set of token counts without preserving word order or contextual usage. When coupled with transparent classifiers (e.g., logistic regression), the model's coefficients help uncover each word's influence on the prediction outcome. For instance, if a logistic regression classifier assigns a large positive coefficient to the token "excellent" in a sentiment analysis task, it signals a strong correlation between that token and a positive sentiment (28; 29).

TF-IDF further refines this representation by assigning greater importance to words that appear frequently in a document but are relatively rare across the entire corpus (30; 31). While these methods are simple and often interpretable, they struggle to encode contextual and syntactic relationships (31–33). Moreover, in TF and TF-IDF systems, interpretability can still be unintuitive: if multiple words frequently co-occur and jointly predict an outcome (e.g., disease codes in discharge letters), each word might receive a small coefficient, even smaller than another word that is less important but does not co-occur as often. Although early in conception, these classical representations paved the way for more complex models that better capture semantic relationships. The drive to incorporate contextual meaning has led researchers toward advanced embedding techniques that combine high performance with more transparent decision processes.

Explainability of Embedding Models

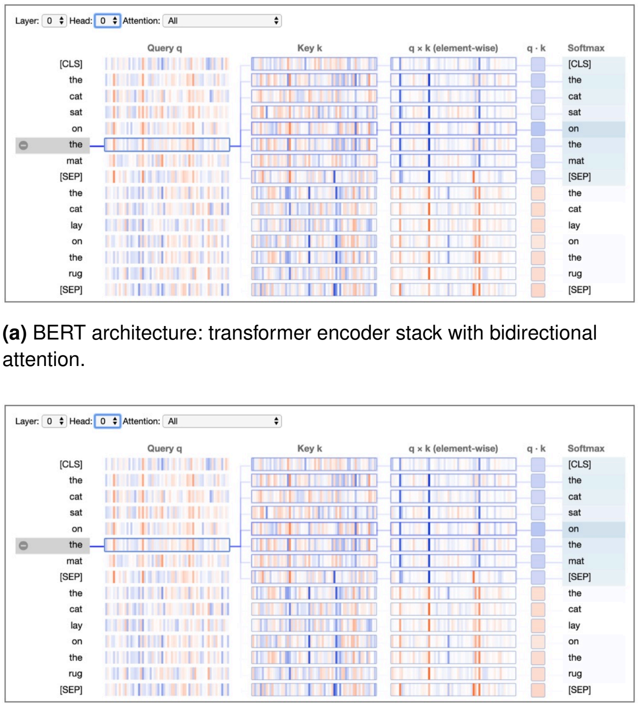
Embedding models revolutionized NLP by mapping words, phrases, or sentences into continuous vector spaces. In contrast to BoW-based approaches, these dense representations capture subtler semantic and syntactic details (34). Word2Vec (35) and GloVe (36), for example,produce vectors where semantically similar words (e.g., “king” and “queen”) reside close together, a property earlier BoW variants could not achieve (37). Sentence-level embeddings (e.g., Universal Sentence Encoder (38)) extend this idea by encoding entire clauses or paragraphs into fixed-dimensional vectors. Despite performance gains, these embedding models add complexity that can obscure their decision-making. Accordingly, researchers employ several explainability strategies:

- **Visualization of Vector Spaces**: Tools like *TensorBoard* (39) and *EmbeddingVis* (40) map high-dimensional embeddings into 2D or 3D layouts, enabling users to visually inspect semantic clusters and language structures. Dimensionality reduction methods such as t-SNE (41) and PCA (42) are commonly employed to reveal meaningful relationships among words or sentences (43; 44).

- **Gradient-Based Methods**: Saliency maps, adapted from computer vision (45), highlight input tokens that produce the largest gradient magnitudes with respect to an embedding layer. By tracing which tokens trigger the strongest change in output, these methods provide local explanations for model predictions.

- **Attention Mechanisms**: Some embedding-based architectures incorporate attention layers, granting insight into how much “focus” the model places on particular words or sub-phrases (46; 47). Unlike gradient-based methods, attention is computed during the forward pass, inherently supporting interpretability.

Balancing interpretability and performance remains challenging as models grow increasingly complex (48). The notion of “explanation” is also context-dependent. Visualizations like PCA or t-SNE may suffice for intuitive overviews, particularly in less technical scenarios (49; 50). However, in high-stakes domains such as medicine or finance, stakeholders often demand deeper, more granular rationales behind each prediction. Consequently, embedding-based NLP poses an ongoing tension: how to develop representations that remain robust and powerful while offering sufficiently transparent insights into their internal structures.



(a) BERT architecture: transformer encoder stack with bidirectional attention.

(b) GPT-2 architecture: transformer decoder stack with causal (unidirectional) attention.

Figure 2. Illustrative comparison of transformer-based model structures and their role in explainability. (a) BERT uses a stack of transformer encoders with bidirectional attention, enabling context-aware explanations for each token. (b) GPT-2 employs a stack of transformer decoders with causal attention, focusing on left-to-right context for generative tasks. Visualizations such as attention maps help interpret which input tokens are most influential in the model’s predictions (55).

Explainability of Transformer-Based Models

Transformers have reshaped modern NLP through *self-attention* mechanisms that handle long-range dependencies without recurrent networks (51). A classic example is BERT (1), which, by using multi-head self-attention, can be pre-trained on massive corpora and then fine-tuned for tasks like sentiment analysis, question answering, or named entity recognition (52). Its successors, such as RoBERTa (53) and ALBERT (54), continue this trend, delivering improvements in various NLP benchmarks.

Although Transformer architectures achieve strong results on NLP tasks, they present unique interpretability challenges. A widely adopted strategy is **attention-weight visualization**, as illustrated in Figure 2. Tools like BERTViz enable multiscale inspection of attention patterns across layers and heads, helping researchers and practitioners understand which input tokens the model attends to when generating its predictions (55). This form of visualization facilitates the interpretation of token interactions and can reveal model reasoning for specific linguistic phenomena. Other notable methods for unpacking transformer-based decisions include:

- **Probing Tasks**: Simplified linguistic evaluations designed to reveal what grammatical or semantic properties the model retains (56).

- **Feature Visualization**: Techniques that visualize learned features or activation patterns, aiding our understanding of which elements are activated by particular words or phrases (57).

- **Attribution Methods**: Integrated Gradients (58) and similar tools offer token-level importance scores by accumulating gradients over input perturbations.

- **Model Simplification & Distillation**: Training smaller, more interpretable student networks to replicate large model outputs, thereby bridging the gap between high performance and clarity (59).

Local interpretation approaches, such as LIME (60), and global methods, including SHAP (61), are also frequently integrated with transformers. Perturbation-based techniques (62) systematically alter input text to identify critical words or phrases, while **contextual decomposition** (63) breaks down output scores into contributions from individual tokens or token interactions. Despite the proliferation of these tools, no single best practice exists for all contexts. Moreover, attention-weight visualization can sometimes be misleading: in sentiment analysis, weights could highlight neutral tokens (e.g., "the"), downplaying salient words like "hate" (64; 65). As a result, any claim of interpretability demands careful validation against the model's actual decision-making mechanisms and the end-user's needs.

Transformer models like BERT (1) and GPT-x (66) have led to major advances in NLP, but their large size and complex inner workings make it hard to understand how they work. Critical Aspects of XNLP section discusses many of these explainability challenges in practice. We also expect that future XNLP methods will include better built-in ways to make their decisions easier to understand.

LLMs and Mechanistic Interpretability

Building on transformer foundations, recent LLMs like Claude 3, GPT-4o, and Gemini have driven major advances in interpretability research, often called *mechanistic* or *feature-level* interpretability. Unlike post-hoc methods that explain model outputs after the fact, mechanistic interpretability seeks to reverse-engineer the internal computations of neural networks, identifying the specific circuits and features responsible for particular behaviors (67).

Sparse Autoencoders and Feature Disentanglement A central challenge in understanding LLMs is *polysemanticity*: individual neurons often encode multiple unrelated concepts, making interpretation difficult (68). **Sparse Autoencoders (SAEs)** address this by learning to decompose neural activations into sparse, interpretable features. The key insight is that while individual neurons may be polysemantic, the underlying representations can be disentangled into monosemantic directions in activation space.

Anthropic's landmark work on Claude 3 Sonnet (67) demonstrated that SAEs can extract millions of interpretable features from LLM activations. Their analysis revealed that approximately 70% of extracted features corresponded to human-interpretable concepts, ranging from abstract notions like "sycophancy" and "deception" to concrete entities like "Golden Gate Bridge." Importantly, these features are not merely correlational; intervening on specific features (e.g., amplifying the "Golden Gate Bridge" feature) systematically altered model outputs in predictable ways.

Recent architectural improvements have enhanced SAE effectiveness. **BatchTopK SAEs** (69) improve training efficiency by applying top-k sparsity across batches rather than individual samples, achieving better reconstruction fidelity with fewer active features. Similarly, work on *transcoders* (67) extends SAEs to capture transformations between layers, providing richer insight into information flow within the model.

Limitations and Critiques Despite promising results, mechanistic interpretability faces significant challenges. Recent critiques (70) have shown that SAEs may fail to reliably detect concepts they purportedly encode, raising questions about whether extracted features genuinely represent causal mechanisms or merely correlate with interpretable concepts. The scalability of these methods to frontier models with hundreds of billions of parameters remains uncertain, and the computational cost of training SAEs on large models is substantial.

Parallel techniques complement SAE-based approaches. *B-cosification* (71) retrofits transformers with constrained linear layers so that each weight contributes a *signed, additive explanation* of the output, advocating for *explainability-by-design*. *Explanation-distillation* methods (72) train smaller student models to replicate both task outputs *and* rationales of LLM teachers, enabling compact models faithful to original reasoning.

The Chain-of-Thought Faithfulness Problem A critical question for LLM explainability is whether **Chain-of-Thought (CoT)** reasoning (73) provides faithful explanations of model behavior. CoT prompting elicits step-by-step reasoning from LLMs, seemingly showing their reasoning process. However, emerging evidence suggests that CoT outputs may not accurately reflect the model's actual computational process—a phenomenon termed the *faithfulness problem* (74).

(75) demonstrated through intervention studies that LLMs frequently produce plausible-sounding reasoning that does not correspond to their true decision-making mechanisms. They introduced biasing features into prompts and found that models often gave correct answers while providing reasoning chains that ignored the biasing information entirely—an "illusion of transparency." More concerningly, (76) found that larger models exhibited *less* faithful reasoning than smaller ones, suggesting that increased capability may worsen rather than improve the faithfulness problem.

These findings have important implications for XNLP. If CoT explanations do not faithfully represent model reasoning, they cannot serve as reliable explanations for high-stakes decisions. Several approaches have been proposed to address this challenge:

- **Faithful CoT Frameworks:** (77) propose a two-stage approach where the LLM first translates the problem into a formal representation, then a deterministic solver produces the final answer. This architectural separation ensures that the reasoning chain is necessarily faithful to the computation.

- **Intervention-Based Validation:** Systematically perturbing CoT steps and measuring the impact on outputs can distinguish faithful from unfaithful reasoning (75).
- **Mechanistic Grounding:** Combining CoT analysis with mechanistic interpretability (e.g., tracing which internal features are active during reasoning) may provide stronger faithfulness guarantees (67).

The CoT faithfulness problem underscores a broader tension in XNLP: the difference between *plausible* explanations that satisfy users and *faithful* explanations that accurately represent model behavior. As LLMs increasingly generate their own explanations, validating faithfulness becomes paramount for trustworthy AI deployment.

Toward Explainability-by-Design This body of research points to a shift toward deeper transparency in LLMs. Instead of relying solely on post-hoc methods like attention or gradient-based saliency, researchers increasingly advocate embedding interpretability directly into model architectures and training. As LLMs grow in complexity and capability, mechanistic interpretability and faithful reasoning validation become crucial for keeping these systems accountable, trustworthy, and aligned with human values (78; 79).

Applications and Domains of XNLP

AI-based applications have been extensively adopted across multiple facets of modern life, including social media, medicine, commerce, customer service, and finance. Common tasks like machine translation, text summarization, and sentiment analysis aim to automate routine processes and improve user experiences (47; 80). In these generic contexts, performance often takes precedence over the transparency of how a model arrives at its conclusions. However, in more sensitive fields such as *medicine*, or in high-stakes areas like *finance*, achieving explainability is critical. For instance, clinicians analyzing EHRs must trust not only the accuracy but also the reasoning behind a model's predictions, given that such insights directly influence patient treatment plans. Unlike traditional AI systems that can provide only “Yes” or “No” type answers, XNLP frameworks offer “Wh-questions” (e.g., “Why,” “When,” and “Where”), enabling richer rationales for model outputs. These rationales must be clear and actionable so that healthcare providers, financial analysts, and other key stakeholders can have confidence in the system.

Beyond interpretability, **fairness** and **equity** rank highly in domains where decisions can drastically impact individuals or communities (81). In medicine, for example, demonstrating how a patient is flagged for a specific intervention is crucial to ensuring unbiased healthcare. Similarly, in CRM, AI-driven personalization should not discriminate based on sensitive attributes. As shown by (82), balancing fairness and explainability can promote greater trust in AI systems. This aligns with the principle that if logical and scientifically grounded arguments reinforce current knowledge, users are more inclined to trust an AI's conclusions (83).

XNLP also benefits human-AI collaboration. When an AI exceeds human capability in certain tasks, such as strategic game play, e.g., AlphaGo (84), transparent explanations allow individuals to glean novel strategies or insights. Conversely, if an AI achieves performance comparable to a human expert, interpretability reinforces end-user trust in the system (85). In finance, for example, a model predicting firm valuations must provide transparent justifications so that shareholders can audit its outputs. Absent transparency, the model could be manipulated to favor particular interests. Similarly, chatbots in customer service must not only respond to user queries but also justify those responses to cultivate user trust and satisfaction.

Table 1 offers an overview of some of the key application domains where XNLP has proven to be a game-changer. These domains range from *medicine* and *finance* to *systematic reviews*, *CRM*, *chatbots*, and *social and behavioral science*, each with distinct subcategories and challenges. HR is also included as a growing domain where XNLP can augment processes like recruitment and performance evaluation.

Medicine

XNLP has become particularly valuable in the medical domain, where decisions can be life-altering. From analyzing patient histories in EHRs to processing physician notes for disease risk assessments, explainability is key because it clarifies the logic behind a model's prediction. For example, (86) devised an interpretable recurrent neural network (RNN) model for heart failure prediction, explicitly highlighting which medical codes contributed most to the risk factors. Similar rule-based approaches have been used to extract clinical evidence from randomized controlled trials (90; 91), ensuring transparency in processes that directly influence patient care.

Recent initiatives also extend to mental health, a critical area of medicine that relies on sensitive textual data, often collected from social media platforms or patient narratives. For instance, LLM-based analyses have been proposed to detect early signs of depression or suicidal ideation, using model explanations to reassure clinicians and researchers about the validity of identified risk factors (122).

These complementary explanation methods can be illustrated through heart failure prediction. The RETAIN model (86), trained on 14 million visits from 263,000 patients, achieves an AUC of 0.8717 while remaining fully interpretable through its two-level attention mechanism: visit-level attention weights (α) identify which clinical encounters contributed most to the prediction, while code-level weights (β) highlight specific diagnoses and medications within each visit. For a patient flagged as high-risk, RETAIN might reveal that the model attended primarily to a recent hospitalization (e.g., visit attention $\alpha = 0.31$), with diagnosis codes for cardiac dysrhythmia and heart failure receiving the highest code-level contributions ($\beta > 0.15$). Applying SHAP (61) to the same prediction produces additive feature attributions, where studies show that serum creatinine levels typically emerge as the strongest predictor, with elevated values contributing positively to mortality risk

Table 1. Overview of XNLP Applications, Subcategories, and Case Studies.

XNLP Applications	Subcategories	Studies
Medicine	EHRs	(86; 87)
	Medical Documents Analysis	(88–91)
Finance	Risk Assessment	(92; 93)
	Fraud Detection	(92; 94–96)
	Firm Valuation	(97; 98)
Systematic Reviews	Review Automation	(99)
	Text Summarization	(100)
CRM	Sentiment Analysis	(101–103)
	Customer Support Automation	(104; 105)
Chatbots and Conversational AI	Conversational Agents	(104; 106)
	Context-Aware Recommendations	(107)
Social and Behavioral Science	Sexism and Hate Speech Detection	(108–114)
	Fake News and AI Generative Detection	(115; 116)
Human Resources	Talent Acquisition and Recruitment	(15; 117)
	Employee Sentiment Analysis	(27; 118)
	Performance Evaluation	(119; 120)
	Diversity and Inclusion	(121)

while normal values reduce predicted risk (123). Meanwhile, LIME (3) fits a local linear approximation, enabling counterfactual reasoning such as estimating how risk would change if a specific diagnosis were absent. Each method illuminates different aspects: RETAIN reveals temporal patterns through reverse-time attention, SHAP provides globally consistent feature attributions grounded in game theory, and LIME offers locally faithful approximations for individual predictions. Recent meta-analyses of XAI in clinical decision support systems (124; 125) emphasize that combining multiple explanation methods provides clinicians with richer decision support than any single method alone.

Table 2 provides a concise overview of major XNLP applications in medicine, showing the interplay among model architectures, explainability techniques, and evaluation metrics.

Table 2. Summary of XNLP Applications in Medicine

Paper	Application	Model	Explainability Method	Dataset	Metrics
(86)	Heart Failure Prediction	RNN	Feature Importance	EHRs	Accuracy, AUC-ROC
(126)	EHRs Classification	CNN	Rationale-based explanations	MIMIC-III	Precision, Recall, F1-score
(90)	RCT Analysis & Extraction	Rule-based NLP	Transparent Rule Extraction	Clinical Trials	Extraction Accuracy
(127)	Biomedical Word Embeddings	fastText + MeSH	MeSH-Based Interpretations	UMNSRS-Sim, UMNSRS-Rel	Embedding Quality
(128)	Disease Progression Prediction	Transformer	Attention Mechanism	Public EHRs	RMSE, MAE
(87)	Cancer Diagnosis	BERT	LIME and SHAP	Cancer Registry	F1-score, Precision, Recall

Alongside EHR classification and rule-based extraction, contemporary work uses deep learning methods, e.g., word embeddings, transformers, to produce more robust predictive power in domains like cancer diagnostics (87), disease progression modeling (128), and ICD-10 coding (126). Nonetheless, the need for interpretability remains pressing, as stakeholders require transparent models to validate medical recommendations and address concerns about potential biases. XAI enhances trust among healthcare professionals by elucidating AI-driven decisions, thereby meeting regulatory transparency requirements and promoting fairness and safety in clinical settings (129).

Research on mental health analysis via XNLP further underscores the domain's complexity and sensitivity. For instance, LLM-based strategies to detect distress patterns depend on clear, faithful explanations that can be passed to mental health practitioners (122). Ensuring patient privacy, dealing with text de-identification (130), and mitigating data biases remain crucial obstacles in this space. Overall, integrating XNLP in the medical domain holds the promise of safer clinical decision support, faster literature reviews, and more equitable patient care by revealing how and why a system recommends certain treatments or diagnoses.

Finance

Financial decision-making involves intricate processes such as risk assessment, fraud detection, and firm valuation, all of which require both accurate and transparent AI-driven insights. Recent studies show that incorporating NLP techniques to analyze financial reports, news articles, or transaction logs can effectively flag risk factors and provide timely alerts (131). However, traditional NLP-based models often lack explainability, which is critical when end users, be they financial experts or customers, need to understand *why* a model has flagged a transaction as fraudulent or assigned a particular credit score. XAI addresses this gap by providing interpretable insights that can build trust in the system's outputs. For instance, integrating XAI techniques into credit scoring models enhances transparency and compliance with regulations like the General Data Protection Regulation (GDPR) and the Equal Credit Opportunity Act (ECOA), ensuring that algorithmic decisions are understandable and coherent (132). Additionally, employing model-agnostic explanation methods such as SHAP and LIME in credit risk management helps stakeholders comprehend the reasoning behind model predictions, thereby fostering trust and facilitating informed decision-making (133).

Risk assessment underpins pivotal tasks in the financial sector, such as loan approvals, insurance rate settings, and investment decisions. Inaccuracies or lack of transparency in this process can have significant ramifications, from unfair interest rates to systemic risks. XNLP can help by elucidating which textual factors, like specific keywords in a credit application or trends in financial reports, contribute to elevated risk scores (92; 93). For example, (93) proposed a graph-based attention model for credit risk assessment, highlighting how interactions and transactions shape the final score. Such clarity is essential not only to justify decisions to regulators and auditors but also to empower clients to take corrective steps to mitigate risk, thereby fostering increased trust in financial institutions (134).

Fraud detection systems have traditionally operated as opaque "black boxes," leaving end users uncertain about what triggers a fraud alert. XNLP tools shine a light on relevant textual features, possibly certain transaction notes or unusual patterns in communications, to clarify why a specific transaction has been flagged (92; 94–96). Interpretable machine learning frameworks, such as decision trees, can provide feature-importance scores, while neural architectures can integrate attention layers that visually highlight suspicious keywords. For instance, (96) introduced an XAI approach for fraud detection, aligning fraud experts' investigative tasks with model-generated explanations. Enhancing user comprehension of these flags not only minimizes false positives but also strengthens collaboration between human fraud analysts and AI systems.

The practical value of combining ensemble methods with explainability is well demonstrated in fraud detection research (135; 136). Stacking ensemble models that integrate random forests, gradient boosting, and neural networks can achieve 99% accuracy with AUC-ROC scores of 0.998 across five-fold cross-validation when trained on datasets containing approximately 590,000 transactions with a 3.5% fraud rate. SHAP-based feature selection identifies the most influential predictors, with engineered features (such as V14 in the commonly used Kaggle credit card dataset) consistently emerging as top contributors, followed by transaction amount and temporal patterns. When such a model flags a transaction, SHAP waterfall plots decompose the prediction into individual feature contributions: for instance, an unusually high transaction amount relative to the cardholder's typical behavior might contribute positively to the fraud score, while established account tenure and consistent merchant category usage reduce it. Notably, exploratory analyses reveal that most fraudulent transactions involve amounts below $1,000, exhibiting more dispersed value distributions compared to legitimate transactions (135). This granular decomposition enables fraud analysts to verify whether flagged features align with known fraud patterns, improving investigative efficiency. However, (136) note that model-agnostic methods like SHAP face scalability challenges in real-time systems processing thousands of transactions per second, motivating research into approximate explanation methods—such as stability-aware SHAP caching that reduces explanation latency to under 50 milliseconds—that trade some precision for computational efficiency (137).

XNLP methods are also gaining traction in firm valuation, a high-stakes arena of finance where annual reports, market data, and corporate disclosures must be parsed to determine a company's worth. A key challenge lies in distinguishing relevant signals from strategic or even misleading language. More advanced NLP models, including *transformer-based* architectures, attempt to capture long-range dependencies and contextual nuances (98). Yet context-dependent meanings, sarcasm, and subtle cues remain challenging to detect, raising concerns that companies might "game the system" by inserting specific words likely to inflate valuations (138). To address this, XNLP solutions like the one proposed by (98) incorporate interpretability mechanisms (e.g., attention-based explanations) to show precisely which textual segments influenced a model's valuation output. Similarly, (97) found that combining explainable lexicon models with sentiment analysis in financial texts improves both accuracy and interpretability, allowing investors and auditors to discern exactly how sentiment-laden phrases affect overall firm valuation.

Table 3. Summary of XNLP Applications in Finance

Paper	Application	Model	Explainability Method	Dataset	Metrics
(139)	Risk Assessment Classification	BERT	Layer-wise relevance propagation	20 Newsgroups	Accuracy, F1-score, Precision, Recall
(93)	Credit Risk Assessment	Graph-based Attention	Probing Methodology	Financial Transactions	Evaluation Metrics
(95)	Fraud Alert Explanation	ML Models (e.g., Decision Trees, SVM)	Feature Importance	Transaction Data	Precision, Recall, AUC
(96)	Fraudulent Transactions Justification	XAI Methods	Explanation Generation	Financial Transactions	Accuracy, F1-score, Precision, Recall
(98)	Firm Valuation	Transformer-based Model	Explanation Generation	Financial Documents	ROUGE, BLEU
(97)	Sentiment Analysis for Valuation	Explainable Lexicon Model	SHAP Explainability	Financial Texts	Accuracy, F1-score, Precision, Recall

In summary, XNLP is reshaping key finance processes by injecting interpretability into risk analyses, fraud alerts, and valuation metrics. Explainable risk assessments help users manage their creditworthiness more effectively, while transparent fraud detection can strengthen the reliability and acceptance of AI-driven alerts. Meanwhile, interpretable firm valuation models demystify how textual content in financial disclosures influences market perceptions. As Table 3 illustrates, the state of the art spans various architectures (from decision trees to transformers) and explanation frameworks (from feature-importance scores to saliency maps), reflecting a rapidly evolving field. Going forward, deeper integration of XNLP principles, coupled with robust evaluation metrics, will be pivotal in boosting stakeholder confidence and ensuring more equitable, auditable financial ecosystems.

Systematic Reviews

Systematic reviews are a cornerstone of evidence-based decision-making in various fields. The process commonly involves screening large volumes of research literature, extracting relevant data, and synthesizing key findings. Although AI-based automation can accelerate tasks such as study identification or data extraction, explainability plays a crucial role in ensuring researchers understand why specific studies are included or excluded. Techniques derived from *XNLP* help demystify this process, offering transparency and trust in the automated workflow.

For instance, some approaches such as (99) use Support Vector Machines (SVMs) and specialized explanation frameworks to elucidate which textual features contributed most to including or excluding a study. These methods, collectively, allow reviewers to focus on interpreting potentially significant papers rather than sifting through thousands of irrelevant ones.

(100) likewise developed RobotReviewer*, which addresses bias assessment within systematic reviews. By using text analysis for bias detection in clinical trials, RobotReviewer can automatically surface key phrases or patterns indicative of methodological flaws. When combined with an explainable component, this process not only speeds up bias assessment but also clarifies the reasons behind each flagged instance. Table 4 highlights notable XNLP applications in systematic reviews, detailing the associated models, explainability methods, data sources, and relevant metrics.

Table 4. Summary of XNLP Applications in Systematic Reviews

Paper	Application	Model	Explainability Method	Dataset	Metrics
(99)	Review Automation	SVM	Explanation Framework	PubMed	WSS†, RRF‡
(100)	Bias Assessment	RobotReviewer	Text Analysis for Bias Detection	Clinical Trials	Bias Assessment Metrics, Accuracy
(140)	Model Interpretation	LSTM	Information Bottleneck	SST-2, IMDb	Accuracy, Mutual Information

Recent large-scale simulation studies have shown that active learning models enhance the efficiency of systematic review screening processes. By prioritizing the most relevant records, these models reduce the manual workload required for reviewing literature. For instance, (141) conducted over 29,000 simulations across various model configurations and datasets, consistently finding that active learning outperformed random screening strategies in identifying pertinent studies.

Quantitative evaluations demonstrate the substantial efficiency gains achievable through these approaches. Studies using SVM classifiers with TF-IDF feature extraction report Work Saved over Sampling (WSS@100) scores of 88.5%, meaning reviewers need to screen only 11.5% of records to achieve complete recall (142). In food safety literature screening, active learning models achieved 99.2% recall while reviewing only 62.6% of total records (143). For bias assessment in clinical trials, RobotReviewer demonstrates inter-rater reliability (Cohen's κ) of 0.42–0.48 across risk-of-bias domains, comparable to agreement levels between human reviewers (100). These metrics underscore that XNLP tools can match human-level performance while dramatically reducing screening burden, though the trade-off between recall and efficiency requires careful calibration based on review objectives. As the volume of global research continues to grow, incorporating XNLP into automated reviewing workflows stands as a promising strategy to enhance speed, clarity, and consistency in evidence synthesis.

Customer Relationship Management (CRM)

CRM involves diverse functions such as tracking user sentiment, generating automated summaries of user feedback, and providing responsive customer support. Recent advances in XNLP have shown promise in improving the transparency and effectiveness of these processes. By revealing the underlying factors that drive model outputs, organizations can better trust, refine, and act upon AI-generated insights.

A key role for XNLP in CRM emerges in **text summarization**, where textual data from various sources, such as product reviews or user comments, must be condensed into concise, informative summaries. For instance, (144) used BERT for extractive summarization in biomedical literature, employing attention visualization to demonstrate how certain sentences contribute to the final output. This approach not only highlights important segments of text but also helps stakeholders understand why particular sentences were chosen. Similarly, the inclusion of **attention-based or rationale-based explanations** promotes transparency, enabling decision-makers to trust automated summaries in areas like market research or product feedback analysis.

Another core application is **sentiment analysis**, which is vital for gauging public perception and refining marketing strategies. Although sentiment analysis can detect emotions or opinions in large-scale text data, it often operates as a "black box." By applying explainability methods such as LIME (LIME), Layer-wise Relevance Propagation (LRP), or attention-based heatmaps, organizations can better interpret why a particular sentiment label was assigned (101–103). For example, (102) integrated LRP into a BiLSTM model to show which words triggered specific sentiment predictions, and (103) used SHAP explanations to reveal token-level contributions, boosting user confidence in the underlying sentiment classification.

Beyond sentiment analysis, XNLP also enhances customer support automation, including chatbots and self-service platforms that respond to consumer inquiries in real time. By making these systems explainable, companies can pinpoint why certain answers or suggestions were given, reducing user frustration and building trust (104; 105). For example, (105) integrated explainability in a transformer-based chatbot, visualizing attention weights to show customers how their queries influenced responses. Table 5 illustrates various CRM-related applications, outlining the models, explainability techniques, datasets, and performance metrics used.

[*] https://www.robotreviewer.net/

Table 5. Summary of XNLP Applications in CRM

Paper	Application	Model	Explainability Method	Dataset	Metrics
(144)	Text Summarization	BERT	Attention Visualization	Biomedical Articles	ROUGE, BLEU
(101)	Sentiment Analysis	BERT	LIME	Yelp Reviews	Accuracy, F1-score
(102)	Sentiment Analysis	BiLSTM	Layer-wise Relevance Propagation	Amazon Reviews	Accuracy, Precision, Recall
(103)	Sentiment Analysis	BiLSTM	SHAP	Movie Reviews	Accuracy, F1-score, Precision
(104)	Customer Support Automation	Seq2Seq	Explanation-by-Example	Customer Service Logs	Customer Satisfaction Score
(105)	Customer Support Automation	Transformer-based Chatbot	Attention Visualization	Customer Service Logs	Response Time, User Satisfaction
(145)	Text Classification, Info Extraction	LSTM	Rationale-based Explanations	Beer Review, CoNLL-2003	Accuracy, F1-score
(146)	Sentiment Analysis	BERT	Feature Visualization	Yelp Polarity	Accuracy, Precision, Recall, F1-score
(147)	Review Automation	LSTM	Shapley Interaction Index	IMDb	Accuracy, F1-score, Precision, Recall
(148)	Fact Verification	BiGRU + Attention	Attention Heatmaps, Explainable Fact-Checking	LIAR, PolitiFact	Accuracy, Precision, Recall, F1-score
(149)	Sentiment Analysis	CNN, BiGRU	Attention Heatmaps, Ablation	Amazon, BeerAdvocate	Accuracy, Precision, Recall, F1-score
(150)	Sentiment Analysis	RL + RNN	Rationalization Generation	TripAdvisor, RateBeer	Accuracy, F1-score

Empirical evaluations demonstrate the strong performance of transformer-based models in CRM applications. BERT-based sentiment classifiers achieve 89–92% accuracy on product review datasets, with fine-tuned variants reaching up to 94% F1-score when combined with attention-based architectures (151; 152). SHAP and LIME integration enables token-level attribution, revealing which words or phrases most influenced sentiment predictions—for instance, identifying that adjectives describing product quality contribute more strongly to positive classifications than generic praise. This granular explainability proves particularly valuable in e-commerce settings, where understanding *why* a review was classified as negative enables targeted product improvements. The interpretable attention mechanisms in these hybrid architectures allow users to trace sentiment predictions back to specific textual features, though enhanced explainability introduces computational overhead during inference (152).

Overall, XNLP's integration into CRM marks a key shift in how businesses collect, interpret, and act on text data. By offering transparent rationales for automated summarization, sentiment detection, and response generation, these systems help organizations forge deeper connections with their customers. Moreover, explainability not only enhances user trust but also empowers developers and stakeholders to refine their NLP pipelines for improved accuracy and reliability. As shown in Table 5, a variety of architectures and techniques, ranging from BiLSTM to Transformer-based models, are being equipped with explainability features, enabling more interpretable and user-aligned CRM solutions.

Chatbots and Conversational AI

Chatbots and conversational AI systems have become increasingly prevalent in areas such as virtual assistance, customer support, and information retrieval. By using XNLP, these systems can offer more transparent and user-centric interactions, as their generated responses or recommendations can be coupled with clear rationales. This transparency fosters greater user trust, enabling individuals to understand the *why* behind the bot's outputs and to feel more confident in adopting its suggestions or insights.

Context-aware recommendation systems aim to deliver personalized suggestions by incorporating elements of user intent, interaction history, or external data. Incorporating XNLP techniques into these recommenders can enhance user confidence by explicitly showing which contextual cues led to particular recommendations. For instance, (107) proposed a context-aware recommendation model that provides explanations about how user-item interactions influenced its outputs, leading to higher satisfaction. Similarly, (106) demonstrated how a chatbot could justify its decisions, improving user trust and engagement.

In 2023, conversational AI took a significant leap with the introduction of GPT-4, known for its larger context window and improved accuracy in both English and code-based tasks. OpenAI's GPT-4o ("o" for "omni") extended these advantages, matching GPT-4 Turbo performance in English while showing further gains in non-English languages (47; 153; 154). Although these advanced models can generate more coherent and context-aware dialogue, they still require effective *explainability* strategies to clarify the internal reasoning paths or chain-of-thought that lead to each response. XNLP methods thus remain vital for building user trust, especially as chatbot complexity and capabilities continue to grow.

Recent studies underscore the importance of explainable conversational AI for enhancing user satisfaction, transparency, and overall effectiveness (155–157). While AI chatbots increasingly excel at generating swift, personalized responses, embedding XNLP capabilities can reveal how the system arrives at each answer. This can be accomplished through mechanisms like **attention distributions**, **rationale highlighting**, or **feature visualization**, each of which helps users grasp the underlying logic. Such interpretability also aids in detecting potential biases or inconsistencies, enabling developers to refine their models and ensuring the system behaves reliably under different conditions (158–160).

Table 6 highlights a range of applications in chatbots and conversational AI, detailing the models, explainability techniques, datasets, and evaluation metrics employed. These examples demonstrate how integrating transparency can bolster user trust, streamline recommendation processes, and refine user engagement strategies.

Table 6. Summary of XNLP Applications in Chatbots and Conversational AI

Paper	Application	Model	Explainability Method	Dataset	Metrics
(64)	Conversational Agents	BiLSTM, CNN	Attention Distributions	SST-5	Accuracy, F1-score
(104)	Automated Customer Service	Transformer Models	Justifications for Responses	IMDb	Efficacy
(106)	Chatbots	Various	Explainability	AG News	User Trust, Engagement
(107)	Recommendation Systems	Deep Learning-based	Attention Mechanism	MovieLens, Amazon Product	User Satisfaction
(156)	Conversational Agents	GPT-4	Transparency	MovieLens, Amazon, Goodreads	User Engagement, Accuracy
(155)	Chatbots in Business	Various	Transparency in Decision-Making	General Business Data	User Trust, Satisfaction
(161)	Task-Oriented Dialogue	Seq2Seq	External DB Integration	MovieLens, Reddit	Response Accuracy

Empirical research on chatbot trust reveals that explainability significantly influences user acceptance and satisfaction. Systematic reviews synthesizing 40 studies identify five categories of trust predictors: user characteristics, machine attributes, interaction quality, social factors, and contextual elements (162). Quantitative studies demonstrate that chatbot usability explains 59% of variance in user trust ($R^2 = 0.59$, $\beta = 0.50$, $p < 0.005$), while trust itself accounts for 9.9% of variance in customer attitudes, 11.4% in behavioral intentions, and 13.6% in user satisfaction (163). Wizard-of-Oz experiments confirm that high explainability in conversational agents has a positive effect on trust and acceptance, particularly when explanations incorporate both dialogue-level context and feature-level rationales. However, balancing explanation detail with conversational fluency remains challenging: overly verbose explanations can disrupt dialogue flow, while sparse explanations may fail to satisfy user curiosity about system behavior.

Overall, the integration of XNLP into conversational AI has the potential to reshape user interactions by providing richer, more transparent exchanges. As large-scale models like GPT-4o continue to grow in complexity, designing robust *explainability* mechanisms will be central to maintaining user trust and improving dialogue outcomes. From clarifying context-aware recommendations to justifying chatbot responses, XNLP methods play a pivotal role in the future trajectory of conversational systems.

Social and Behavioral Science

Social and behavioral science research often addresses highly complex societal challenges, such as hate speech, sexism, misinformation, and emotional well-being. *XNLP* provides valuable tools for enhancing the reliability of automated methods in this domain, enabling clearer rationales behind model outputs. This transparency allows researchers and practitioners to gain deeper insights into underlying language patterns and the decision-making processes of NLP models.

Transformer-based architectures, including BERT, have significantly advanced the detection of offensive language. (108) relied on BERT to accurately analyze annotated data for hate speech, improving both detection performance and interpretability. Likewise, (109) developed domain-specific word embeddings focused on hate speech data, offering nuanced insights into how such content manifests linguistically. Efforts such as the sEXism Identification in Social neTworks (EXIST) and SemEval-2023 Task 10 competitions demonstrate the multilingual capabilities of XNLP for tackling sexist or hateful content in English and Spanish (110–113). These studies highlight the value of *human-in-the-loop* validation, where user feedback helps align model outputs with human judgments and mitigate biases (114).

Benchmark datasets with integrated explainability annotations have proven valuable for advancing research in this area. The HateXplain dataset (164) provides 20,000 social media posts annotated with three-way classification labels (hate, offensive, normal), target community identification, and human rationales indicating which text spans influenced labeling decisions. BERT-based models trained with these rationales achieve 0.698 accuracy and 0.687 F1-score, with AUROC of 0.851—modest improvements over standard BERT (0.690 accuracy, 0.674 F1)—but critically, models incorporating human rationales during training demonstrate reduced unintended bias toward target communities. In the SemEval-2023 Task 10 competition on explainable sexism detection, the best-performing systems achieved 0.84 macro F1-score for binary classification, though performance dropped substantially (0.43 F1) for fine-grained categorization across 11 sexism subcategories (111). These results underscore a key insight: models achieving high classification accuracy do not necessarily score well on explainability metrics such as plausibility (overlap with human rationales) and faithfulness (whether explanations reflect actual model reasoning).

Misinformation poses an escalating threat in digital communication. XNLP offers robust techniques to identify fabricated or AI-generated text, thereby safeguarding the integrity of online information. (115) illustrate the efficacy of XNLP in distinguishing genuine content from deceptive or machine-generated posts. Similarly, (116) created a multilingual ensemble model, tested under a shared task by Computational Linguistics in the Netherlands (CLIN), to discern AI-generated text from human-authored writing (165). Beyond purely text-based strategies, multimodal systems like *SceneFND* integrate textual, contextual, and visual cues, showing enhancements in identifying misinformation across diverse datasets (166).

Though frequently classified under *medicine* (see Medicine section), research in social and behavioral contexts also applies XNLP to assess mental health conditions and public sentiment. (167) employed NLP features from social media posts to predict emotional or psychological well-being. Meanwhile, (168) experimented with transformer-based architectures to classify depression severity, illustrating how explainable output can guide mental health interventions. Additionally, (169) leveraged large-scale XNLP methods for real-time analysis of political sentiment on Twitter during the 2020 U.S. Presidential Elections, demonstrating how these techniques can capture public opinion dynamics.

(170) introduce *L-Defense*, a framework that partitions crowd discourse on a news claim into two opposing camps. It then extracts salient evidence for each side and prompts a large language model (LLM) to debate and defend their respective narratives. The "winning" defenseproduces a concise natural-language explanation along with the final veracity label, achieving higher detection accuracy than previous methods and offering users a transparent, evidence-based rationale.

(171) compare GPT-4 and Claude 3 in a tumor-board scenario, finding that both models produce expert-level treatment plans, each accompanied by step-by-step rationales that clinicians can readily audit. In parallel, (172) study social-media language to evaluate depression severity, identifying the phrases that most influenced each assessment. Such research underscores an emerging focus on verifying not only the plausibility but also the *practical utility* of XNLP explanations in real-world healthcare settings.

Domain-Specific Challenges. The social and behavioral science domain presents unique challenges for XNLP that distinguish it from other application areas (173; 174):

- **Cultural and Linguistic Variation:** Hate speech and offensive language manifest differently across cultures, languages, and platforms. What constitutes hate speech in one cultural context may be acceptable discourse in another, and XAI methods must account for these variations rather than applying universal standards (175). Low-resource languages present additional challenges, as training data and lexicons are often unavailable.

- **Annotation Bias and Subjectivity:** Ground-truth labels in hate speech datasets often reflect annotator demographics and perspectives. When models learn from biased annotations, their explanations may perpetuate those biases. Explaining *why* a model flagged content as hateful requires acknowledging the inherent subjectivity of the labeling process itself.

- **Evolving Language and Adversarial Evasion:** Users attempting to spread hate or misinformation continuously adapt their language to evade detection—using coded language, deliberate misspellings, or memes. XAI methods must not only detect current patterns but also provide explanations that help human moderators understand novel evasion tactics.

- **Platform-Specific Context:** The same text may have different meanings on Twitter versus Reddit versus messaging apps. Explanations must incorporate platform-specific context (thread structure, community norms, user history) to be actionable for moderators.

These challenges highlight that effective XNLP in social science requires not just technical sophistication but also deep engagement with the sociocultural contexts in which models are deployed.

Table 7. Summary of XNLP Applications in Social and Behavioral Science

Paper	Application	Model	Explainability Method	Dataset	Metrics
(108)	Hate Speech Detection	BERT	Transformer Models	Annotated Datasets	Detection Accuracy
(109)	Hate Speech Analysis	Domain-Specific Models	Word Embedding	Hate Speech Websites	Delicate Language Insights
(110)	Sexism Detection	ML Algorithms	EXIST Competition	EXIST-2023	Classification Accuracy
(108)	Hate Speech Detection	BERT + Bias Mitigation	Bias Mitigation Mechanism	Davidson & Waseem	F1-measure
(109)	Hate Speech Detection	BERT & Deep Models	LIME	Hate Speech Word Embedding	Detection Accuracy
(111)	Explainable Online Sexism Detection	BERT	SHAP	SemEval-2023 Task 10, EXIST-2023	F1 Score
(113)	Explainable Sexism Detection	Ensemble (BERT, XLM-R, DistilBERT)	SHAP	EXIST-2023	Token Influence Analysis
(116)	AI-Generated Text Detection	Ensemble + Multilingual BERT	SHAP	Various Genres	Detection Accuracy
(167)	Mental Health Analysis	NLP Techniques	Feature Importance	Social Media Posts	Predictive Power
(170)	Fake-News Detection (L-Defense)	LLM-based Debate	Natural-Language Justification	Crowd Discourse	Detection Accuracy
(171)	Tumor-Board Treatment Plans	GPT-4, Claude 3	Step-by-step Rationales	Clinical Case Reports	Expert-Level Validity
(172)	Depression Severity Assessment	Transformer-based	Highlighted Key Phrases	Social Media Posts	Clinical Utility

Overall, integrating XNLP into social and behavioral science marks a significant advance in addressing issues such as hate speech, sexism, misinformation, and mental health monitoring (164; 176). By using the interpretability features of advanced transformer models, domain-specific embeddings, and multilingual datasets, researchers can develop classifiers that not only identify problematic content or behaviors but also explain their predictions. This approach reinforces the credibility of automated systems and aids stakeholders, ranging from social scientists to clinicians, in better understanding the nuanced language patterns that drive complex social phenomena.

Human Resources (HR)

HR encompasses processes ranging from talent acquisition and employee sentiment analysis to performance reviews and diversity initiatives. XNLP has recently gained traction in optimizing these functions, offering transparent decision-making tools that can build trust, reduce biases, and streamline various HR operations.

Transformer-based architectures like BERT and GPT are increasingly used to automate the resume-screening and candidate-ranking process, matching required skills more effectively with job descriptions. For instance, (117) demonstrate how attention mechanisms within BERT can make the initial recruitment phase more efficient and less prone to subjective bias. In parallel, (15) show how attention visualization explains model outputs, thereby enhancing stakeholder trust in the screening process. (177) further highlight how a transformer-based ensemble model can accurately extract both technical and non-technical competencies, significantly boosting the precision and transparency of candidate matching.

In understanding workforce morale, XNLP provides robust sentiment analysis techniques to analyze feedback from surveys, internal forums, and social media. Using transformer-based models like RoBERTa, (118) process large volumes of employee comments, pinpointing negative sentiments and emerging workplace issues. Additionally, attention-driven methods clarify which textual factors influenced each sentiment label, allowing HR departments to better understand employee concerns (27).

Traditional performance reviews often lack consistency and can harbor biases. By incorporating XNLP to analyze written feedback and performance data, organizations achieve more objective and transparent evaluations. (119) demonstrate how attention mechanisms in transformer models highlight the most relevant textual feedback, providing a clear rationale for each performance score. Recent work by (120) similarly suggests that integrating explainabilityproduces more equitable outcomes, as HR managers can interpret the exact factors driving each model's assessment.

XNLP further contributes to fostering diversity and inclusion within companies by detecting linguistic biases in job postings, internal policies, and employee communications. For example, (121) propose an XNLP framework that not only flags biased content but also clarifies its reasoning through interpretable outputs. This transparency helps organizations address the root causes of inequality and maintain more inclusive hiring practices.

Despite these advances, significant challenges remain regarding bias in AI-driven recruitment. Large-scale evaluations reveal that AI resume screening systems exhibit gender, racial, and intersectional biases, with some models showing an 18% higher rejection rate for candidates from underrepresented groups compared to majority-group candidates with equivalent qualifications (178). Sentence-BERT embeddings used for resume-job matching achieve approximately 92% top-1 accuracy when evaluated on curated datasets, using cosine similarity thresholds around 0.7 for candidate selection (179). However, the "black-box" nature of many systems complicates bias auditing. As of 2024, only 21% of companies report rejecting candidates without human review, suggesting widespread adoption of human-in-the-loop approaches (178). Regulatory frameworks are emerging in response: New York City mandates auditing of AI hiring systems since 2023, while Colorado's comprehensive AI employment law takes effect in 2026. These developments underscore the need for XNLP methods that not only improve recruitment efficiency but also provide transparent, auditable decision rationales to satisfy both ethical standards and legal requirements.

Table 8. Summary of XNLP Applications in HR

Paper	Application	Model	Explainability Method	Dataset	Metrics
(117)	Recruitment Automation	BERT	Attention Mechanisms	Resume Data	Match Accuracy
(15)	Recruitment	BERT	Attention Visualization	Resume Data	Bias Reduction
(177)	Skill Extraction	Transformer Ensemble	Attention Mechanisms	Job Descriptions	Precision, Recall
(118)	Employee Sentiment Analysis	RoBERTa	Attention Mechanisms	Employee Feedback	Sentiment Insights
(27)	Sentiment Analysis	RoBERTa	Attention Mechanisms	Employee Feedback	Sentiment Accuracy
(119)	Performance Evaluation	Transformer Models	XAI	Performance Reviews	Objectivity Scores
(120)	Performance Evaluation	Transformer Models	XAI	Performance Reviews	Fairness Scores
(121)	Diversity	Custom NLP	XAI	Company Policies	Bias Detection

As illustrated in Table 8, the use of XNLP across various HR functions, from screening resumes to promoting inclusivity, underscores its capacity to bring objectivity, transparency, and efficiency to people-centric tasks. Studies consistently show that leveraging explainability features (e.g., attention heatmaps, rationale generation) not only strengthens user trust but also aligns HR strategies more closely with organizational values. Looking ahead, ongoing research into transformer-based and explainable methods promises further refinements in reducing biases, understanding employee concerns, and improving overall talent management.

Other Applications

Beyond the domains discussed earlier, XNLP finds utility in a wide range of tasks and sectors. These include language generation, text classification, machine translation, summarization, visual question answering, and more. In such contexts, explainability is vital for understanding *how* and *why* a model produces specific outputs, offering transparency and trustworthiness in settings that may require critical decision-making or that deal with large-scale user interactions.

(180) employed Transformer-based models (e.g., GPT-2) to visualize attention mechanisms and detect biases in generated language. Such visualization not only reveals which tokens or phrases drive certain outputs but also highlights how societal or dataset biases might surface in generation processes.

(181) proposed rationale-based explanations for text classification tasks across multiple datasets (BoolQ, e-SNLI, etc.), evaluating the quality of explanations on fidelity, comprehensiveness, and sufficiency. Similarly, (182) and (183) extended the discussion to machine translation, summarization, and visual question answering, using (multi)transformer models for *faithfulness* and *rationale generation*, respectively.

Several works have delved into annotation of word importance and visualization of model internals. For instance, (184) and (185) examined neural machine translation and reading comprehension using BERT and Transformer-based architectures, annotating tokens with importance weights and generating explanations tested on datasets such as 20NG, AGNews, IMDB, SQuAD, and e-SNLI. Likewise, (186) and (187) employed BERT and other Transformer variants to study learned self-attention, providing interactive visual analyses of attention patterns. Beyond purely text-based challenges, multimodal Transformer approaches such as (183) and (188) addressed visual commonsense reasoning, indicating how explainable components can strengthen cross-domain understanding.

Commonsense-related tasks and interpretability-oriented methods have also drawn increasing attention. For example, (189) introduced FiD-Ex to generate extractive rationales, while (190) used T5-based joint models to produce free-text explanations in tasks like commonsense question-answering and natural language inference. Additional works by (191), (192), and (193) concentrated on translation quality estimation, text classification, and universal rationalization frameworks, respectively. Table 9 highlights diverse XNLP applications along with key models, methods, datasets, and metrics.

XNLP's application in these diverse areas underscores its growing impact, including tasks that may not carry as high a risk as medical or financial domains but still benefit from transparency (e.g., language bias detection, rationale generation, and VQA). By illuminating *how* decisions are reached, XNLP fosters trust and reliability in NLP-driven solutions, particularly in critical endeavors such as summarization, machine translation, or text classification, where hidden biases could otherwise remain undiscovered. The consistent theme across studies is that explainability strengthens user confidence and supports auditing and refining model performance.

From the compiled studies, we observe that *risk-sensitive* areas (e.g., healthcare) rely strongly on actionable explanations (feature importance, rule extraction) to ensure trustworthy decision-making. In contrast, sectors like finance emphasize attention-based methods for spotting anomalies in vast data streams. CRM and chatbot applications often focus on generating user-facing explanations, thereby building trust in real-time interactions. Regardless of domain, the shared principle is that *explainability* underpins system reliability and user acceptance, paving the way for broader adoption of NLP-based automation.

Cross-Domain Analysis of Divergent Explainability Requirements

While the preceding sections examined XNLP applications within individual domains, a cross-domain perspective reveals both common threads and critical differences in how explainability is conceptualized, implemented, and evaluated. This synthesis addresses a key gap in the literature: the lack of systematic comparison across domains (194; 195). Table 10 provides a comparative overview of domain-specific XNLP requirements.

Table 9. Summary of XNLP Applications in Other Domains

Paper	Task	Model	Explainability Method	Dataset	Metrics
(180)	Language Generation	GPT-2, etc.	Attention Visualization	WebText	Automated Bias Metrics
(184)	Neural Machine Translation	BERT	Word Importance Annotation	Various (20NG, AGNews, IMDb)	Accuracy, F1
(181)	Text Classification & Rationale	Various Models	Rationale-based Explanations	BoolQ, e-SNLI, etc.	Fidelity, Comprehensiveness
(182)	Machine Translation, Summ.	Transformer Models	QA for Faithfulness	CNN/Daily Mail, XSum	Fidelity Score
(183)	Visual Question Answering	Multimodal Transformer	Rationale Generation	VQA v2.0, VizWiz	Answer Accuracy
(185)	Machine Reading Comprehension	Transformer-based	Explanation Generation	SQuAD 1.1, e-SNLI	Fidelity, Sensibility
(186)	Natural Language Understanding	BERT	Activation & Attention Visualization	BERT-based Tasks	(Varied)
(187)	Transformer Model Analysis	BERT, etc.	Visualization, Self-Attention	Not Specified	Not Specified
(189)	Extractive Rationale Generation	LSTM, BERT	FiD-Ex	VCR	Exact Match, Rationale F1
(188)	Visual Commonsense Reasoning	DMVCR	Dynamic Working Memory	VCR, Revisited-VQA	(Task-Specific)
(190)	Free-text Rationales	T5-based Joint Models	Natural Language Rationales	Commonsense QA, NLI	Feature Importance Agreement
(191)	NMT Quality Estimation	Transformer	LIME, Integrated Gradients	MLQE-PE	Pearson, Spearman, MAE
(192)	Neural Machine Translation	BERT-based	Feature Visualization	WMT Metrics Shared Task	Pearson, Kendall's Tau
(193)	Text Classification	UniREX	Rationalization	FEVER, Movie Reviews	Precision, Recall, F1

Convergent Themes. Despite domain-specific variations, several themes emerge across all application areas:

- **Trust as a Universal Goal:** Across medicine, finance, HR, and social science, building user trust in model predictions remains paramount. However, trust operationalization varies: clinicians require explanations that map to medical concepts (196), while financial regulators demand audit trails that satisfy compliance frameworks (137).

- **Feature Importance Dominance:** SHAP and LIME remain the most widely adopted methods across domains due to their model-agnostic nature and intuitive output format. However, their computational cost limits real-time deployment in finance and CRM settings (136).

Table 10. Cross-Domain Comparison of XNLP Requirements, Methods, and Challenges

Domain	Primary XAI Need	Preferred Methods	Key Metrics	Unique Challenges
Medicine	Clinical actionability; patient safety	Feature importance, rule extraction, attention visualization	AUC, clinical validity, user trust	HIPAA/GDPR privacy constraints; concept drift in EHR data; class imbalance; integration with clinical workflows (124; 125)
Finance	Regulatory compliance; fraud detection	SHAP, LIME, attention heatmaps, ensemble methods	AUC-ROC, accuracy, compliance score	Adversarial gaming of explanations; real-time scalability; temporal dynamics in markets (135; 137)
CRM	User satisfaction; personalization	Attention visualization, rationale generation, dialogue explanations	Accuracy, F1-score, satisfaction scores	Personalization vs. transparency trade-off; multi-language requirements; real-time response needs
HR	Fair hiring; bias reduction	Feature importance, counterfactual explanations	Match accuracy, bias metrics, fairness scores	Legal liability for biased decisions; sensitive personal data; cross-cultural validity of criteria
Social Science	Content moderation; safety	SHAP, LIME, attention-based methods	Accuracy, F1-score, fairness metrics	Cultural sensitivity in hate speech detection; annotation bias; platform-specific language (173; 174)
Systematic Reviews	Reproducibility; efficiency	Rule-based extraction, feature highlighting	WSS, inter-rater agreement (κ), recall	Scalability across heterogeneous corpora; integration with existing review tools
Chatbots	User understanding; trust building	Dialogue-level explanations, attention visualization	User trust, satisfaction, task completion rate	Balancing explanation detail with conversational flow; real-time generation

- **Evaluation Gap:** A recurring challenge is the disconnect between technical metrics (fidelity, faithfulness) and practical utility. Recent work on unified benchmarks like M^4 (195) and comprehensive fidelity studies (194) addresses this gap but domain-specific validation remains essential.

Divergent Requirements. The analysis also reveals fundamental differences:

- **Risk Tolerance:** Medicine and finance operate under strict regulatory oversight with low tolerance for unexplained decisions, while CRM applications prioritize user experience over exhaustive explanations.
- **Temporal Dynamics:** Financial applications must handle rapidly changing data distributions and adversarial actors attempting to game explanations, whereas medical applications face slower concept drift but stricter validation requirements.
- **Stakeholder Diversity:** Healthcare explanations must serve clinicians, patients, and regulators simultaneously, each with different information needs and technical expertise. In contrast, HR systems primarily serve recruiters and compliance officers.
- **Explanation Granularity:** Social science applications (hate speech, misinformation) often require fine-grained token-level explanations to understand why specific phrases triggered classification, while systematic reviews benefit from document-level summaries.

Methodological Implications. These cross-domain insights suggest that a "one-size-fits-all" approach to XNLP is insufficient. Future research should develop domain-adaptive explanation frameworks that can adjust their granularity, format, and validation criteria based on the specific deployment context. The growing availability of domain-specific benchmarks and evaluation frameworks (194; 195) provides a foundation for such adaptive approaches.

The empirical evidence presented throughout these application domains reveals that explainability contributes measurably to system effectiveness. In healthcare, explainable models such as RETAIN achieve AUC of 0.8717 for heart failure prediction while maintaining interpretable attention mechanisms that clinicians can audit (197). Financial fraud detection systems combining ensemble methods with SHAP explanations report high accuracy while maintaining interpretability, demonstrating that transparency need not compromise performance (135). BERT-based sentiment classifiers in CRM contexts achieve 89–92% accuracy, with attention-enhanced variants reaching 94% F1-score (151; 152). Active learning approaches in systematic reviews achieve Work Saved over Sampling (WSS@100) of 88.5%, while bias assessment tools demonstrate inter-rater reliability ($\kappa = 0.42$–0.48) comparable to human reviewers (100; 142). Social science applications benefit from human rationales, with models trained on HateXplain achieving 0.698 accuracy and 0.687 F1-score (164). Chatbot research demonstrates that usability explains 59% of variance in user trust ($R^2 = 0.59$), highlighting the connection between explainability and user acceptance (163). In HR, documented biases in AI resume screening—with up to 18% higher rejection rates for underrepresented groups—underscore the critical role of explainability in enabling bias auditing and ensuring fair outcomes (178). Collectively, these findings establish that explainability serves not merely as a desirable feature but as a measurable contributor to the reliability, fairness, and effectiveness of NLP systems in practice.

Critical Aspects of XNLP

Evaluation Metrics for Quantifying Understanding

What does it mean to *understand* a model's output? In order to evaluate the effectiveness of XNLP techniques, it is necessary to *quantify* the level of comprehension provided by the model's explanations, typically through a blend of quantitative and qualitative metrics (198). While many of these metrics can be described conceptually, formalizing them mathematically can offer a more precise view of how such measures are actually computed.

Quantitative Metrics These approaches aim to measure how closely an explanation mirrors a model's underlying reasoning, often focusing on the following concepts:

Fidelity reflects how accurately an explanation captures the *true* behavior of a model. High-fidelity explanations yield the same predictions as the original model across an evaluation set. Formally, suppose $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ is a dataset of n instances, where x_i is an input and y_i is the model's predicted label. Let M be the black-box model and E be the explainability function (e.g., a local surrogate or rationale generator). One way to measure fidelity is to track how well the surrogate's output $\hat{y}_i = E(x_i)$ aligns with $y_i = M(x_i)$. A simplified discrete version might be:

$$\text{Fidelity}(E, M, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(x_i, y_i) \in \mathcal{D}} \mathbf{1}\left[\hat{y}_i = y_i\right], \tag{1}$$

where $\mathbf{1}[\cdot]$ is the indicator function. The higher this value, the more the explanation's outcomes match the original model's predictions. For instance, LIME (3) can approximate M locally with a simpler model to gauge this agreement.

Coherence assesses the logical consistency or readability of an explanation in natural language form. Established language metrics (e.g., BLEU (199), ROUGE (200), BERTScore (201)) are commonly used. For instance, the BLEU score between an automatically generated explanation $E(x)$ and a reference explanation $R(x)$ can be expressed as:

$$\text{BLEU} = \exp\Big(\min\big(0, 1 - \tfrac{r}{c}\big) + \sum_{n=1}^{N} w_n \ln p_n\Big), \tag{2}$$

where r is the reference length, c is the candidate length, p_n is the precision of matched n-grams, and w_n are weights (often uniform). Higher BLEU scores indicate closer alignment between generated explanations and reference texts.

Completeness determines whether an explanation includes all salient factors behind a decision. One example arises in SHAP (61), where each feature contribution ϕ_i in a set of N features is aggregated to reconstruct the model's prediction $M(x)$. A simplified Shapley-based formula is:

$$\phi_i(M, x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \Big(M(S \cup \{i\}) - M(S)\Big). \tag{3}$$

Here, completeness implies that $\sum_{i=1}^{N} \phi_i = M(x)$, capturing each feature's marginal contribution. Explanations lacking key contributions may have lower completeness scores.

Qualitative Metrics While quantitative metrics address how well the explanation aligns with the model's internal logic or reference texts, qualitative measures focus on user perceptions and real-world usability:

User trust (202) gauges how confident users feel about an explanation. It can be approached via questionnaires or controlled user tests, but is not usually reduced to a single formula. Instead, a rating scale (e.g., 1–5) is often used to approximate trust. **Satisfaction** deals with how helpful or intuitive users find an explanation (203). It may be measured as a difference in user performance or preference when they have access to explanations versus when they do not. **Transparency** captures whether users believe they understand the model's reasoning. This can be elicited via self-report items: for instance, "On a scale of 1–7, rate how well you grasp the model's decision rationale." (202). Aggregated responses create a transparency index, though it remains subjective by nature.

In practice, both quantitative and qualitative assessments are needed to create a holistic view of an explanation's quality (204; 205). However, the reliance on subjective judgments can introduce bias, and users may feel satisfied with explanations that *appear* coherent but are not causally accurate (206). This tension underscores the complexity of evaluating XNLP methods in real-world contexts. (207) emphasizes that context-aware metrics, those tailored to specific domain needs, may further improve the relevance and robustness of XNLP evaluations.

Building XNLP systems often entails balancing model performance with interpretability. Although simpler models (e.g., linear regression, decision trees) are more transparent, they may fail to capture the complexities of language data as accurately as neural networks (203). In high-stakes domains like healthcare and finance, navigating this *explainability-performance* trade-off is critical; inaccurate predictions can have serious real-world consequences, but opaque models are less likely to be trusted by stakeholders.

Time Complexity. Some XNLP methods, such as gradient-based approaches (e.g., SHAP in Eq. 3), require computational overhead comparable to the training phase itself. Others, like attention-based mechanisms, exploit existing model components to highlight important tokens with less added cost. As model sizes grow (e.g., GPT-4-level architectures), computational demands for generating explanations increase significantly (208). Scalability thus becomes a key bottleneck for widespread adoption of XNLP, particularly when handling large datasets or complex tasks.

Scalability and Standardization. Large-scale LLMs have heightened concerns about the feasibility of real-time explanations (209). Ongoing research focuses on optimizing explanation modules, exploring ways to produce faithful insights without unduly burdening memory or compute cycles (207). Meanwhile, standardized evaluation protocols are needed to compare XNLP approaches across domains, ensuring consistency in how metrics (e.g., fidelity, completeness) are applied. Moreover, effective knowledge management practices are shown to enhance software development outcomes, in part by improving software processes (210).

Fidelity versus Faithfulness: A Critical Distinction A fundamental conceptual issue in XNLP evaluation is the distinction between **fidelity** and **faithfulness** (194). While often used interchangeably, these concepts address different aspects of explanation quality:

- **Fidelity** measures whether an explanation accurately predicts the model's outputs—i.e., whether a surrogate model or attribution method can reproduce the original model's behavior on test inputs.
- **Faithfulness** asks whether the explanation reflects the *actual causal mechanisms* the model uses to arrive

Table 11. Overview of Evaluation Metrics for XNLP Techniques

Type	Metric	Description	Study
Quantitative	Fidelity	Measures how accurately the explanations reflect the model's behavior; see Eq. 1 for a simplified version.	(3)
	Coherence	Assesses clarity and logical consistency in generated explanations (e.g., BLEU in Eq. 2).	(199)
	Completeness	Evaluates whether all relevant factors are included, e.g., Shapley-based sums in Eq. 3.	(61)
Qualitative	User Trust	Measures the confidence users have in the model's explanations (often via surveys).	(202)
	Satisfaction	Gauges user acceptance and perceived usefulness; frequently collected through rating scales.	(203)
	Transparency	Evaluates whether users feel they understand the model's reasoning.	(202)

at its predictions—not merely correlations or post-hoc rationalizations.

This distinction matters greatly. An explanation can have high fidelity (accurately predicting model outputs) while having low faithfulness (not representing the true reasoning process). Recent work on Chain-of-Thought faithfulness (75; 76) demonstrates this gap: LLMs can generate plausible reasoning chains that correctly predict outputs but do not reflect their internal computations. Similarly, attention weights may correlate with model predictions without causally determining them (64).

Addressing the Out-of-Distribution Perturbation Problem A persistent challenge in evaluating explanation methods is the **out-of-distribution (OOD) perturbation problem** (211). Traditional fidelity metrics perturb input features to measure their importance, but these perturbations often create inputs that fall outside the model's training distribution, leading to unreliable assessments. (211) propose **F-Fidelity**, a robust framework that explicitly accounts for distribution shifts during evaluation, providing more reliable fidelity estimates for text classifiers.

Unified Benchmarks: Toward Standardized Evaluation The lack of standardized evaluation has long hindered progress in XNLP. Recent benchmarks address this gap:

- **M^4 (Measuring Model Interpretability via Faithfulness)** (195): A unified benchmark that evaluates explanation methods across text, image, and multimodal domains. M^4 introduces standardized faithfulness metrics and reveals that many popular explanation methods perform inconsistently across modalities, highlighting the need for domain-adaptive evaluation.
- **Comprehensive Fidelity Studies** (194): The first objective benchmark for comparing fidelity metrics themselves, rather than explanation methods. This meta-evaluation reveals significant disagreements between metrics, suggesting that multi-metric evaluation is essential.

Domain-Specific Evaluation Requirements Effective XNLP evaluation must account for domain-specific constraints and stakeholder needs. Table 12 summarizes how evaluation requirements vary across application domains.

These domain-specific requirements underscore that XNLP evaluation cannot rely solely on generic technical metrics. Healthcare applications require clinical validation that goes far beyond algorithmic fidelity, while financial applications must demonstrate robustness against adversarial manipulation. Future evaluation frameworks should incorporate domain-specific validation protocols alongside standardized technical metrics.

Rationalization Techniques and Current Challenges

As NLP models become more intricate, providing explicit *rationales* for their outputs has garnered growing attention (212; 213). Often referred to as **Rational NLP (RNLP)**, this subfield seeks to generate human-readable *explanations* of model decisions. While the foundational concept dates back to (214), the surge in interest follows the wider availability of neural methods that can produce or highlight textual justifications.

Extractive Rationalization. Techniques like LIME, Grad-CAM, or SHAP highlight the parts of the input most influential for a model's prediction (3; 145). These methods are relatively straightforward to implement but can oversimplify the decision process. As models scale (e.g., large transformers), ensuring that these extractive saliency maps remain faithful is a growing challenge (215).

Abstractive Rationalization. Future directions in rationalization include generating free-form natural language explanations that may deviate from the exact phrasing of the input text (216; 217). This approach allows for more context-rich narratives but risks producing plausible yet inaccurate summaries if not rigorously tested for alignment with the model's internal states. (218) introduced ideas like *Semantically Equivalent Adversarial Rules* to refine rationalization by ensuring the model's textual explanations genuinely mirror its decision boundary.

Despite these advances, *post-hoc* rationalizations can still be misleading if they do not reflect the true causal pathways in the model (219). Furthermore, widely used metrics often emphasize precision but rarely capture coherence, consistency, or domain relevance (182). Tackling these gaps is crucial for developing dependable rationalization frameworks.

Advances in Evaluation Metrics and Benchmarks Two large-scale meta-evaluation suites are shaping how researchers assess XNLP methods:

- **LATEC** (220): Benchmarks 17 explainers on 20 metrics across 7,500 settings, revealing frequent metric disagreements and advocating a multi-faceted evaluation approach.

Table 12. Domain-Specific Evaluation Requirements for XNLP

Domain	Primary Metrics	Validation Method	Key Challenges	Stakeholders
Medicine	Clinical validity, fidelity, actionability	Expert clinician review; clinical trial validation	Regulatory approval (FDA/EMA); integration with clinical workflows; liability concerns (124)	Clinicians, patients, regulators
Finance	Fidelity, compliance score, robustness	Regulatory audit; stress testing; adversarial evaluation	Adversarial robustness; temporal stability; real-time requirements (137)	Analysts, auditors, regulators
Social Science	Fairness, faithfulness, cultural validity	Crowdsource annotation; expert review; cross-cultural validation	Annotation bias; cultural variation in interpretation; platform specificity (173)	Researchers, moderators, users
HR	Fairness metrics, disparate impact, legal compliance	Legal review; bias audits; demographic parity testing	Legal liability; protected attribute handling; cross-jurisdictional requirements	HR professionals, legal counsel, candidates
CRM	User satisfaction, trust, engagement	A/B testing; user surveys; behavioral metrics	Real-time generation; personalization trade-offs; multilingual support	Customer service agents, end users

- **BEExAI** (221): An open leaderboard where contributors upload saliency maps or rationales, receiving a comprehensive scorecard on fidelity, interpretability, and compute cost.

On the user side, (222) compiled a systematic review of 73 user studies in XAI, highlighting the need to merge objective performance metrics with subjective user assessments. Collectively, these efforts signal a shift toward *reasoning-aware, user-centered* evaluation protocols that extend beyond single numeric proxies.

Data and Code Availability for Open Science

In parallel with technical innovations, *open science* principles significantly affect the transparency and reproducibility of XNLP studies (156; 223). Public release of datasets and code fosters rigorous validation, enabling other researchers to replicate, critique, or extend the work. Tools like Ecco (224) facilitate the interpretation of transformer activations, aligning with open-source philosophies that lower barriers for scientific collaboration. Nonetheless, partial access or commercial constraints (e.g., proprietary GPT-4 models) can hamper in-depth analysis. During this survey, many XNLP projects provided open-source code on GitHub or Zenodo, but far fewer offered publicly accessible datasets, underscoring a persistent bottleneck. (207) further highlights the value of standardized tasks and benchmarks, promoting fair comparisons across various XNLP domains.

Future Directions and Research Opportunities in XNLP

Reinforcement learning (RL) and Chain-of-Thought Reasoning. RL increasingly intersects with advanced models like GPT, creating opportunities to refine both performance and interpretability (225; 226). Chain-of-thought prompting (73), wherein a model articulates its intermediate reasoning steps, can bolster interpretability—though these intermediate rationales are not always genuinely employed by the model's internal mechanics (76). Future research may embed *faithfulness tests* into the training objective, penalizing spurious or ungrounded reasoning steps.

Hybrid Neuro-Symbolic Systems. Combining neural networks with rule-based logic canproduce interpretable, high-performing NLP pipelines. (227) illustrate a question-answering system that merges Prolog-style inferences with neural expansions, balancing explicit symbolic reasoning with the flexibility of learned representations. Similarly, NELLIE (228) uses a Prolog-like proof tree guided by an LLM retriever, demonstrating how symbolic inference can offer auditable explanations while preserving neural adaptability.

Explainable Dialogue and Social Media Analytics. (229) showcases an explainable transformer-based dialogue system capable of clarifying its reasoning, improving user trust. In social media contexts, XNLP can disclose the driving motivations behind viral content (230), interpret emotional or political trends, and combat fake news. Here, personalized or context-driven explanations can play a pivotal role in aligning NLP outputs with diverse user needs.

Personalized and Adaptive Explainability. As user demographics and tasks grow more diverse, one-size-fits-all explanations may not suffice (231). Systems like TELL-ME (232) let users specify whether they prefer analogical or factual styles, adapting future explanations accordingly. Tailoring the level of detail or communication style to match user domain expertise or cultural nuances can boost comprehension and adoption across varied environments.

Rational AI (RAI). Ongoing work in rational AI moves beyond surface-level justification, aiming to ensure that textual or visual rationales truly align with the model's underlying logic (233). Verified chain-of-thought reasoning, step-by-step evaluation, and semantically consistent rationales all converge to make model outputs not merely comprehensible but also trustworthy in real operational settings.

In sum, XNLP's future hinges on bridging performance and transparency, leveraging advanced techniques (e.g., chain-of-thought, RL) while upholding robust evaluation and open-science practices. Continued progress will likely depend on refining how we measure explanation quality, adapt them to user needs, and ensure that rationales faithfully mirror model processes.

Conclusion

In this paper, we explored the landscape of XNLP by focusing on how it can be effectively applied across multiple domains to enhance user understanding, transparency, and trust in machine learning models. We began by examining the increasing sophistication of NLP systems and the intrinsic opacity of advanced architectures such as transformers. Particularly in high-stakes sectors like healthcare and finance, understanding *why* a model predicts certain outcomes is often as crucial as the predictions themselves.

We then traced the evolution of XNLP modeling techniques, starting from traditional methods like BoW and TF-IDF and advancing to embedding-based and transformer architectures. We highlighted various interpretability approaches, attention visualization, gradient-based explanations, and rationalization methods, that aim to demystify the inner workings of NLP systems. Additionally, we demonstrated how these techniques can be employed to align complex models with end-user requirements for clarity and reliability.

From here, we examined XNLP's implementation in distinct domains:

- **Medicine:** Deployments in EHRs and clinical text analysis illustrate how XNLP can generate actionable insights for patient care. It provides interpretable outputs that can strengthen medical decision-making and build confidence among clinicians.
- **Finance:** We explored XNLP solutions for risk assessment, fraud detection, and firm valuation, underscoring the need for transparent predictive models in a domain where accountability is critical.
- **Systematic Reviews:** By explaining which studies are included or excluded, XNLP can enhance both the efficiency and clarity of evidence-based research.
- **CRM and Chatbots:** Employing XNLP to boost sentiment analysis or to create context-aware chatbotsproduces more trustworthy and user-friendly systems, emphasizing customer satisfaction.
- **Social and Behavioral Science:** We saw how XNLP can detect hate speech, sexism, misinformation, and mental health signals, thus promoting ethical and transparent analysis of social data.
- **Human Resources:** In HR, XNLP can automate tasks like recruitment or performance evaluation, offering explicit rationales that foster fair and unbiased decisions.
- **Other Applications:** Tasks like language generation, machine translation, and visual question answering demonstrate the breadth of XNLP's impact and the diverse challenges in making advanced models interpretable.

We then delved into *critical aspects* of XNLP, discussing evaluation metrics (both quantitative and qualitative), potential trade-offs in model complexity and interpretability, and rationalization methods that strive for transparency while maintaining performance. We also highlighted the pivotal role of human-in-the-loop designs—where user feedback, bias detection, and user satisfaction can refine explanations, as well as the importance of open science practices for data and code availability. Finally, we outlined future research directions, including RL integrations, chain-of-thought reasoning, hybrid neuro-symbolic architectures, explainable dialogue systems, and personalized or adaptive explainability. These frontiers show how XNLP may further develop to meet real-world needs for trustworthy, user-aligned natural language solutions.

Glossary of Terms and Abbreviations

Table 13. Common Definitions and Abbreviations

Term/Abbrev.	Definition
SVM	Support Vector Machine
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
Transformer	A neural network architecture using self-attention
BERT	Bidirectional Encoder Representations from Transformers
DistilBERT	A smaller, distilled version of BERT
XLM-R	A multilingual variant of RoBERTa
BiGRU	Bidirectional Gated Recurrent Unit
BiLSTM	Bidirectional Long Short-Term Memory
Ensemble Model	A combination of multiple models (e.g., BERT, XLM-R, DistilBERT)
GPT-4	Generative Pre-trained Transformer 4, a large language model by OpenAI
MAE	Mean Absolute Error
AUC / AUC-ROC	Area Under the ROC (Receiver Operating Characteristic) Curve
Accuracy	Ratio of correct predictions to total predictions
Precision	$\frac{\text{True Positives}}{\text{True Positives}+\text{False Positives}}$
Recall	$\frac{\text{True Positives}}{\text{True Positives}+\text{False Negatives}}$
F1-score	Harmonic mean of Precision and Recall
RMSE	Root Mean Squared Error
ROUGE	Recall-Oriented Understudy for Gisting Evaluation (a summarization metric)
XAI	Explainable Artificial Intelligence
LIME	Local Interpretable Model-agnostic Explanations
SHAP	SHapley Additive exPlanations
MLQE-PE	Machine Translation Quality Estimation - Post Editing dataset
NMT	Neural Machine Translation
Probing Methodology	Analyzing internal representations (e.g., attention) to assess model behavior
RCT	Randomized Controlled Trial
SST-2 / SST-5	Stanford Sentiment Treebank (binary/5-class)
UMNSRS	University of Minnesota Semantic Relatedness Set
WSS	Work Saved over Sampling (systematic review metric)

Research Methodology

To initiate our review, we conducted an extensive search of reputable databases such as *Google Scholar*, *IEEE Xplore*, *ACL*, and *ACM Digital Library*, using keywords related to *explainable AI* and *natural language processing*. This initial search provided a broad corpus of relevant academic work, which we subsequently refined through a **bibliometric analysis** performed with ***VOSviewer***[§]. This tool facilitated the identification of core articles, keywords, and research interconnections, thereby producing a more coherent keyword set for our dataset.

[§] https://www.vosviewer.com/

Following this, we employed the open-source tool ***ASReview***[¶], which leverages active learning, to further sift through our refined collection of papers. By iteratively screening them, ASReview helped pinpoint the most pertinent articles for each specific application domain in XNLP. This dual-stage methodology—*VOSviewer* for visual bibliometric insights and *ASReview* for targeted article retrieval—proved robust and efficient in capturing the breadth and depth of XNLP research. Below is a concise summary of the methodological steps:

1. *Data Sources:* We scoured databases like *Google Scholar*, *IEEE Xplore*, *ACL*, and *ACM Digital Library*, focusing on publications from 2018 to 2025.

2. *Search Strategy:* Relevant terms (``explainable AI'' and ``natural language processing'') were searched within titles, abstracts, and keywords.

3. *Bibliometric Analysis:* Using VOSviewer, we built co-occurrence networks of keywords and identified major thematic clusters.

4. *Initial Results:* We initially obtained 217 candidate papers, forming a provisional survey of the XNLP landscape.

5. *Data Cleaning:* Duplicates and marginally related items were removed, resulting in a streamlined dataset of 191 papers.

6. *Targeted Literature Retrieval:* Through ASReview, we systematically screened the dataset to isolate articles that offered the most relevant insights for each XNLP application.

Cross-referencing these tools (VOSviewer for bibliometric visualization and ASReview for systematic screening) offered a balanced approach that sped up the literature analysis and clarified the focal points in XNLP research. In particular, this methodology aided in revealing the primary ways XNLP is applied, the challenges observed, and the prospective frontiers for further inquiry.

Table 14 displays the final distribution of papers across different XNLP application domains, illustrating how we curated a representative yet succinct overview of current research. This approach not only streamlined the literature survey but also enabled a sharper focus on the domains covered in this review.

Table 14. Number of Final Related Papers in Different Applications

No.	Application	No. Papers
1	Medicine	26
2	Finance	22
3	Systematic Reviews	14
4	CRM	31
5	Chatbots and Conversational AI	17
6	Social and Behavioral Science	29
7	HR	18
8	Other Applications	34

Terminology

To ensure clarity throughout the paper, we define the key terms that frequently appear in our discussions:

- **Natural Language Processing (NLP):** A field of AI involving the interpretation and generation of human language by computational means.
- **XNLP:** A specific domain within NLP emphasizing transparent and interpretable machine learning models. Its purpose is to clarify how models arrive at their conclusions.
- **Rational NLP (RNLP):** An extension of NLP focusing on generating explicit logical justifications or rationales for model decisions.
- **Explainability-Performance Balance:** The tension between maximizing predictive accuracy and preserving interpretability in model design.
- **Human-in-the-Loop (HITL):** A paradigm where human feedback is integrated into model training and validation to ensure alignment with practical, ethical, or domain-specific criteria.
- **Interpretability Metrics:** Quantitative measures (e.g., fidelity, completeness) used to assess how well an explanation reflects the internal logic of a model or informs end-users.
- **Explanatory Visualization:** Approaches (e.g., attention heatmaps, saliency maps) that visually clarify which features or tokens influence a model's decision-making process.

Together, these definitions and methodological steps underlie our exploration of XNLP across various application domains, shaping the paper's discussion on explainability, transparency, and the evolving research directions in this field.

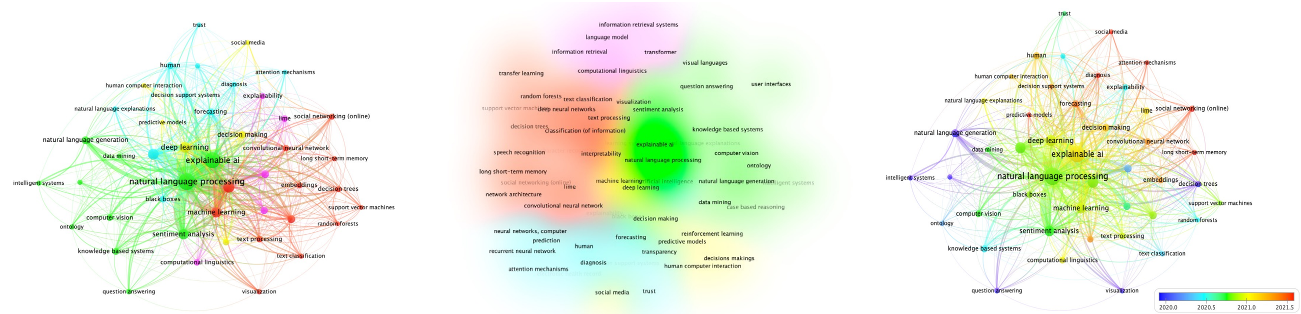
References

[1] Devlin J, Chang MW, Lee K et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*. ACL Anthology. URL https://api.semanticscholar.org/CorpusID:52967399.

[2] Tyagi AK and Bhushan B. Demystifying the role of natural language processing (NLP) in smart city applications: Background, motivation, recent advances, and future research directions. *Wireless Personal Communications* 2023; 130: 857–908. DOI:10.1007/s11277-023-10312-8.

[3] Ribeiro MT, Singh S and Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16, New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-4232-2, pp. 1135–1144. DOI: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

[¶] https://asreview.nl/



(a) Keyword Network **(b)** Item Density **(c)** Overlay Visualization

Figure 3. VOSviewer-based visualizations of the bibliometric data: **(a)** Network of co-occurring keywords, **(b)** Item density revealing concentrated regions of research interest, **(c)** Overlay map where color corresponds to average publication year (blue = earlier, red = more recent).

[4] Guidotti R, Monreale A, Ruggieri S et al. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 2018; 51(5): 1–42.

[5] Oniani D, Wu X, Visweswaran S et al. Enhancing large language models for clinical decision support by incorporating clinical practice guidelines. *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)* 2024; : 694–702URL https://api.semanticscholar.org/CorpusID:267069462.

[6] Li B, Meng T, Shi X et al. Meddm: Llm-executable clinical guidance tree for clinical decision-making. *ArXiv* 2023; abs/2312.02441. URL https://api.semanticscholar.org/CorpusID:265658947.

[7] Zhao WX, Zhou K, Li J et al. A survey of large language models. *ArXiv* 2023; abs/2303.18223. URL https://api.semanticscholar.org/CorpusID:257900969.

[8] Rudin C and Radin J. Why are we using black box models in ai when we don't need to? a lesson from an explainable ai competition. *Harvard Data Science Review* 2019; 1(2): 1–9.

[9] Bolukbasi T, Chang KW, Zou JY et al. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 2016; 29.

[10] Caliskan A, Bryson JJ and Narayanan A. Semantics derived automatically from language corpora contain human-like biases. *Science* 2017; 356(6334): 183–186.

[11] Barocas S, Hardt M and Narayanan A. *Fairness in machine learning*. NIPS Tutorial, 2016.

[12] Raghavan M, Barocas S, Kleinberg J et al. Mitigating bias in algorithmic hiring: Evaluating claims and practices. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* 2020; : 469–481DOI:10.1145/3351095.3372828.

[13] Igoe K. Algorithmic bias in health care exacerbates social inequities—how to prevent it. *Executive and Continuing Professional Education* 2021; .

[14] Kozodoi N, Jacob J and Lessmann S. Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research* 2022; 297(3): 1083–1094. DOI:10.1016/j.ejor.2021.06.023.

[15] Gurrapu S, Kulkarni A, Huang L et al. Rationalization for explainable nlp: A survey. *Frontiers in Artificial Intelligence* 2023; 6.

[16] Qian K, Danilevsky M, Katsis Y et al. Xnlp: A living survey for xai research in natural language processing. In *26th International Conference on Intelligent User Interfaces-Companion*. pp. 78–80.

[17] Jain R, Kumar A, Nayyar A et al. Explaining sentiment analysis results on social media texts through visualization. *Multimedia Tools and Applications* 2023; : 1–17.

[18] Gramegna A and Giudici P. Shap and lime: an evaluation of discriminative power in credit risk. *Frontiers in Artificial Intelligence* 2021; 4: 752558.

[19] Tenney I, Wexler J, Bastings J et al. The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models. In Liu Q and Schlangen D (eds.) *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: ACL Anthology, pp. 107–118. DOI:10.18653/v1/2020.emnlp-demos.15. URL https://aclanthology.org/2020.emnlp-demos.15.

[20] Islam MR, Ahmed MU, Barua S et al. A systematic review of explainable artificial intelligence in terms of different application domains and tasks. *Applied Sciences* 2022; 12(3): 1353.

[21] Zirikly A, Resnik P, Uzuner O et al. Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology*. pp. 24–33.

[22] Černevičienė J and Kabašinskas A. Explainable artificial intelligence (xai) in finance: a systematic literature review. *Artificial Intelligence Review* 2024; 57(8): 216.

[23] Tjoa E and Guan C. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems* 2020; 32(11): 4793–4813.

[24] Arrieta AB, Díaz-Rodríguez N, Del Ser J et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion* 2020; 58: 82–115.

[25] Puiutta E and Veith EM. Explainable reinforcement learning: A survey. In *International cross-domain conference for machine learning and knowledge extraction*. Springer, pp. 77–95.

[26] Ali S, Abuhmed T, El-Sappagh S et al. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion* 2023;

99: 101805.

[27] Danilevsky M, Qian K, Aharonov R et al. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, pp. 447–459.

[28] Heap B, Bain M, Wobcke W et al. Word vector enrichment of low frequency words in the bag-of-words model for short text multi-class classification problems 2017; abs/1709.05778. URL https://api.semanticscholar.org/CorpusID:29143028.

[29] Liu Z, Lin Y and Sun M. *Representation learning for natural language processing*. Springer Nature, 2020. ISBN 9811555737.

[30] Shimomoto EK, Souza LS, Gatto BB et al. Text classification based on word subspace with term-frequency. In *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE. ISBN 1-5090-6014-6, pp. 1–8.

[31] Zubiaga A. Tf-cr: Weighting embeddings for text classification 2020; abs/2012.06606. URL https://api.semanticscholar.org/CorpusID:229156006.

[32] DAGAN I, MARCUS S and MARKOVITCH S. Contextual word similarity and estimation from sparse data. *Comput speech lang (Print)* 1995; 9(2): 123–152.

[33] Jurafsky D and Martin JH. Vector semantics and embeddings. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* 2019; : 270–85.

[34] Song C and Raghunathan A. Information leakage in embedding models. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*. pp. 377–390.

[35] Mikolov T, Sutskever I, Chen K et al. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.

[36] GloVe: Global Vectors for Word Representation. Doha, Qatar.

[37] Sileo D and Moens MF. Analysis and prediction of NLP models via task embeddings. In Calzolari N, Béchet F, Blache P et al. (eds.) *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 633–647. URL https://aclanthology.org/2022.lrec-1.67.

[38] Cer D, Yang Y, Kong Sy et al. Universal sentence encoder for English. In Blanco E and Lu W (eds.) *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: ACL Anthology, pp. 169–174. DOI:10.18653/v1/D18-2029. URL https://aclanthology.org/D18-2029.

[39] Vogelsang DC and Erickson BJ. Magician's corner: 6. TensorFlow and TensorBoard. *Radiology: Artificial Intelligence* 2020; 2(3): e200012. DOI:10.1148/ryai.2020200012.

[40] Li Q, Njotoprawiro KS, Haleem H et al. Embeddingvis: A visual analytics approach to comparative network embedding inspection. In *2018 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, pp. 48–59.

[41] Van der Maaten L and Hinton G. Visualizing data using t-SNE. *Journal of machine learning research* 2008; 9(11).

[42] Jolliffe IT and Cadima J. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences* 2016; 374(2065): 20150202.

[43] Braşoveanu AM and Andonie R. Visualizing and explaining language models. In *Integrating Artificial Intelligence and Visualization for Visual Knowledge Discovery*. Springer, 2022. pp. 213–237.

[44] Wang Y. Single training dimension selection for word embedding with pca 2019; abs/1909.01761. URL https://api.semanticscholar.org/CorpusID:202541526.

[45] Simonyan K, Vedaldi A and Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR* 2013; abs/1312.6034. URL https://api.semanticscholar.org/CorpusID:1450294.

[46] Bahdanau D, Cho K and Bengio Y. Neural machine translation by jointly learning to align and translate. *CoRR* 2014; abs/1409.0473. URL https://api.semanticscholar.org/CorpusID:11212020.

[47] Vaswani A, Shazeer N, Parmar N et al. Attention is all you need. *Advances in neural information processing systems* 2017; 30.

[48] Sonkar S, Waters AE and Baraniuk R. Attention word embedding. In *International Conference on Computational Linguistics*. URL https://api.semanticscholar.org/CorpusID:219176825.

[49] Conklin AC, Nishi H, Schlamp F et al. Meta-analysis of smooth muscle lineage transcriptomes in atherosclerosis and their relationships to in vitro models. *Immunometabolism* 2021; 3(3).

[50] Wattenberg M, Viégas F and Johnson I. How to use t-sne effectively. *Distill* 2016; 1(10): e2.

[51] Kalyan KS, Rajasekharan A and Sangeetha S. Ammus : A survey of transformer-based pretrained models in natural language processing 2021; abs/2108.05542. URL https://api.semanticscholar.org/CorpusID:236987275.

[52] McCann B, Keskar NS, Xiong C et al. The natural language decathlon: Multitask learning as question answering 2018; abs/1806.08730. URL https://api.semanticscholar.org/CorpusID:49393754.

[53] Liu Y, Ott M, Goyal N et al. Roberta: A robustly optimized bert pretraining approach 2019; abs/1907.11692. URL https://api.semanticscholar.org/CorpusID:198953378.

[54] Lan Z, Chen M, Goodman S et al. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=H1eA7AEtvS.

[55] Vig J. A multiscale visualization of attention in the transformer model. *ACL 2019* 2019; : 37.

[56] Hewitt J and Liang P. Designing and interpreting probes with control tasks 2019; abs/1909.03368. URL https://api.semanticscholar.org/CorpusID:202538609.

[57] Olah C, Mordvintsev A and Schubert L. Feature visualization. *Distill* 2017; 2(11): e7.

[58] Sundararajan M, Taly A and Yan Q. Axiomatic attribution for deep networks. In *International conference on machine learning*. PMLR, pp. 3319–3328.

[59] Hinton GE, Vinyals O and Dean J. Distilling the knowledge in a neural network 2015; abs/1503.02531. URL https://api.semanticscholar.org/CorpusID:7200347.

[60] Ribeiro MT, Singh S and Guestrin C. Model-agnostic interpretability of machine learning 2016; abs/1606.05386. URL https://api.semanticscholar.org/CorpusID:8561410.

[61] Lundberg SM and Lee SI. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 2017; 30.

[62] Liang B, Li H, Su M et al. Deep text classification can be fooled. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. IJCAI'18, Stockholm, Sweden: AAAI Press. ISBN 9780999241127, p. 4208–4215.

[63] Murdoch WJ, Liu PJ and Yu B. Beyond word importance: Contextual decomposition to extract interactions from lstms 2018; abs/1801.05453. URL https://api.semanticscholar.org/CorpusID:25717172.

[64] Jain S and Wallace BC. Attention is not Explanation. In Burstein J, Doran C and Solorio T (eds.) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: ACL Anthology, pp. 3543–3556. DOI:10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357.

[65] Wiegreffe S and Pinter Y. Attention is not not explanation. In Inui K, Jiang J, Ng V et al. (eds.) *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: ACL Anthology, pp. 11–20. DOI:10.18653/v1/D19-1002. URL https://aclanthology.org/D19-1002.

[66] Radford A, Wu J, Child R et al. Language models are unsupervised multitask learners. *OpenAI blog* 2019; 1(8): 9.

[67] Templeton A, Conerly T, Marcus J et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread* 2024; .

[68] Cunningham H, Ewart A, Riggs L et al. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*.

[69] Bussmann B, Leask P and Nanda N. Batchtopk: A simple improvement for topk-sparseautoencoders. In *ICML 2024 Workshop on Mechanistic Interpretability*.

[70] Chanin D, Hasenzahl P, Popp S et al. Do sparse autoencoders dream of concepts? when saes fail at concept detection. *arXiv preprint arXiv:241008608* 2024; .

[71] Arya S, Rao S, Boehle M et al. B-cosification: Transforming deep neural networks to be inherently interpretable. In *38th Conference on Neural Information Processing Systems*.

[72] Fang L, Yu X, Cai J et al. Knowledge distillation and dataset distillation of large language models: Emerging trends, challenges, and future directions. *arXiv preprint arXiv:250414772* 2025; .

[73] Wei J, Wang X, Schuurmans D et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 2022; 35: 24824–24837.

[74] Barez F, Wu TY et al. Chain-of-thought is not explainability. *Oxford AI Governance Institute* 2025; .

[75] Turpin M, Michael J, Perez E et al. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., pp. 74952–74965.

[76] Lanham T, Chen A, Radhakrishnan A et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:230713702* 2023; .

[77] Lyu Q, Havaldar S, Stein A et al. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 305–329.

[78] Keles H, Keles A, Aksoy B et al. The explainability of transformers: Current status and directions. *Computers* 2024; 13(4): 92.

[79] Sun Y et al. Towards stable and explainable attention mechanisms. *IEEE Transactions on Knowledge and Data Engineering* 2025; .

[80] Covington P, Adams J and Sargin E. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*. pp. 191–198.

[81] Ledro C, Nosella A and Vinelli A. Artificial intelligence in customer relationship management: literature review and future research directions. *Journal of Business & Industrial Marketing* 2022; 37(13): 48–63.

[82] Brandl S, Bugliarello E and Chalkidis I. On the interplay between fairness and explainability. In Ovalle A, Chang KW, Cao YT et al. (eds.) *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*. Mexico City, Mexico: ACL Anthology, pp. 94–108. DOI:10.18653/v1/2024.trustnlp-1.10. URL https://aclanthology.org/2024.trustnlp-1.10.

[83] Islam MR, Ahmed MU and Begum S. Local and global interpretability using mutual information in explainable artificial intelligence. In *2021 8th International Conference on Soft Computing & Machine Intelligence (ISCMI)*. IEEE, pp. 191–195.

[84] Silver D, Schrittwieser J, Simonyan K et al. Mastering the game of go with deep neural networks and tree search. *Nature* 2016; 529(7587): 484–489.

[85] Glikson E and Woolley AW. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals* 2020; 14(2): 627–660.

[86] Choi E, Bahadori MT, Schuetz A et al. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. In *Proceedings of the 1st Machine Learning for Healthcare Conference*. PMLR, pp. 301–318. URL https://proceedings.mlr.press/v56/Choi16.html. ISSN: 1938-7228.

[87] Alabi RO, Elmusrati M, Leivo I et al. Machine learning explainability in nasopharyngeal cancer survival using lime and shap. *Scientific Reports* 2023; 13(1): 8984.

[88] Li I, Pan J, Goldwasser J et al. Neural natural language processing for unstructured data in electronic health records: A review. *Computer Science Review* 2022; 46: 100511. DOI:https://doi.org/10.1016/j.cosrev.2022.100511. URL https://www.sciencedirect.com/science/article/pii/S1574013722000454.

[89] Moradi M and Samwald M. Deep learning, natural language processing, and explainable artificial intelligence in the biomedical domain 2022; abs/2202.12678. URL https://api.semanticscholar.org/CorpusID:247154684.

[90] EliIE: An open-source information extraction system for clinical trial eligibility criteria ; 24.

[91] Weng WH, Wagholikar KB, McCray AT et al. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. *BMC Medical Informatics and Decision Making* 2017; 17(1): 155. DOI:10.1186/s12911-017-0556-8. URL https://doi.org/10.1186/s12911-017-0556-8.

[92] Fritz-Morgenthal S, Hein B and Papenbrock J. Financial Risk Management and Explainable, Trustworthy, Responsible AI. *Frontiers in Artificial Intelligence* 2022; 5. URL https://www.frontiersin.org/articles/10.3389/frai.2022.779799.

[93] Wallace E, Wang Y, Li S et al. Do nlp models know numbers? probing numeracy in embeddings. In *Conference on Empirical Methods in Natural Language Processing*. URL https://api.semanticscholar.org/CorpusID:202583694.

[94] Psychoula I, Gutmann A, Mainali P et al. Explainable Machine Learning for Fraud Detection. *Computer* 2021; 54(10): 49–59. DOI:10.1109/MC.2021.3081249. Conference Name: Computer.

[95] Varshney KR and Alemzadeh H. On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. *Big data* 2017; 5(3): 246–255.

[96] Cirqueira D, Helfert M and Bezbradica M. Towards design principles for user-centric explainable AI in fraud detection. In *Artificial Intelligence in HCI: Second International Conference, AI-HCI 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings*. Springer, pp. 21–40.

[97] Rizinski M, Peshov H, Mishev K et al. Sentiment analysis in finance: From transformers back to explainable lexicons (xlex). *IEEE Access* 2024; .

[98] Bagga P and Stathis K. Towards explainable strategy templates using nlp transformers 2023; abs/2311.14061. URL https://api.semanticscholar.org/CorpusID:265445738.

[99] Rosemblat G, Fiszman M, Shin D et al. Towards a characterization of apparent contradictions in the biomedical literature using context analysis. *J Biomed Inform* 2019; 98: 103275. DOI:10.1016/j.jbi.2019.103275.

[100] Marshall IJ, Kuiper J and Wallace BC. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association* 2016; 23(1): 193–201.

[101] Capuano N, Greco L, Ritrovato P et al. Sentiment analysis for customer relationship management: an incremental learning approach. *Applied Intelligence* 2021; 51: 3339–3352.

[102] Du M, Liu N and Hu X. Techniques for interpretable machine learning. *Communications of the ACM* 2019; 63(1): 68–77.

[103] Bacco L, Cimino A, Dell'Orletta F et al. Explainable sentiment analysis: a hierarchical transformer-based extractive summarization approach. *Electronics* 2021; 10(18): 2195.

[104] Bar-Haim R, Eden L, Friedman R et al. From arguments to key points: Towards automatic argument summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: ACL Anthology, pp. 4029–4039. DOI:10.18653/v1/2020.acl-main.371. URL https://aclanthology.org/2020.acl-main.371.

[105] Jenneboer L, Herrando C and Constantinides E. The impact of chatbots on customer loyalty: A systematic literature review. *Journal of theoretical and applied electronic commerce research* 2022; 17(1): 212–229.

[106] COPPO F and GUIDOTTI A. Artificial intelligence: analysis and evolution of the international startup ecosystem 2019; .

[107] Zhang S, Yao L, Sun A et al. Deep learning based recommender system: A survey and new perspectives. *ACM computing surveys (CSUR)* 2019; 52(1): 1–38.

[108] Mozafari M, Farahbakhsh R and Crespi N. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one* 2020; 15(8): e0237861.

[109] Saleh H, Alhothali A and Moria K. Detection of hate speech using bert and hate speech word embedding with deep model. *Applied Artificial Intelligence* 2023; 37(1): 2166719.

[110] Plaza L, Carrillo-de Albornoz J, Morante R et al. Overview of exist 2023: sexism identification in social networks. In *European Conference on Information Retrieval*. Springer, pp. 593–599.

[111] Kirk HR, Yin W, Vidgen B et al. Semeval-2023 task 10: Explainable detection of online sexism 2023; abs/2303.04222. URL https://api.semanticscholar.org/CorpusID:257405434.

[112] Mohammadi H, Giachanou A, Bagheri A et al. Towards robust online sexism detection: a multi-model approach with bert, xlm-roberta, and distilbert for exist 2023 tasks. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023)*, volume 3497. CEUR Workshop Proceedings, pp. 1000–1011.

[113] Mohammadi H, Giachanou A and Bagheri A. A transparent pipeline for identifying sexism in social media: Combining explainability with model prediction. *Applied Sciences* 2024; 14(19): 8620.

[114] Anjum and Katarya R. Hate speech, toxicity detection in online social media: a recent survey of state of the art and opportunities. *International Journal of Information Security* 2023; : 1–32.

[115] Capuano N, Fenza G, Loia V et al. Content based fake news detection with machine and deep learning: a systematic review. *Neurocomputing* 2023; .

[116] Mohammadi H, Giachanou A and Bagheri A. AI-Generated Text Detection Using Ensemble and Combined Model Training, 2023. DOI:10.5281/zenodo.10079010. URL https://doi.org/10.5281/zenodo.10079010.

[117] Patwardhan N, Marrone S and Sansone C. Transformers in the real world: A survey on nlp applications. *Information* 2023; 14(4): 242.

[118] Jim JR, Talukder MAR, Malakar P et al. Recent advancements and challenges of nlp-based sentiment analysis: A state-of-the-art review. *Natural Language Processing Journal* 2024; : 100059.

[119] Chang KH. Natural language processing: Recent development and applications. *Applied Sciences* 2023; 13(20): 11395. DOI:10.3390/app132011395.

[120] Herrewijnen E, Nguyen D, Van Deemter K et al. Human-annotated rationales and explainable text classification: A survey. *Frontiers in Artificial Intelligence* 2023; 7: 1260952.

[121] Longo L. *Explainable Artificial Intelligence: First World Conference, xAI 2023, Lisbon, Portugal, July 26–28, 2023, Proceedings, Part II*. Springer Nature, 2023.

[122] Yang K, Ji S, Zhang T et al. Towards interpretable mental health analysis with large language models. In Bouamor H, Pino J and Bali K (eds.) *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: ACL Anthology, pp. 6056–6077. DOI:10.18653/v1/2023.emnlp-main.370. URL https://aclanthology.org/2023.emnlp-main.370.

[123] Moreno-Sánchez PA. Improvement of a prediction model for heart failure survival through explainable artificial intelligence. *Frontiers in Cardiovascular Medicine* 2023; 10: 1219586. DOI:10.3389/fcvm.2023.1219586.

[124] Nazir A et al. Explainable ai in clinical decision support systems: A meta-analysis of methods, applications, and usability challenges. *Healthcare* 2025; 13(17): 2154.

[125] Noor K et al. Unveiling explainable ai in healthcare: Current trends, challenges, and future directions. *WIREs Data Mining and Knowledge Discovery* 2025; .

[126] Mullenbach J, Wiegreffe S, Duke J et al. Explainable prediction of medical codes from clinical text. In Walker M, Ji H and Stent A (eds.) *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: ACL Anthology, pp. 1101–1111. DOI:10.18653/v1/N18-1100. URL https://aclanthology.org/N18-1100.

[127] Zhang Y, Chen Q, Yang Z et al. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific data* 2019; 6(1): 52.

[128] Yang L, Wang X, Guo Q et al. Deep learning based multimodal progression modeling for alzheimer's disease. *Statistics in Biopharmaceutical Research* 2021; 13(3): 337–343.

[129] Abgrall G, Holder AL, Chelly Dagdia Z et al. Should ai models be explainable to clinicians? *Critical Care* 2024; 28(1): 301.

[130] Bagheri A, Giachanou A, Mosteiro P et al. *Natural Language Processing and Text Mining (Turning Unstructured Data into Structured)*. Cham: Springer International Publishing. ISBN 978-3-031-36678-9, 2023. pp. 69–93. DOI:10.1007/978-3-031-36678-9_5. URL https://doi.org/10.1007/978-3-031-36678-9_5.

[131] Wang L, Cheng Y, Xiang A et al. Application of natural language processing in financial risk detection 2024; abs/2406.09765. URL https://api.semanticscholar.org/CorpusID:270522037.

[132] Demajo LM, Vella V and Dingli A. Explainable ai for interpretable credit scoring. *arXiv preprint arXiv:201203749* 2020; .

[133] Misheva BH, Osterrieder J, Hirsa A et al. Explainable ai in credit risk management. *arXiv preprint arXiv:210300949* 2021; .

[134] Rudin C, Chen C, Chen Z et al. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys* 2022; 16: 1–85.

[135] Hassan M et al. Financial fraud detection using explainable ai and stacking ensemble methods. *arXiv preprint arXiv:250510050* 2025; .

[136] Hassan M et al. Model-agnostic explainable ai methods in finance: Challenges and directions. *Artificial Intelligence Review* 2025; .

[137] Giudici P and Raffinetti E. Explainable artificial intelligence in finance: A systematic review. *Artificial Intelligence Review* 2025; 57(1): 1–38.

[138] Abro AA, Talpur MSH and Jumani AK. Natural language processing challenges and issues: A literature review. *Gazi University Journal of Science* 2023; : 1–1.

[139] Ghai B, Liao QV, Zhang Y et al. Explainable active learning (xal) toward ai explanations as interfaces for machine teachers. *Proceedings of the ACM on Human-Computer Interaction* 2021; 4(CSCW3): 1–28.

[140] Chen J, Song L, Wainwright M et al. Learning to explain: An information-theoretic perspective on model interpretation. In *International conference on machine learning*. PMLR, pp. 883–892.

[141] Teijema JJ, de Bruin J, Bagheri A et al. Large-scale simulation study of active learning models for systematic reviews. *International Journal of Data Science and Analytics* 2025; : 1–22.

[142] Ferdinands G, Schrama R, de Bruin A Jonathan and"; Bagheri et al. Performance of active learning models for screening prioritization in systematic reviews: A simulation study into the average time to discover relevant records. *Systematic Reviews* 2023; 12: 100. DOI:10.1186/s13643-023-02257-7.

[143] Burgard DR and Doyle MP. Active learning models to screen articles as part of a systematic review of literature on digital tools in food safety. *Journal of Food Protection* 2025; 88(3): 100448. DOI:10.1016/j.jfp.2025.100448.

[144] Nye B, Li JJ, Patel R et al. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018. NIH Public Access, p. 197.

[145] Lei T, Barzilay R and Jaakkola T. Rationalizing neural predictions. In Su J, Duh K and Carreras X (eds.) *Proceedings of the 2016 Conference on Empirical Methods*

in Natural Language Processing. Austin, Texas: ACL Anthology, pp. 107–117. DOI:10.18653/v1/D16-1011. URL https://aclanthology.org/D16-1011.

[146] Lage I, Ross A, Gershman SJ et al. Human-in-the-Loop Interpretability Prior. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/0a7d83f084ec258aefd128569dda03d7-Abstract.html.

[147] Tsai CP, Yeh CK and Ravikumar P. Faith-shap: The faithful shapley interaction index. *Journal of Machine Learning Research* 2023; 24(94): 1–42.

[148] Rana A, Khanna D, Singh M et al. Rerrfact: Reduced evidence retrieval representations for scientific claim verification 2022; abs/2202.02646. URL https://api.semanticscholar.org/CorpusID:246634344.

[149] Yu M, Zhang Y, Chang S et al. Understanding interlocking dynamics of cooperative rationalization. *Advances in Neural Information Processing Systems* 2021; 34: 12822–12835.

[150] Antognini D and Faltings B. Rationalization through concepts. *ArXiv* 2021; abs/2105.04837. URL https://api.semanticscholar.org/CorpusID:234357794.

[151] Alsulami B et al. Enhancing product design through ai-driven sentiment analysis of amazon reviews using bert. *Algorithms* 2024; 17(2): 59. DOI:10.3390/a17020059.

[152] Li W et al. A hybrid transformer and attention based recurrent neural network for robust and interpretable sentiment analysis of tweets. *Scientific Reports* 2024; 14: 24567. DOI:10.1038/s41598-024-76079-5.

[153] OpenAI. Gpt-4o mini: Advancing cost-efficient intelligence, 2024. URL https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence. Accessed: 2024-08-04.

[154] OpenAI. Hello gpt-4o, 2024. URL https://openai.com/index/hello-gpt-4o/. Accessed: 2024-08-04.

[155] Guerreiro J and Loureiro SMC. Human-computer interaction in customer service: The experience with AI chatbots—a systematic literature review. *Electronics* 2022; 11(10): 1579. DOI:10.3390/electronics11101579.

[156] Seminck O. Conversational ai: Dialogue systems, conversational agents, and chatbots by michael mctear. *Computational Linguistics* 2023; 49(1): 257–259.

[157] Miglani V, Yang A, Markosyan A et al. Using captum to explain generative language models. In Tan L, Milajevs D, Chauhan G et al. (eds.) *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*. Singapore: ACL Anthology, pp. 165–173. DOI:10.18653/v1/2023.nlposs-1.19. URL https://aclanthology.org/2023.nlposs-1.19.

[158] Soni A and Dubey S. The impact of ai-powered chatbots on customer satisfaction in e-commerce marketing (tam approach). *Journal of Public Relations and Advertising* 2024; 3(1): 12–18.

[159] Cheng Y and Jiang H. The effects of chatbot service recovery with emotion words on customer satisfaction, repurchase intention, and positive word-of-mouth. *Frontiers in Psychology* 2022; 13: 922503. DOI:10.3389/fpsyg.2022.922503.

[160] Følstad A and Brandtzæg PB. Users' experiences with chatbots: findings from a questionnaire study. *Quality and User Experience* 2020; 5(1): 3. DOI:10.1007/s41233-020-00033-2.

[161] Ghazvininejad M, Brockett C, Chang MW et al. A knowledge-grounded neural conversation model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

[162] Li M et al. Trust in ai chatbots: A systematic review. *Computers in Human Behavior* 2025; 152: 108059. DOI:10.1016/j.chb.2025.108059.

[163] Gu Y et al. Trust in the chatbot: A semi-human relationship. *Future Business Journal* 2023; 9: 288. DOI:10.1186/s43093-023-00288-z.

[164] Mathew B, Saha P, Yimam SM et al. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35. pp. 14867–14875.

[165] Fivez P, Daelemans W, Van de Cruys T et al. The clin33 shared task on the detection of text generated by large language models. *Computational Linguistics in the Netherlands Journal* 2024; 13: 233–259.

[166] Zhang G, Giachanou A and Rosso P. Scenefnd: Multimodal fake news detection by modelling scene context information. *Journal of Information Science* 2024; 50(2): 355–367.

[167] Han N, Li S, Huang F et al. Sensing psychological well-being using social media language: Prediction model development study. *Journal of Medical Internet Research* 2023; 25: e41823.

[168] Ahmed T, Ivan S, Munir A et al. Decoding depression: Analyzing social network insights for depression severity assessment with transformers and explainable ai. *Natural Language Processing Journal* 2024; 7: 100079. DOI:https://doi.org/10.1016/j.nlp.2024.100079. URL https://www.sciencedirect.com/science/article/pii/S294971912400027X.

[169] Ali RH, Pinto G, Lawrie E et al. A large-scale sentiment analysis of tweets pertaining to the 2020 us presidential election. *Journal of big Data* 2022; 9(1): 79.

[170] Wang B, Ma J, Lin H et al. Explainable fake news detection with large language model via defense among competing wisdom. In *Proceedings of the ACM Web Conference 2024*. pp. 2452–2463.

[171] Schmidl B, Hütten T, Pigorsch S et al. Assessing the use of the novel tool claude 3 in comparison to chatgpt 4.0 as an artificial intelligence tool in the diagnosis and therapy of primary head and neck cancer cases. *European Archives of Oto-Rhino-Laryngology* 2024; 281(11): 6099–6109.

[172] Wang Y, Inkpen D and Buddhitha P. Explainable depression detection using large language models on social media data. In *9th Workshop on Computational Linguistics and Clinical Psychology*. p. 108.

[173] Gongane V, Munot M and Anuse A. Decoding fake news and hate speech: A survey of explainable ai techniques. *ACM Computing Surveys* 2025; .

[174] Rawat R, Chakraborty S et al. Explainable ai for hate speech detection: A survey. *Journal of Computational Social Science* 2024; 7: 587–623.

[175] Rawat R et al. Hate speech detection in low-resource languages: Challenges and solutions. *Social Network*

Analysis and Mining 2024; 14(1): 89.

[176] Yang Y, Kim J, Kim Y et al. Hare: Explainable hate speech detection with step-by-step reasoning 2023; abs/2311.00321. URL https://api.semanticscholar.org/CorpusID:264832755.

[177] Akkasi A. Job description parsing with explainable transformer based ensemble models to extract the technical and non-technical skills. *Natural Language Processing Journal* 2024; 9: 100102. DOI:https://doi.org/10.1016/j.nlp.2024.100102. URL https://www.sciencedirect.com/science/article/pii/S2949719124000505.

[178] Kaushal A et al. Gender, race, and intersectional bias in ai resume screening via language model retrieval. *Brookings Institution Report* 2024; URL https://www.brookings.edu/articles/gender-race-and-intersectional-bias-in-ai-resume-screening/.

[179] Mukherjee A et al. Ai-powered resume screening: A comparative study of traditional vs. ai-based recruitment methods. *International Journal of Engineering Science and Advanced Technology* 2024; 25(4): 15–28.

[180] Sheng E, Chang KW, Natarajan P et al. The woman worked as a babysitter: On biases in language generation 2019; abs/1909.01326. URL https://api.semanticscholar.org/CorpusID:202537041.

[181] DeYoung J, Jain S, Rajani NF et al. ERASER: A benchmark to evaluate rationalized NLP models. In Jurafsky D, Chai J, Schluter N et al. (eds.) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: ACL Anthology, pp. 4443–4458. DOI:10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408.

[182] Jacovi A and Goldberg Y. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Annual Meeting of the Association for Computational Linguistics*. ACL Anthology. URL https://api.semanticscholar.org/CorpusID:215416110.

[183] Ayyubi HA, Tanjim MM, McAuley J et al. Generating rationales in visual question answering 2020; abs/2004.02032. URL https://api.semanticscholar.org/CorpusID:214802897.

[184] He S, Tu Z, Wang X et al. Towards understanding neural machine translation with word importance 2019; abs/1909.00326. URL https://api.semanticscholar.org/CorpusID:202539954.

[185] Alishahi A, Chrupała G and Linzen T. Analyzing and interpreting neural networks for nlp: A report on the first blackboxnlp workshop. *Natural Language Engineering* 2019; 25(4): 543–557.

[186] Zhou W, Hu J, Zhang H et al. Towards interpretable natural language understanding with explanations as latent variables. *Advances in Neural Information Processing Systems* 2020; 33: 6803–6814.

[187] Hoover B, Strobelt H and Gehrmann S. exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer Models. In Celikyilmaz A and Wen TH (eds.) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Online: ACL Anthology, pp. 187–196. DOI:10.18653/v1/2020.acl-demos.22. URL https://aclanthology.org/2020.acl-demos.22.

[188] Tang X, Huang X, Zhang W et al. Cognitive visual commonsense reasoning using dynamic working memory. In *Big Data Analytics and Knowledge Discovery: 23rd International Conference, DaWaK 2021, Virtual Event, September 27–30, 2021, Proceedings 23*. Springer, pp. 81–93.

[189] Lakhotia K, Paranjape B, Ghoshal A et al. Fid-ex: Improving sequence-to-sequence models for extractive rationale generation. In *Conference on Empirical Methods in Natural Language Processing*. URL https://api.semanticscholar.org/CorpusID:229923145.

[190] Wiegreffe S, Marasović A and Smith NA. Measuring association between labels and free-text rationales. In *Conference on Empirical Methods in Natural Language Processing*. URL https://api.semanticscholar.org/CorpusID:225068329.

[191] Plyler M, Green M and Chi M. Making a (counterfactual) difference one rationale at a time. *Advances in Neural Information Processing Systems* 2021; 34: 28701–28713.

[192] Fomicheva M, Specia L and Aletras N. Translation error detection as rationale extraction. In Muresan S, Nakov P and Villavicencio A (eds.) *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: ACL Anthology, pp. 4148–4159. DOI:10.18653/v1/2022.findings-acl.327. URL https://aclanthology.org/2022.findings-acl.327.

[193] Chan A, Sanjabi M, Mathias L et al. Unirex: A unified learning framework for language model rationale extraction. In *International Conference on Machine Learning*. PMLR, pp. 2867–2889.

[194] Amparore E, Perotti A and Bajardi P. A comprehensive study on fidelity metrics for xai. *Information Processing & Management* 2024; 61(6): 103900.

[195] Zhang X, Wang S, Sixt L et al. M4: A unified xai benchmark for faithfulness evaluation of feature attribution methods across metrics, modalities and models. In *Advances in Neural Information Processing Systems*.

[196] Abdullah T et al. Leveraging chatgpt and explainable ai for enhancing clinical decision support. *Scientific Reports* 2025; 15.

[197] Choi E, Bahadori MT, Sun J et al. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In Lee D, Sugiyama M, Luxburg U et al. (eds.) *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2016/file/231141b34c82aa95e48810a9d1b33a79-Paper.pdf.

[198] Fan F, Xiong J, Li M et al. On interpretability of artificial neural networks: A survey. *IEEE Transactions on Radiation and Plasma Medical Sciences* 2020; 5: 741–760. URL https://api.semanticscholar.org/CorpusID:227240484.

[199] Papineni K, Roukos S, Ward T et al. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. pp. 311–318.

[200] Lin CY. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. pp. 74–81.

[201] Kaster M, Zhao W and Eger S. Global explainability of BERT-based evaluation metrics by disentangling along linguistic factors. In Moens MF, Huang X, Specia L et al. (eds.) *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: ACL Anthology, pp. 8912–8925. DOI:10.18653/v1/2021.emnlp-main.701. URL https://aclanthology.org/2021.emnlp-main.701.

[202] Lipton ZC. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 2018; 16(3): 31–57.

[203] Doshi-Velez F and Kim B. Towards a rigorous science of interpretable machine learning 2017; URL https://api.semanticscholar.org/CorpusID:11319376.

[204] Hoffman RR, Mueller ST, Klein G et al. Metrics for explainable ai: Challenges and prospects 2018; abs/1812.04608. URL https://api.semanticscholar.org/CorpusID:54577009.

[205] Mohseni S, Zarei N and Ragan ED. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2021; 11(3-4): 1–45.

[206] Miller T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 2019; 267: 1–38.

[207] Søgaard A. *Explainable natural language processing*. Morgan & Claypool Publishers, 2021.

[208] Brown T, Mann B, Ryder N et al. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc., pp. 1877–1901. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[209] Singh S, Ribeiro MT and Guestrin C. Programs as black-box explanations. *ArXiv* 2016; abs/1611.07579. URL https://api.semanticscholar.org/CorpusID:13345472.

[210] Chugh M, Chanderwal N, Upadhyay R et al. Effect of knowledge management on software product experience with mediating effect of perceived software process improvement: An empirical study for indian software industry. *Journal of Information Science* 2020; 46(2): 258–272.

[211] Xu C et al. F-fidelity: A robust framework for faithfulness evaluation of explainable ai. *arXiv preprint arXiv:241002970* 2024; .

[212] Atanasova P. *Generating Fact Checking Explanations*. Cham: Springer Nature Switzerland. ISBN 978-3-031-51518-7, 2024. pp. 83–103. DOI:10.1007/978-3-031-51518-7_4. URL https://doi.org/10.1007/978-3-031-51518-7_4.

[213] Rajani N, McCann B, Xiong C et al. Explain yourself! leveraging language models for commonsense reasoning 2019; abs/1906.02361. URL https://api.semanticscholar.org/CorpusID:174803111.

[214] Zaidan O, Eisner J and Piatko C. Using “annotator rationales” to improve machine learning for text categorization. In *Human language technologies 2007: The conference of the North American chapter of the association for computational linguistics; proceedings of the main conference*. ACL Anthology, pp. 260–267.

[215] Majumder BP, Camburu OM, Lukasiewicz T et al. Knowledge-grounded self-rationalization via extractive and natural language explanations. In *International Conference on Machine Learning*. URL https://api.semanticscholar.org/CorpusID:248963782.

[216] Gurrapu S, Huang L and Batarseh FA. Exclaim: Explainable neural claim verification using rationalization. *2022 IEEE 29th Annual Software Technology Conference (STC)* 2022; : 19–26URL https://api.semanticscholar.org/CorpusID:253629794.

[217] Sha L, Camburu OM and Lukasiewicz T. Rationalizing predictions by adversarial information calibration. *Artificial Intelligence* 2023; 315: 103828. DOI: https://doi.org/10.1016/j.artint.2022.103828. URL https://www.sciencedirect.com/science/article/pii/S0004370222001680.

[218] Ribeiro MT, Singh S and Guestrin C. Semantically Equivalent Adversarial Rules for Debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: ACL Anthology, pp. 856–865. DOI:10.18653/v1/P18-1079. URL https://aclanthology.org/P18-1079.

[219] Siddhant A and Lipton ZC. Deep Bayesian active learning for natural language processing: Results of a large-scale empirical study. In Riloff E, Chiang D, Hockenmaier J et al. (eds.) *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: ACL Anthology, pp. 2904–2909. DOI:10.18653/v1/D18-1318. URL https://aclanthology.org/D18-1318.

[220] Klein L, Lüth C, Schlegel U et al. Navigating the maze of explainable ai: A systematic approach to evaluating methods and metrics. *Advances in Neural Information Processing Systems* 2024; 37: 67106–67146.

[221] Sithakoul S, Meftah S and Feutry C. Beexai: Benchmark to evaluate explainable ai. In *World Conference on Explainable Artificial Intelligence*. Springer, pp. 445–468.

[222] Kim J, Maathuis H and Sent D. Human-centered evaluation of explainable ai applications: a systematic review. *Frontiers in Artificial Intelligence* 2024; 7: 1456486.

[223] Brinkman L, de Haan JJ, van Hemert D et al. Open science monitor 2020 utrecht university: Commissioned by the utrecht university open science programme 2021; .

[224] Alammar J. Ecco: An open source library for the explainability of transformer language models. In Ji H, Park JC and Xia R (eds.) *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*. Online: ACL Anthology, pp. 249–257. DOI:10.18653/v1/2021.acl-demo.30. URL https://aclanthology.org/2021.acl-demo.30.

[225] Li J, Monroe W, Ritter A et al. Deep reinforcement learning for dialogue generation. In Su J, Duh K and Carreras X (eds.) *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin,

Texas: ACL Anthology, pp. 1192–1202. DOI:10.18653/v1/D16-1127. URL https://aclanthology.org/D16-1127.

[226] Liu CW, Lowe R, Serban I et al. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: ACL Anthology, pp. 2122–2132. DOI:10.18653/v1/D16-1230. URL https://aclanthology.org/D16-1230.

[227] Weber L, Minervini P, Munchmeyer J et al. Nlprolog: Reasoning with weak unification for question answering in natural language. In *Annual Meeting of the Association for Computational Linguistics*. ACL Anthology. URL https://api.semanticscholar.org/CorpusID:189898046.

[228] Weir N, Clark P and Van Durme B. Nellie: a neuro-symbolic inference engine for grounded, compositional, and explainable reasoning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*. pp. 3602–3612.

[229] Sarkar S, Gaur M, Chen L et al. Towards explainable and safe conversational agents for mental health: A survey 2023; abs/2304.13191. URL https://api.semanticscholar.org/CorpusID:258332026.

[230] Bovet A and Makse HA. Influence of fake news in Twitter during the 2016 US presidential election. *Nature communications* 2019; 10(1): 7.

[231] Kuhl N, Lobana J and Meske C. Do you comply with ai?–personalized explanations of learning algorithms and their impact on employees' compliance behavior 2020; .

[232] Jeck J, Leiser F, Hüsges A et al. Tell-me: Toward personalized explanations of large language models. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. pp. 1–18.

[233] Yu M, Chang S, Zhang Y et al. Rethinking cooperative rationalization: Introspective extraction and complement control. In *Conference on Empirical Methods in Natural Language Processing*. URL https://api.semanticscholar.org/CorpusID:202235037.