Open Challenges in NLP for NFRs: A Focus on Semantics, Generalization, and Interpretability

Rrezarta Krasniqi[1]

[1]Department of Software and Information Systems, University of North Carolina at Charlotte, Charlotte, NC, USA

Abstract

Leveraging natural language processing (NLP) models within the non-functional requirements (NFR) domain has proven highly effective in addressing various issues, including automated traceability, classification of NFR compliance documents, NFR prioritization, among others. Despite these significant advancements, there remain open challenges associated with the full integration of NLP models in the NFR domain. For example, using NLP models to capture the semantics of complex phrases present in safety-critical NFRs must ensure that they do not lead to misinterpretations and potential safety risks. Therefore, this paper focuses on three key challenges related to semantic soundness, ontology generalizability, and the interpretability of model outcomes. These challenges have been chosen for several reasons. First, the absence of semantic precision can result in the misinterpretation of NFRs. Second, given that NFRs cover diverse domains, NLP models must generalize across these domains. Lastly, many problems within the NFR domain rely on decision-making based on predictions from NLP models. However, frequently adapted traditional NLP models such as ensemble models or kernel models are often regarded as 'black-boxes,' with output predictions that are challenging to interpret. Guided by these insights, we present a roadmap agenda through 10 implicit system-based scenarios drawn from the NFR perspective. These scenarios illustrate gaps where these NLP challenges become evident within the NFR domain. Additionally, we suggest solutions, strategies, and alternative approaches to better address these NLP challenges.

Keywords

Requirements Engineering, Non-Functional Requirements, Semantic Soundness, Generalizability, Interpretability

1. Introduction

In the early stages of requirements engineering, descriptions of requirements specifications often take an informal approach [1]. Typically, these requirements are expressed in natural language [2]. However, written requirements are inherently ambiguous, inconsistent and lack structure [3]. This issue becomes more evident when dealing with non-functional requirements (NFRs). For example, consider the NFR specification: 'The product shall retrieve query results in a reasonable time.' is vague and fails to describe what 'reasonable time' means. Existing traditional natural language processing (NLP) models are unable to fully draw contextual semantic inferences from such texts. Hence, this brings another issue in perspective, the need for building more refined ontologies suitable to NFRs. While various NLP-based techniques attempt to address this issue, they have concurrently become more challenging to interpret due to their complex underlying design [4]. Due to their dense internal model representations, they have almost become unusable among practitioners [5, 6]. Their internal calculations that lead to predictive outputs often resemble black-boxes [7], raising concerns about their interpretability. In light of these developments, requirements analysts must exercise caution to ensure that the output predictions generated by these models align with human reasoning. Consequently, a sole reliance on NLP predictive models can carry consequences within the NFR domain that base decisions upon those predictions. We reason that NLP models should built upon strong attributes of transparency and explain how they arrived at those output predictions. Because this is such a major challenge nowadays, the need for human-in-the-loop is a necessity rather than a choice [8]. Based on these observations, the

In: Muhammad Abbas, Fatma Başak Aydemir, Maya Daneva, Renata Guizzardi, Jens Gulden, Andrea Herrmann, Jennifer Horkoff, Marc Oriol Hilari, Sylwia Kopczyńska, Patrick Mennig, Elda Paja, Anna Perini, Alexander Rachmann, Kurt Schneider, Laura Semini, Paola Spoletini, Andreas Vogelsang. Joint Proceedings of REFSQ-2025 Workshops, Doctoral Symposium, Posters & Tools Track, and Education and Training Track. Co-located with REFSQ 2025. Barcelona, Spain, April 7, 2025.

rrezarta.krasniqi@charlotte.edu (R. Krasniqi)

0000-0001-6884-6131 (R. Krasniqi)

following research question is formulated.

What are the key challenges to leveraging existing NLP models within the NFR domain, with respect to **semantic soundness**, **ontology generalizability**, and **output interpretability**?

While other challenges within NLP models, such as efficiency and scalability problems [9, 10], do exist, they are perceived as less critical compared to the lack of semantic soundness, ontology generalizability, and interpretability that such models entail if they were to be used within NFR domain [11, 12]. We reason that the lack of semantic soundness in NLP models can lead to misinterpretations of complex phrases, such as those found in regulatory documents for safety-critical systems. Moreover, NFRs span across diverse domains, each characterized by domain-specific terminologies. In such cases, NLP models need to expand domain boundaries. hence, the lack of ontology generalizability can diminish the applicability of NLP models with the NFR domain [13]. Furthermore, because NFRs are so heterogeneous and exceed domain boundaries, they rely on predictions derived from traditional NLP models such as ensemble and/or VSM models on their final decision-makings. However, these models are often regarded as 'black-boxes' with output predictions that are challenging to interpret. Their lack of interpretability can prevent both practitioners and researchers from using them due to potential reliability implications. We explore 10 implicit, system-based scenarios (from an NFR perspective) illustrating key challenges, and then provide recommendations and solutions.

2. Background

NLP domain has advanced both from the technical scope and applicability in multiple disciplinary areas such as medicine, finance, marketing, education, machine translation, text summarization, speech recognition, and chatbots among others. Despite of its systematic progress, the unresolved issues of semantic soundness, ontology generalizability, and interpretability pose a substantial risk to its full potential for solving problems in the NFR domain. To ensure clarity, we briefly define semantic soundness, ontology generalizability, and interpretability within the NLP context.

2.1. Semantic Soundness

Semantic soundness in NLP pertains to the models' capability to accurately understand and interpret the meaning of words, phrases, and sentences in natural language [14]. Improving NLP semantic soundness is essential when analyzing intricate safety-critical regulatory documents requiring precise comprehension. Challenges in semantic soundness frequently arise due to word ambiguity, polysemy, or inconsistencies stemming from terminological usage. As a result, NLP models lack precision in discerning the semantic context of words or meaning of complex NFR phrases, that potentially lead to wrong interpretations and unsatisfactory outcomes.

2.2. Ontology Generalizability

Ontology generalizability pertains to the adaptability of NLP models across various domains and contexts [15]. From an NLP perspective, these models are frequently tailored to a particular domain or subset of domains. They provide pre-trained word-embeddings, specific terminologies and/or vocabularies that may not be applicable in many disciplines, including the conventional NFR knowledge base. As a result, NLP models often lack the necessary semantic knowledge and understanding to be effectively used within the NFR domain, highlighting the need for research into more generalizable NLP models for requirements engineering.

2.3. Model Output Interpretability

Model output interpretability pertains to the extent of transparency in the decision-making processes computed by NLP models [16]. Within the NFR domain, where NLP models play a crucial role in making critical decisions based on user inputs, the ability to interpret model output predictions becomes a human accountability concern. For example, safety standards for self-driving cars are stringent [17]. If NLP models predict issues related to sensor data, additional interpretation might be necessary to ensure the safety of passengers. The lack of interpretability can lead to a lack of trust among stakeholders, including end-users, architects, developers, and safety operators. This lack of understanding about how these models arrive at specific outcomes can question human confidence in relying on them.

3. Open NLP Challenges in the Context of NFRs

In this section, we turn our focus to three core NLP challenges and their respective shortcomings: (A) lack of semantic soundness, (B) issues with ontology generalizability, and (C) the level of interpretability inherent in the models. We present illustrative problem scenarios that span a range of 10 diverse domains, all with an emphasis on NFRs. These scenarios serve multiple purposes. First, they provide brief insights into each domain. Second, they bring a user-centric perspective. Finally, they serve as a bridging point connecting NLP challenges to the broader context of NFRs.

3.1. NLP Semantic Soundness

We explore and analyze NLP models' semantic soundness, focusing on two dimensions (1) semantic similarity and (2) semantic interpretability.

(1) Semantic Similarity—Measuring the semantic similarity of requirements is challenging because requirements are often written as short sentences, which means they contain implicit information that can only be understood in the limited context. Typically, implicit information can be viewed by the requirement analyst in a perceptive way.

• **NFR Scenario #** [18]: A financial organization is developing a trading platform, and one of the non-functional requirements is defined as *'NFR-1: The system should ensure high security for financial transactions.'* The description of NFR-1 uses a semantics that can be open to multiple interpretations. If we closely examine the term 'high security,' is open to multiple interpretations, such as encryption protocols, authentication methods, data confidentiality and integrity, compliance, authorization, authentication, availability or access controls. Current NLP models will not be able to adequately identify the security aspects crucial for financial transactions. As a result, the generated solutions might not align with the stakeholders' actual expectations.

• **NFR Scenario #2** [19]: For example, if we analyze *'R-1: The product shall preclude personal data from being printed'* and *'R-2: The system shall grant the user to print the invoice summary.'* Both R-1 and R-2 share a similar semantic meaning, as both refer to the same task ('print'). However, in the eyes of requirement analysts, R-1 and R-2 differ since R-1 conveys security matters and R2 is purely functional.

(2) Semantic Interpretability—While semantic representation models have enhanced on solving many RE tasks, including NFR ones, a deeper semantic analysis is necessary. A chief problem with most of such models relies on the usage of distributional semantics. The idea of distributional semantics is that words that appear in similar contexts tend to have similar meanings. However, inaccurate semantic relationships can skew NLP similarity scores, leading to unreliable interpretations, especially in NFR domains that rely heavily on these scores for decision-making.

• **NFR Scenario #3** [20]: Consider an NFR scenario in the context of a medical diagnosis system that uses NLP to interpret patient symptoms and provide accurate diagnoses. One of the non-functional requirements is defined as: *'NFR-3: The medical diagnosis system should accurately identify rare diseases based on patient symptoms.'* The NLP model uses distributional semantics to analyze patient records and identify symptoms associated with rare diseases. However, relying solely on context may not guarantee the necessary soundness for detecting rare diseases. Certain symptoms may be contextually similar to those of common diseases, leading to misdiagnoses and potential patient harm.

• **NFR Scenario #4** [21]: A software development team is working on an e-commerce platform with a specific NFR problem pertaining to security: *'NFR-4: The system should ensure robust protection against SQL injection attacks.'* In this scenario, the NFR-4 involves the phrase 'SQL injection attacks,' which is a critical security concern [22]. Semantic interpretability relies on contextual information from documents to understand the meaning of words, and it may identify common contexts where the phrase 'SQL injection' appears, such as discussions about web application security or external vulnerabilities. However, in these types of scenarios, NLP models that apply semantic interpretability fail to capture the depth of knowledge required to interpret 'SQL injection attacks.' This lack of semantic soundness can lead to potential misinterpretations, such as overlooking specific security measures needed to address SQL injection attacks.

• **NFR Scenario #5** [23]: A software development team is working on a new platform with an NFR requirement: *'NFR-5: The website should have low latency response times during high traffic events, ensuring smooth user experience.'* In this scenario, the NFR-5 requires the system to handle high traffic events efficiently, ensuring low latency response times. The team employs existing NLP models to interpret user feedback and performance reports to identify issues with response times during peak loads. The NLP model relying on semantic interpretability will not be able to accurately capture the context and meaning of 'slow' during 'sales.' Due to the lack of soundness in interpreting the NFR context, the NLP models will fail to identify the importance of addressing the specific performance issue during high traffic events. This can lead to subsequent delays in optimizing the website's performance during peak loads, resulting in non-satisfactory user experience during critical events.

3.2. NLP Ontology Generalizability

NLP models are often built to fit certain problems and do not apply to a wide spectrum of problems. In essence, they lack incorporating robust heterogeneous multi-models that meet diverse users' needs. These limitations impede the comprehensive extraction of knowledge from various domains, including the NFR domain. Consequently, the NLP knowledge base lacks compositional aspects both in terms of linguistic knowledge (e.g., morphological, syntactic, and lexical) and non-linguistic knowledge (i.e., pragmatic inference). That said, existing specific ontologies need to expand domain boundaries to be effectively considered for NFR cross-domain demands.

• **NFR Scenario #6** [24]: A healthcare organization is developing a medical diagnosis system to assist doctors in diagnosing various diseases. One of the non-functional requirements is *'NFR-6: The system should provide accurate and efficient diagnosis for a wide range of medical conditions.'* Existing NLP models used in medical diagnosis systems are often text-based and may not always handle other data modalities and/or capture the full complexity of several medical conditions that potentially a patient may be suffering. Hence, existing NLP models, will fail to integrate and use information from multiple sources or heterogeneous environments, subsequently leading to limited ontology generalizability.

• **NFR Scenario #7** [25]: An educational institution is developing an e-learning platform, and one of the non-functional requirements is *NFR-7: The system should support personalized learning experiences for students.'* It is evident that NFR-7 requires the e-learning platform to provide personalized learning experiences, tailoring educational content and resources to each student's individual needs and learning styles. However, current NLP models may not fully comprehend the intricacies of various subject domains and educational contexts, limiting their ability to adapt and cater to the unique learning preferences of different students effectively.

3.3. NLP Model Interpretability

In this section, our focus shifts to interpretability, emphasizing two key aspects: 'faithfulness' (i.e., does the explanation provided by the NLP model accurately represent its behavior?) and 'transparency' (i.e., is the explanation valuable for requirement analysts utilizing NLP models for decision-making?).

Often, NFRs intertwine with functional requirements, leading to subsequent changes at the implementation level [26, 27, 28]. These scenarios hinder developers from holistically understanding how NFRs interact and potentially impact functional aspects of the code [29]. This situation occasionally becomes unavoidable, particularly when NFR modifications occur at the code-level. Conversely, the adherence to good development practices for synchronizing or updating requirements at the document level is not consistently enforced by developers [30]. This naturally prompts essential questions: *'how can NLP methods effectively bridge this gap, especially in scenarios where intricate relationships exist between NFRs and FRs at the code level and lack of their documentation at requirement level?'*. Furthermore, *how can we employ NLP models to facilitate iterative refinement and clarification of NFRs when they are intertwined with functional requirements at the code level?'*. To illustrate this with a simple example, consider a scenario where security can be further refined into confidentiality and integrity. These subsets should undergo further refinement until specific design methods can be applied to satisfy and comprehend the rationale behind requirements. The challenge lies in the opacity of the NLP model's decision-making process. Mere output predictions are insufficient in revealing the explanations behind these decisions or the causal factors influencing these predictions. We assert that, for diverse NFR tasks, the interpretability of NLP models is necessary.

• **NFR Scenario #8** [31]: A software development team is building a virtual assistant system with various voice functionalities, including setting reminders, sending messages, and controlling smart home devices. One of the non-functional requirements is *'NFR-8: The virtual assistant should respond to user commands accurately and with high speed, ensuring real-time interactions.'* In this scenario, the NFR-8 of real-time interactions is closely interwoven with functional requirements for different tasks, such as setting reminders or controlling smart home devices. The challenge arises when NLP models are used to interpret user commands and generate responses for different tasks while considering the real-time interaction constraint. NLP models, especially those based on deep learning or complex neural networks, can be inherently opaque and lack interpretability. When NFRs such as real-time interactions are tightly integrated with functional requirements, it becomes challenging to discern how the NLP model processes and prioritizes tasks based on user inputs. For instance, if a user says, 'set a reminder for my meeting at 3 PM,' the NLP model needs to accurately interpret the time constraint and prioritize the task for a real-time interaction. However, due to the complexity of the NLP model, the developers might not have clear insights into how the model processes the time constraint and makes decisions about real-time interactions. This lack of transparency can lead to delays in setting the reminder or even missing the real-time interaction requirement.

• **NFR Scenario #9** [32]: A software development team is developing a chatbot to assist customers with various financial banking tasks, including account inquiries, fund transfers, and investment advice. One of the non-functional requirements is *'NFR-9: The chat-bot should ensure data privacy and compliance with financial regulations while providing seamless user interactions.'* In this scenario, the NFR-9 of data privacy and regulatory compliance is interwoven with functional requirements. The challenge arises when NLP models are used to interpret customer queries and generate responses while considering the data privacy and compliance constraints. For example, if a customer asks the chatbot to transfer funds from one account to another, the NLP model has to ensure that the transaction is executed securely and in compliance with financial regulations. Existing NLP models even the traditional ones such as Random Forest model lack transparency in their decision-making process. Consequently, for requirement analysts, comprehending how these NLP models handle sensitive information and ensure compliance with financial regulations becomes exceedingly complex. Lack of interpretability in model decisions can lead to data privacy breaches or regulatory violations, directly contradicting NFR-9.

• **NFR Scenario #10** [33]: The autonomous medical diagnosis system is designed to analyze various medical data: patient symptoms, lab test results, and medical history, to provide accurate diagnoses. *'NFR-10: The system must ensure high interpretability of its predictions to build trust among healthcare professionals and patients.'* But a patient is presented with a combination of symptoms (e.g., fever, cough, and fatigue). The NLP system predicts a possible respiratory infection, but the doctor, with years of experience, thinks the symptoms could also be indicative of a more severe condition. However, the current NLP-based models used for diagnosis lack a human-in-the-loop component. Without a human-in-the-loop component, the doctor cannot hone into the patient's medical history or conduct additional tests to validate the model's predictions.

4. Research Directions: Roadmap

The three forefront challenges that we discussed in Section 3.1, Section 4.2, and Section 3.3 have shown to have implications when they are examined from the NFR context. Thus, we provide several research directions and recommendations complemented by alternative strategies and solutions that can be adopted in a flexible manner.

4.1. Directions on NLP Semantic Soundness

The recommendations and alternative solutions that we consider for leveraging NLP models within the NFR domain when they are constrained by the challenges tied to **semantic similarity** are as follows.

Recommended Approach #1: Directly integrate contextual information into NLP models to enhance their semantic interpretability. This endeavor involves incorporating advanced contextual embeddings [34] such as BERT [35] and RoBERTa [36] into the NLP models. These embeddings are able to capture the contextual meanings of words and phrases. Subsequently, we can fine-tune the NLP models using domain-specific NFR datasets. This step is essential to adapt to the specific contextual complexities of the NFR domain. The subsequent step involves the adaptation of a domain-specific ontology that encodes the relationships between terms and concepts within the NFR domain. This ontology is instrumental in helping the NLP models comprehend the contextual connections among terms. To integrate this ontology, we can employ two techniques. The first is the GNTM [37], which creates semantic correlation graphs to capture the co-occurrence patterns of terms within NFR documents. GNTM provides underlying semantic relationships among terms in diverse contexts. The second technique is the DWGTM [38], which extracts topics from semantic correlation graphs. Unlike GNTM, DWGTM explores how the semantics of NFR terms evolve across various periods or contexts. By combining GNTM and DWGTM, we can create semantic interpretability layers within the NFR context.

Recommended Approach #2: Develop domain-specific NLP models tailored explicitly to NFRs. These models would be designed with a specific awareness of the intricate linguistic and semantic structures inherent in NFR-related text. Specific steps that we recommend are the followings. Initially, we should collect a diverse and extensive NFR dataset representative of the specific domain(s) of interest. This dataset should entail a range of NFR types. Then, we should consider annotating the dataset to provide clear labels indicating the semantic relationships between different NFRs, capturing various levels of similarity and dissimilarity. Once that is done, we could develop NLP models that take into account the inherent complexities of NFRs. This could involve using advanced architectures such as transformers [34], recurrent neural networks [39], or hybrid models that incorporate both word and phrase-level embeddings [40]. Afterwards, we could train the NLP models on the domain-specific NFR dataset, fine-tuning them to understand the specific linguistic context and semantic variations present in NFRs. Finally, using a separate validation dataset, we can evaluate the domain-specific NLP models' ability to capture NFR semantic similarity and compare their performance against traditional NLP models to demonstrate the improvements.

4.2. Directions on NLP Ontology Generalizability for NFR

Human feedback and expertise are crucial, especially for understanding the relationships and rationales among NFRs, as well as their taxonomy relations with other NFRs. The same principle needs to be applied to predictive NLP models, which are designed, viewed, and treated as black-box models. However, existing NLP models have not been designed to incorporate a 'humans-in-the-loop' approach.

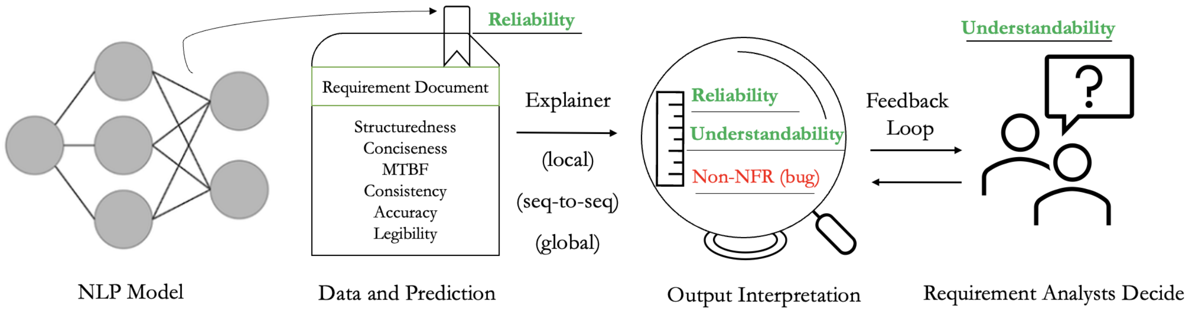

Figure 1: An NFR Example of an Interpretable NLP Model with Humans-in-the-Loop

To address the interpretability problem in the context of NFRs, we require models that can interpret output predictions and involve humans-in-the-loop for final decision-making. Frequently, we apply NLP to requirements—such a model would make probabilistic inferences as to why a specific requirement is detected as NFR-related or not. While these models often perform well, it is essential to understand how NLP models process and reach such conclusions. If we closely examine Figure 1, it illustrates a scenario where an NLP model predicts that the requirement document is related to 'Reliability' based on correlated features selected by the NLP model in the document, such as 'structuredness, conciseness, error/fault, MTBF, consistency, accuracy.' However, after the requirement analyst carefully reviews the document and its NFR-related content, they may disagree with the prediction, considering it as an 'understandability' document. This disagreement arises because most features, such as 'conciseness, structuredness, legibility, and consistency,' are subcategories of 'understandability,' despite the fact that 'consistency' is a sub-category that belongs to both 'understandability' and 'reliability.' Without explanations of how the NLP model arrived at these predictions, which cannot be solely revealed by the output predictions, insights are lacking. Based on these observations, we recommend the following.

Recommended Approach #1: Leverage word embeddings and embedding-based techniques to enhance the ontology. These techniques can enhance the ontology with richer semantic information and provide contextual meaning of NFR terms. Specifically, Word2Vec [41], GloVe [42], and/or FastText [43] can be employed to train word embeddings that encapsulate contextual meanings. Subsequently, these trained word embedding models can be applied to execute Named Entity Recognition (NER) [44] on the NFR-related dataset. NER has the capability to detect and extract NFR-related entities such as 'security,' 'performance,' 'usability,' and others from the NFR-related dataset. Following this, text mining can be utilized as a complementary technique to extract NFR-related terms both from the NFR dataset and the output of the NER process. This procedure involves identifying concepts and phrases specific to NFRs, which should then be incorporated into the NFR ontology. For example, the approach could identify the dependency relationship between 'security' and 'data encryption' and add them into the NFR ontology. This in turn, facilitates the reuse of foundational knowledge.

Recommended Approach #2: Use compositional embeddings (CE) to improve NLP ontology generalizability. CE creates phrase/sentence embeddings from individual words, capturing semantic relationships within text. A hierarchical phrase embedding model can be trained on NFR data to compose NFR-specific phrase embeddings, generalizing across domains even with limited NFR ontology coverage. For example, it could recognize the similarity between "fluctuating transportation demands" and "peak load." Contextual adaptation can further update the ontology with new NFR knowledge. This approach addresses the limitations of standard ontologies like WordNet, which often lack NFR-specific terms (e.g., "offline mode.") However, CE effectiveness depends on the quality and quantity of NFR training data and the complexity of NFR expressions.

4.3. Directions on NLP Interpretability and Human Feedback

Recommended Approach #1: Use post-hoc local explanations, such as the 'Local Interpretable Model-Agnostic Explanations' (LIME) [45]. LIME's goal is to train local models on specific instances, obtain the model's decisions for those instances, and then weigh them based on their proximity to the instance being explained by humans. This is achieved by initially perturbing the inputs and subsequently monitoring the outputs derived from this 'black box' to understand how predictive outcomes change. Incorporate global explanations by providing explanations in the context of a system's general behavior through independent 'if-then rules' [46]. With this approach, the model learns essential rules that elucidate the classification of particular instances, with these rules propagating from instance-level to class-level rules. These rules are then processed to obtain the best set. In the realm of global explanations, the primary interest is understanding how the model arrives at selecting the best set and, more crucially, how it generates the final rule set for specific instances in the dataset. Another approach to consider is incorporating Sequence-to-Sequence models for explaining NLP models [47]. The Sequence-to-Sequence approach employs a variational auto-encoder to produce meaningful input perturbations. Analyzing input variations through perturbing inputs is becoming a reliable method for generating explainable predictive models. Furthermore, there have been advances in enhancing the meaningfulness of word embeddings using word intrusion [6]. This work builds on the prior research of Chang *et al.* [48], which interprets probabilistic topic models. A common method to interpret embeddings is to enforce sparsity during training [49].

Recommended Approach #2: Build an interactive explanation. This involves implementing an interactive explanation mechanism that offers comprehensive insights into the decision-making process of the NLP model. Users should be able to pose questions and seek explanations for the model's predictions. Throughout this process, incorporate uncertainty estimation techniques to quantify the level of confidence in the model's predictions. For instance, techniques like SHAP (SHapley Additive exPlanations) [50, 51] can be employed for such estimations. SHAP is capable of providing explanations for individual predictions, allowing users to comprehend specific model outcomes. The subsequent step would entail establishing a feedback loop involving domain experts to review the model's outputs and assess its interpretability. Their feedback can shed light on any deficiencies in the model. By adopting these approaches, NLP models can be effectively integrated into the NFR domain. Both of these recommendations aim to augment user comprehension, transparency, and trust in the model's predictions, thereby enhancing the user-friendliness and reliability of NLP applications.

5. Conclusion

This paper provides insights into the research gaps at the intersection of the NLP and NFR domains, focusing on three key NLP challenges such as semantic soundness, ontology generalizability, and interpretability. By raising awareness of these challenges, researchers can exercise greater caution when leveraging NLP models as primary solutions within the NFR domain, especially in safety-critical contexts. We illustrate these challenges with ten typical scenarios. Finally, we propose a research agenda outlining recommendations and strategies to advance research within requirements engineering and potentially the broader software engineering community.

References

[1] C. J. Neill, P. A. Laplante, Requirements Engineering: The State of the Practice, IEEE Software 20 (2003) 40–45.

[2] M. Kassab, C. Neill, P. Laplante, State of Practice in Requirements Engineering: contemporary data, Innovations in Systems and Software Engineering 10 (2014) 235–241.

[3] S. F. Tjong, D. M. Berry, The Design of SREE—A Prototype Potential Ambiguity Finder for

Requirements Specifications and Lessons Learned, in: International Conference on Requirement Engineering for Software Quality, Springer, 2013, pp. 80–95.
[4] C. Rosset, Turing-NLG: A 17-billion-parameter language model by Microsoft, hhttps://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/, 2020. [Online; accessed 03/16, 2023].
[5] A. Kumar, P. Howlader, R. Garcia, D. Weiskopf, K. Mueller, Challenges in Interpretability of Neural Networks for Eye Movement Data, in: Symposium on Eye Tracking and Applications, 2020, pp. 1–5.
[6] L. K. Şenel, I. Utlu, V. Yücesoy, A. Koc, T. Cukur, Semantic Structure and Interpretability of Word Embeddings, Transection on Audio, Speech, & Language Processing 26 (2018) 1769–1779.
[7] S. Choudhary, N. Chatterjee, S. K. Saha, Interpretation of Black Box NLP Models: A Survey, arXiv:2203.17081 (2022).
[8] M. T. Ribeiro, S. Singh, C. Guestrin, "Why Should I Trust You?" Explaining the Predictions of Any Classifier, in: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
[9] K. S. Clark, Efficient and Scalable Transfer Learning for Natural Language Processing, Stanford University, 2021.
[10] W. Zhao, H. Peng, S. Eger, E. Cambria, M. Yang, Towards Scalable and Reliable Capsule Networks for Challenging NLP Applications, arXiv:1906.02829 (2019).
[11] K.-W. Chang, H. He, R. Jia, S. Singh, Robustness and Adversarial Examples in Natural Language Processing, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 22–26.
[12] I. Mollas, N. Bassiliades, I. Vlahavas, G. Tsoumakas, Lionforests: Local Interpretation of Random Forests, arXiv preprint arXiv:1911.08780 (2019).
[13] D. Dermeval, J. Vilela, I. I. Bittencourt, J. Castro, S. Isotani, P. Brito, A. Silva, Applications of Ontologies in Requirements Engineering: A Systematic Review of the Literature, Requirements Engineering 21 (2016) 405–437.
[14] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep Contextualized Word Representations, arXiv:1802.05365 (2018).
[15] F. Giunchiglia, M. Fumagalli, Entity Type Recognition–Dealing with the Diversity of Knowledge, in: International Conference on Principles of Knowledge Representation and Reasoning, volume 17, 2020, pp. 414–423.
[16] Z. C. Lipton, The Mythos of Model Interpretability: In Machine Learning, the Concept of iIterpretability is both Important and Slippery., Queue 16 (2018) 31–57.
[17] M. F. Lohmann, Liability Issues Concerning Self-Driving Vehicles, European Journal of Risk Regulation 7 (2016) 335–340.
[18] A. Stockel, Securing data and financial transactions, in: Proceedings The Institute of Electrical and Electronics Engineers. 29th Annual 1995 International Carnahan Conference on Security Technology, IEEE, 1995, pp. 397–401.
[19] A. Alessi, G. Ciccarelli, L. Cipolli, L. Guidotti, A. Marsano, A. Hanganu, Privacy by design and by default in software development in order to prevent unlawful processing of personal data. privacy certifications impact on software development and liabilities., 2021.
[20] J. A. Swets, Measuring the accuracy of diagnostic systems, Science 240 (1988) 1285–1293.
[21] J. Clarke-Salt, SQL injection attacks and defense, Elsevier, 2009.
[22] R. M. Thiyab, M. Ali, F. Basil, et al., The Impact of SQL Injection Attacks on the Security of Databases, in: 6th International Conference of Computing and Informatics, School of Computing, 2017, pp. 323–331.
[23] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, P. Tran-Gia, A survey on quality of experience of http adaptive streaming, IEEE Communications Surveys & Tutorials 17 (2014) 469–492.
[24] J. R. Ball, B. T. Miller, E. P. Balogh, Improving diagnosis in health care, National Academies Press,

2015.
[25] M. Bulger, Personalized learning: The conversations we're not having, Data and Society 22 (2016) 1–29.
[26] R. Krasniqi, Detecting scattered and tangled quality concerns in source code to aid maintenance and evolution tasks, in: 2023 IEEE/ACM 45th International Conference on Software Engineering: Companion Proceedings (ICSE-Companion), IEEE, 2023, pp. 184–188.
[27] R. Krasniqi, H. Do, Towards semantically enhanced detection of emerging quality-related concerns in source code, Software Quality Journal 31 (2023) 865–915.
[28] R. Krasniqi, H. Do, Capturing contextual relationships of buggy classes for detecting quality-related bugs, in: 2023 IEEE International Conference on Software Maintenance and Evolution (ICSME), IEEE, 2023, pp. 375–379.
[29] J. Eckhardt, A. Vogelsang, D. M. Fernández, Are" Non-functional" Requirements really Non-functional? An Investigation of Non-functional Requirements in Practice, in: International Conference on Software Engineering, 2016, pp. 1–11.
[30] I. J. Jureta, S. Faulkner, P.-Y. Schobbens, A More Expressive Softgoal Conceptualization for Quality Requirements Analysis, in: International Conference on Conceptual Modeling, Springer, 2006, pp. 281–295.
[31] M. B. Hoy, Alexa, siri, cortana, and more: an introduction to voice assistants, Medical reference services quarterly 37 (2018) 81–88.
[32] L. Anaya, A. Braizat, R. Al-Ani, Implementing ai-based chatbot: Benefits and challenges, Procedia Computer Science 239 (2024) 1173–1179.
[33] M. Sarwar Kamal, N. Dey, A. S. Ashour, Large scale medical data mining for accurate diagnosis: A blueprint, in: Handbook of large-scale distributed Computing in smart healthcare, Springer, 2017, pp. 157–176.
[34] A. Chernyavskiy, D. Ilvovsky, P. Nakov, Transformers:"the end of history" for natural language processing?, in: Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III 21, Springer, 2021, pp. 677–693.
[35] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv (2018).
[36] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, ArXiv (2019).
[37] D. Shen, C. Qin, C. Wang, Z. Dong, H. Zhu, H. Xiong, Topic Modeling Revisited: A Document Graph-based Neural Network Perspective, in: Neural IPS, 2021, pp. 1–13.
[38] Y. Wang, X. Li, X. Zhou, J. Ouyang, Extracting Topics with Simultaneous Word Co-occurrence and Semantic Correlation Graphs: Neural Topic Modeling for Short Texts, in: Association for Computational Linguistics, 2021, pp. 18–27.
[39] H. Salehinejad, S. Sankar, J. Barfett, E. Colak, S. Valaee, Recent advances in recurrent neural networks, arXiv preprint arXiv:1801.01078 (2017).
[40] M. Tang, L. Zhuang, H. Li, J. Yang, Y. Guo, Phrase-level global-local hybrid model for sentence embedding, in: Int'l Conference on Multimedia and Expo, IEEE, 2020, pp. 1–6.
[41] K. W. Church, Word2Vec, Natural Language Engineering 23 (2017) 155–162.
[42] J. Pennington, R. Socher, C. D. Manning, Glove: Global Vectors for Word Representation, in: EMNLP, 2014, pp. 1532–1543.
[43] B. Athiwaratkun, A. G. Wilson, A. Anandkumar, Probabilistic Fasttext for Multi-Sense Word Embeddings, arXiv preprint arXiv:1806.02901 (2018).
[44] X. Liu, H. Chen, W. Xia, Overview of Named Entity Recognition, Journal of Contemporary Education 6 (2022) 65–68.
[45] S. Mishra, B. L. Sturm, S. Dixon, Local Interpretable Model-Agnostic Explanations for Muusic Content Analysis, in: ISMIR, volume 53, 2017, pp. 537–543.
[46] N. Liu, X. Huang, J. Li, X. Hu, On Interpretation of Network Embedding via Taxonomy Induction, in: International Conference on Knowledge Discovery and Data Mining, 2018, pp. 1812–1820.

[47] D. Alvarez-Melis, T. S. Jaakkola, A Causal Framework for Explaining the Predictions of Black-Box Sequence-to-Sequence Models, arXiv (2017).
[48] J. Chang, S. Gerrish, C. Wang, J. Boyd-Graber, D. Blei, Reading Tea Leaves: How Humans Interpret Topic Models, Advances in Neural Information Processing Systems 22 (2009).
[49] V. Trifonov, O.-E. Ganea, A. Potapenko, T. Hofmann, Learning and Evaluating Sparse iIterpretable Sentence Embeddings, 1809.08621 (2018).
[50] Y. Nohara, K. Matsumoto, H. Soejima, N. Nakashima, Explanation of machine learning models using improved shapley additive explanation, in: Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, 2019, pp. 546–546.
[51] I. Ekanayake, D. Meddage, U. Rathnayake, A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using shapley additive explanations (shap), Case Studies in Construction Materials 16 (2022) e01059.