A Literature Review: Bridging the Gap Between Natural Language Processing and Explainable Artificial Intelligence in Language-Based AI Systems

ZAYN JAMEEL ABBAS, University of Guelph, Canada

Table of Contents

1. Introduction

Within the vast landscape of Artificial Intelligence (AI), two distinct yet complementary fields, Natural Language Processing (NLP) and Explainable AI (XAI), have emerged with the potential to reshape our understanding and interactions with intelligent systems. The goal of NLP is to enable machines to understand, interpret, and output what we call natural language (human language) (Hjorth, 2021; Nagarhalli et al., 2022; Shivahare et al., 2022). This allows for the development of various applications, including language translation, sentiment analysis, question-answering systems, and information retrieval (Nagarhalli et al., 2022). XAI seeks to clarify the decision-making process of AI models by providing comprehensible and justifiable explanations for their outcomes (Okolo et al., 2022). This literature review intends to provide a solid understanding of the current and emerging approaches in both fields and to bridge the gap between them by exploring the potential for future collaboration based on human-centric goals.

1.1 Scope

AI can accomplish various tasks across multiple domains, such as education, transportation, health care, and many more (Feng & Boyd-Graber, 2019; Hughes et al., 2020). While this is exciting and opens up doors to many opportunities, trusting the output of AI blindly is dangerous. Truly understanding the decision-making process of AI is challenging due to the lack of transparency in most models (Hughes et al., 2020; Okolo et al., 2022; Pafla, 2020). XAI aims to address this issue by providing insight into how an AI system arrives at its conclusions by offering interpretable explanations (Barredo Arrieta et al., 2020; Hughes et al., 2020). This allows users to understand the output provided, allowing them to trust the system and make a conscious decision about what to do with the information provided. The importance of XAI lies in building trust and confidence in AI, ensuring the key concepts of interpretability and transparency (Barredo Arrieta et al., 2020). Although the intent of XAI is good, this literature review will delve into the true state of XAI and what limits it from meeting these goals regarding language-based systems.

NLP, a subfield of AI, aims to enable computers to understand, interpret, and generate natural language in a way that is meaningful and useful (Danilevsky et al., 2021; Hjorth, 2021; Luekhong et al., 2019; Tsvetkov et al., 2019). Natural language is the form of communication that humans use in everyday life, whether it be spoken or written word (Danilevsky et al., 2021). NLP seeks to bridge the gap between computers and humans by using algorithms and models capable of processing and analyzing text and speech (Chaman Kumar et al., 2020; Maruf et al., 2022; Shivahare et al., 2022). Some common applications of NLP include text generation, information retrieval, question-answering systems, machine translation, and

sentiment analysis (Hjorth, 2021; Nagarhalli et al., 2022; Shivahare et al., 2022). These applications are commonly integrated into systems. For example, chatGPT by OpenAi implements text generation, information retrieval, and question & answer. Although I will briefly describe each of these applications, this review will primarily focus on machine translation and sentiment analysis.

1.2 Goal

This literature review aims to cover significant aspects of NLP translation and sentiment analysis, including individual and combined approaches and challenges, to understand their intersectionality with XAI further. To gain insights into the current state of NLP translation and sentiment analysis and their intersections with XAI, I focused on the following key questions:

1. How can XAI be used to enhance understandability, interpretability, and transparency in NLP translation and sentiment analysis systems?

To answer this question, we must first understand the current state of XAI and NLP. The first section explores the current and emerging XAI approaches, along with their limitations, with particular emphasis on methods and strategies that improve the interpretability and transparency of language-based AI systems. The next section delves into NLP translation and sentiment analysis approaches and applications. The third section further explores the limitations and challenges within NLP approaches. Lastly, the fourth section investigates the potential synergy between XAI and NLP and how they can be effectively combined to enhance transparency and interpretability in language-based AI models. The main contributions of this review are as follows:

1. Insights into best practices: Identifying and highlighting the most effective approaches in XAI and NLP can contribute to the interpretability of language-based systems.
2. Highlighting ethical concerns: The lack of interpretability and transparency can negatively impact human interaction. This reviews sheds light on potential ethical concerns related to language-based systems.
3. Bridging the gap between XAI and NLP: As both fields have traditionally developed independently, demonstrating their interconnectedness, and advocating for their integration can encourage researchers to explore this further.

2. Research Methodology

The data corpus collected for this review was primarily from the Association for Computing Machinery Digital Library. Additionally, papers were sourced from the Web of Science Index, Google Scholar and IEEE Digital Library. Papers were selected by title initially, then filtered through abstract, and lastly filtered by a complete read-through of the papers themselves. This process resulted in an accumulation of 76 total papers and articles. The Mendeley[1] online reference manager tool was used to facilitate the literature review process and to organize the data corpus.

2.1 Search Terms

Literature review search terms were generated and iterated to produce as many relevant articles

as possible. The search terms used are as follows:

"(Natural Language Processing OR NLP) AND (XAI OR Explainable AI) AND Translation AND Machine Learning AND Sentiment Analysis"

2.2 Inclusion & Exclusion Criteria

Initially, the search included papers from the past five years. However, the most recent papers were mostly from the global South and primarily focused on language translation. Due to this, I decided to expand the search to include the past 20 years, which resulted in a higher quantity of relevant papers.

3. Results

This section presents a comprehensive synthesis and analysis of the findings from the data corpus. By examining a wide range of research articles, reports, and scholarly works, this section aims to provide a clear overview of the key insights, trends, and outcomes of NLP and XAI. This is presented in four sections, as described prior, common and emerging XAI approaches, common and emerging NLP approaches, challenges and biases, and the intersection of XAI and NLP.

3.1 Common & Emerging Explainable Artificial Intelligence (XA) Approaches

As the AI field continues to expand, the demand for interpretability and transparency increases alongside it. Researchers have developed various techniques and methodologies to help better understand the decision-making process of these state-of-the-art AI systems (Hughes et al., 2020; Okolo et al., 2022; Pafla, 2020).

The capabilities of AI systems continue to grow significantly, creating waves of transformative applications across multiple domains. However, as these systems expand, so does the challenge of interpretability and transparency due to the opaque nature of machine learning (ML) systems (Hughes et al., 2020; Okolo et al., 2022; Pafla, 2020). Meaning that the decision-making process of the model is unknown to the user, as depicted in Figure 1.

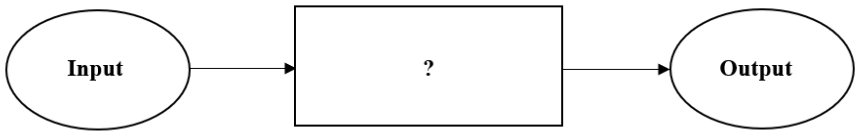

Figure 1. Opaque ML

The field of XAI has emerged to address this by providing human-understandable explanations for AI system outputs (Gohel et al., 2021; Hughes et al., 2020; Jacovi, 2023; Okolo et al., 2022; van der Waa et al., 2021). By bridging the gap between AI algorithms and human cognition, XAI has the potential to help build the trust and reliability currently missing in the field (Gohel et al., 2021; Jacovi, 2023).

Barredo et al. (2020) conducted an extensive literature review of concepts related to XAI

by analyzing multiple studies, which led to the concept of "Responsible Artificial Intelligence". They stressed that to help make AI systems, ethical principles need to be embedded into AI applications and processes. XAI can help build trust and convey the message of responsibility through the transparency of these systems. They go into extensive detail in defining terminology and determine the transparency of current AI models based on the following criteria: simulatability (the ability of a model to be simulated by a human), decomposability (the ability to break down a model and explain each of its parts), and algorithmic transparency (the ability for the user to understand the model's algorithm).

The models deemed transparent tended to be based on Rule-Based algorithms such as Decision Trees or Linear Regression. They found that the following models lack transparency based on their criteria: Tree Ensembles, Support Vector Machines (SVM), Multi-Layer Neural Networks, Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs). They propose post-hoc approaches to make these models more interpretable and transparent. Post-hoc approaches are based on the idea of generating explanations by humans or from data after the AI has already produced an output (van der Waa et al., 2021). These post-hoc approaches include visualization, local explanations, model simplification, feature relevance, text explanations, and explanations by example. Table 1 summarizes the recommended approaches to improve the transparency of these models. While there are many other XAI approaches, I will primarily explain these approaches as they are the most commonly seen in the data corpus.

Model	**Post-Hoc Approach Recommended**					
	Visualization	**Local Explanations**	**Model Simplification**	**Feature Relevance**	**Text Explanations**	**Explanations by Example**
Tree Ensembles	✗	✗	✓	✓	✗	✗
SVM	✗	✓	✓	✗	✗	✗
Multi-Layer Neural Networks	✓	✗	✓	✓	✗	✗
CNNs	✓	✗	✗	✓	✗	✗
RNNs	✗	✗	✗	✓	✗	✗

Table 1. XAI Approaches

3.1.1 Visualization

This approach refers to using visual representations, whether it be graphs, diagrams or images, to help users understand the decision-making process of an AI system (Barredo Arrieta et al., 2020; Hudon et al., 2021). Current XAI focuses on explaining the math behind AI models, which neglects the real-world users as there is an assumed understanding of algorithms or concepts used to explain the model's behaviour (Barredo Arrieta et al., 2020; Hudon et al., 2021; Okolo et al., 2022; van der Waa et al., 2021).

Hudon et al. (2021) conducted a study in an attempt to provide human-centred explanations using visualizations of the AI system's decision-making model. They wanted to

see if the visualization of explanations would affect a user's cognitive load and confidence in the AI system by presenting the user with an AI system that identifies animals and explains how it is doing so. They based this on the Cognitive Fit theory, which suggests that when the information presented matches the task, people are more likely to understand it more accurately (Vessey, 1991). The study involved measuring the user's pupil dilation to indicate when there was more cognitive function. Figure 2 shows their task design and process. They found that presentation order and visualization explanation affect cognitive load. They initially hypothesized that simpler visualization explanations would lead to a lower cognitive load and a higher confidence rate. While this was proven true, they interestingly found a negative correlation between cognitive load and confidence rates. In one experiment, where they used a more complex explanation, the users had higher confidence in the AI system. The authors hypothesize that this may be because all the information helped the users identify and understand the decision-making process of the system; however, further research is needed to investigate this.

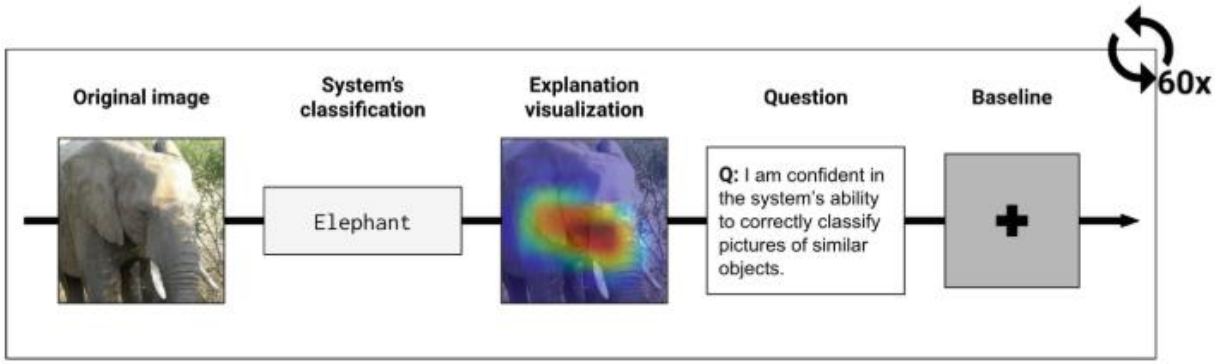

Figure 2. Task Design. (Ref: Hudon et al., 2021)

3.1.2 Local Explanations

Local Explanations can explain the prediction made at the specific time instead of the general behaviour of a model (Barredo Arrieta et al., 2020). The most common Local Explanation methods include SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and Gradient-Weighted Class Activation Mapping (Grad-CAM) (Le et al., 2023; Okolo et al., 2022).

SHAP uses a game-theory approach to explain how individual predictions are generated regarding the model's input variables (*LIME vs. SHAP: Which Is Better for Explaining Machine Learning Models?*, 2021; Panati et al., 2022). Unlike many Local Explanation methods, SHAP can be globally interpretable because it explains the behaviour of a model based on its inputs (*LIME vs. SHAP: Which Is Better for Explaining Machine Learning Models?*, 2021).

LIME explains the individual prediction of a model by approximating it locally with an interpretable model (*LIME vs. SHAP: Which Is Better for Explaining Machine Learning Models?*, 2021; Panati et al., 2022). It approximates an interpretable model because the original model is too complex to understand. It produces explanations in the form of feature importance scores that show how much each feature contributes to the specific prediction (Okolo et al., 2022; Panati et al., 2022).

Grad-CAM is a combination of Local Explanation and Visualization, used for CNN

models, that uses gradients of a target concept to produce local heatmaps that highlight important regions of an input image (Mohammadreza (Reza) Salehi, 2020). The idea is that by visualizing the parts of an image the CNN is focusing on, it can help gain insight into how the model makes its predictions (Mohammadreza (Reza) Salehi, 2020). For example, if the CNN is required to identify a cow in an image, the Grad-CAM heatmap would highlight the cow in that image. If the CNN makes an incorrect prediction, the Grad-CAM heatmap can help us understand why by highlighting what part of the image the model was focused on.

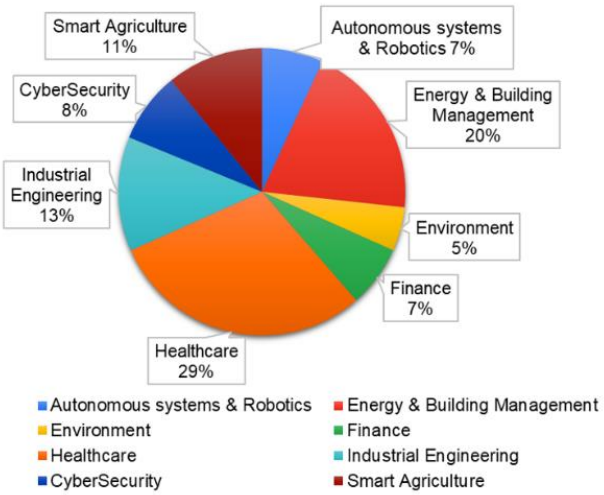

Figure 3. AI Industrial Applications (Ref: Le et al., 2023)

Le et al. (2023) conducted an extensive literature review of Local Explanation techniques and their practical applications in various industry sectors. Figure 3 depicts the AI industrial applications found in their data corpus. They found that Local Explanation techniques, with SHAP and LIME being the most dominant methods, can potentially improve the interpretability and transparency of opaque models in the industry setting. However, they identified limitations and needs for improvement. Some of these include trade-offs between model accuracy and interpretability and computational costs. They stress that more consistent evaluation measures, datasets, and benchmarks are needed to improve the interpretability of current models. They recommend that future work should focus on improving and developing more efficient and effective Local Explanation techniques that can be used on large-scale datasets and complex models.

3.1.3 Model Simplification

Model Simplification is a post-hoc XAI method that explains AI models by simplifying them through architecture or by using simpler AI models to explain the more complex ones (Barredo Arrieta et al., 2020; Owens et al., 2022). Knowledge distillation is a common Model Simplification technique which involves transferring knowledge from a large model to a smaller model (Owens et al., 2022). Transferring knowledge from a complex model to a simple one can help us understand the capabilities and predictions of the larger model.

Another commonly used Model Simplification technique is rule extraction (Owens et al., 2022). This is the process of representing knowledge learned by a model during training in the form of rules, primarily IF-THEN rules (Hailesilassie, 2016). This can help us understand complex models by using rules to represent how a model makes its decisions in a human-

readable way. For example, there is an AI model that is used to make predictions on whether or not someone will file an insurance claim. We can use rule extraction to help us understand the decisions made by the model. This process may reveal that the model uses the following two rules to make its predictions:

1. IF the customer is over 45 AND has had no traffic violations in the past five years, THEN they are unlikely to file a claim.
2. IF the customer is under 25 AND has more than one traffic violation in the past year, THEN they are likely to file a claim.

3.1.4 Feature Relevance

Feature Relevance is used to explain the behaviour of complex models by measuring how much each input feature contributes to the output of the model (Barredo Arrieta et al., 2020; Tritscher et al., 2023). Analyzing the relevance of each input feature can help us understand which ones are the most important to the model's decision-making process, in turn improving the interpretability and transparency of the model.

Tritscher et al. (2023) conducted a literature review on Feature Relevance techniques for anomaly detection. *Figure 4* represents their reviewed approaches and depicts which approaches are more relevant for data access versus model access. A key finding is that many of these XAI approaches require manual selection of a reference data point (used as a basis for comparison when explaining ML behaviour), which can be problematic as these reference data points do not transfer over. To address this, the authors suggest finding the optimal reference data through optimization. They stress that more work needs to be put in XAI to ensure the approaches are reliable in the context of anomaly detection.

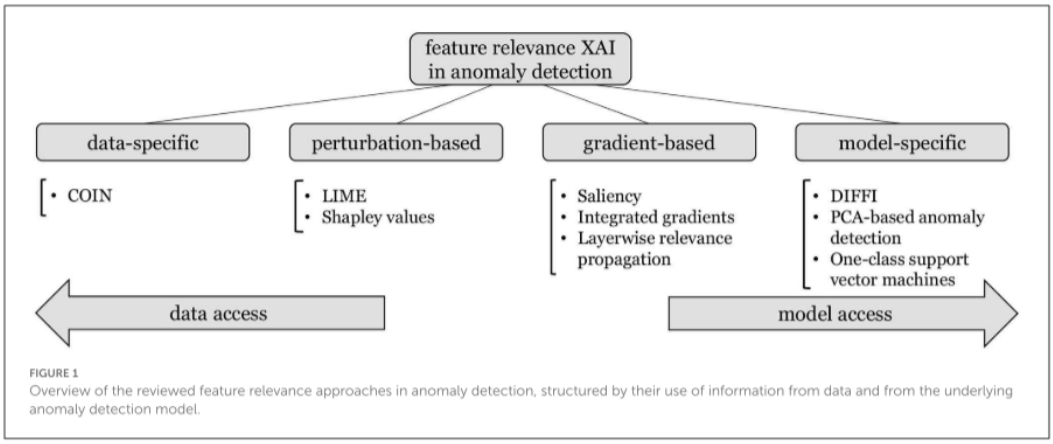

Figure 4. Feature Relevance Approaches (Ref: Tritscher et al. 2023)

3.1.5 Text Explanations

Text Explanations is an XAI approach that generates text to help explain the output of a model, this includes generating symbols, simply providing human-readable explanations (Barredo Arrieta et al., 2020). This approach can be beneficial for non-technical users as the explanations are provided in text, meaning the user does not require previous knowledge if the generated text is simplified as opposed to using complex terms and explanations (Barredo Arrieta et al., 2020; Bennetot et al., 2019).

Bennetot et al. (2019) propose an integration of connectionist and symbolic paradigms to

produce explanations of neural network models for users. They propose a reasoning model that is based off of definitions of XAI found in a paper by Doran et al. (2017) to explain the decisions of a neural network. In *Figure 5*, from their study, we see how an opaque model produces a caption output compared to their reasoning model, which provides an explanation.

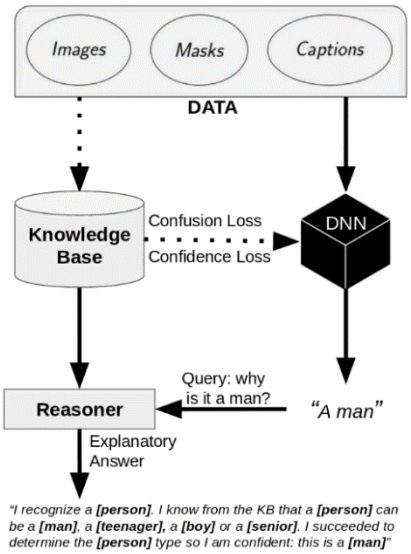

Figure 5. Reasoning Model (Ref: Bennetot et al. 2019)

3.1.6 Explanations by Example

The Explanations by Example XAI approach explains complex models by using transparent ones or representative examples that clarify the relationships of the complex model (Barredo Arrieta et al., 2020; Kenny et al., 2021). This is similar to how humans tend to explain more complex concepts, we use examples of other scenarios that has similar correlations to that of the one we are attempting to explain.

A study by Kenny et al. (2021) focused on using Explanations by Examples to improve the transparency of opaque ML models. They identify three different techniques: factual, semi-factual, and counterfactual. Figure 6 best depicts the difference between these techniques.

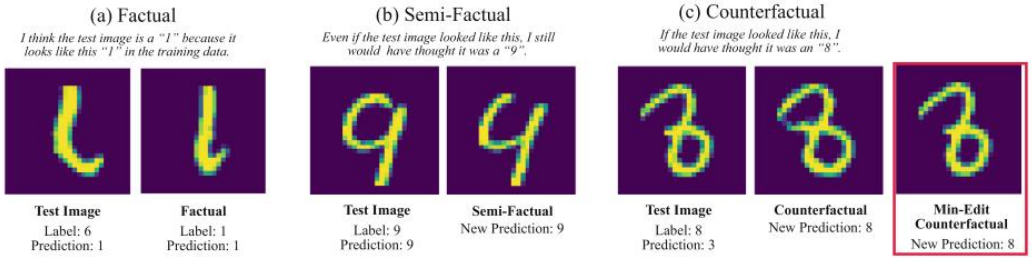

Figure 6. Post-hoc Explanation-by-Example (Ref: Kenny et al. 2021)

In this example, the factual explanation shows what the model thinks a "1" looks like an explains that is why it predicted the image to be a 1. The semi-factual explanation shows an image similar to a 9, to show how the model would still predict the image to be a 9 as it has a similar shape. The counterfactual approach shows what the model would interpret as an 8 compared to the input image and the model's prediction. By providing an example of how the

model comes to its conclusions, we are able to improve the interpretability and transparency of models.

3.2 Common & Emerging NLP Approaches

As mentioned previously, in *Section 1.1*, NLP tries to help computers understand, analyze, and generate meaningful human-readable language to bridge the gap between humans and computers (Chaman Kumar et al., 2020; Danilevsky et al., 2021; Hjorth, 2021; Luekhong et al., 2019; Maruf et al., 2022; Shivahare et al., 2022; Tsvetkov et al., 2019).

Figure 7 provides a brief overview of the applications of NLP. The integration of multiple NLP applications enables the development of robust systems that are capable of comprehending, evaluating, and generating human language (Nagarhalli et al., 2022; Shivahare et al., 2022). Machine translation systems aim to translate spoken or text data from one language to another while maintaining its original context (Chaman Kumar et al., 2020; Maruf et al., 2022). Sentiment analysis is a method that is used to determine whether textual data is positive, negative, or neutral (Raja Subramanian et al., 2021a; Thakur et al., 2022). Information retrieval systems respond to text-based input, by evaluating and retrieving textual information (Nasukawa & Yi, 2003; Raja Subramanian et al., 2021). Question-answering systems are a subset of information retrieval, that in a textual query and outputs the most appropriate response in a human-readable way (Nagarhalli et al., 2022). Conversational agents are systems that attempts to perform human-readable dialogue with a user (Nagarhalli et al., 2022). Topic modelling is used to analyze text data to determine word patterns to identify groups of similar words within a body of text (Nagarhalli et al., 2022).

One potential NLP application could be using sentiment analysis in conversational agents to analyze the sentiment of a user's message and generate an appropriate response based on this (Wahde & Virgolin, 2022). To further understand NLP approaches, a base understanding of text generation is necessary.

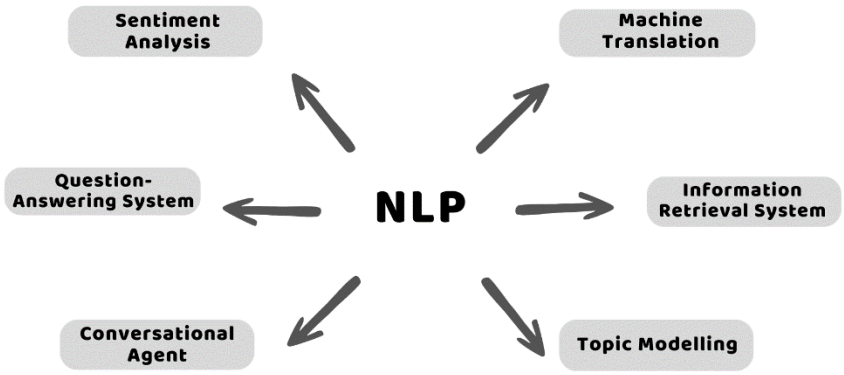

Figure 7. Applications of NLP

3.2.1 Text Generation

Text generation is one of the most challenging tasks of NLP, as it aims to produce human-

readable and plausible text (Dong et al., 2023; J. Li et al., 2021a). There are many approaches to text generation, including statistical, deep generative, and, more recently, transformer models (Abdelwahab & Elmaghraby, 2018; Dong et al., 2023).

Markov processes are a type of statistical model that generates text by predicting the next word based on the current state, such as the number of fixed previous words (Abdelwahab & Elmaghraby, 2018; Szymanski & Ciota, 2014). Deep generative models are neural networks that are trained to generate text by predicting the next word in the sequence based on the previous words (Iqbal & Qureshi, 2022). Although they seem similar, because Markov models are mathematical, the output depends on the current state of the model, meaning there is a finite number of outputs that the model is able to produce (Abdelwahab & Elmaghraby, 2018; Szymanski & Ciota, 2014b). Due to this, they have limitations in modeling the relationship between words, which can result in incoherent and non-fluent text (Szymanski & Ciota, 2014). Deep generative models, such as Long Short-Term Memory (LSTMs), are better at capturing and modeling complex relationships between words (Abdelwahab & Elmaghraby, 2018; Iqbal & Qureshi, 2022; J. Li et al., 2021).

Transformer models, introduced by Vaswani et al. (2017), are the recent game-changers in the text generation field. They are a type of neural network architecture that uses an attention mechanism to weigh the importance of different parts of input when making decisions (Amatriain et al., 2023; Vaswani et al., 2017). Vaswani et al. (2017) conducted experiments on two machine translation tasks that showed the transformer models exhibit higher quality, increased parallelizability, and reduced training times compared to other models. Transformer models have gained significant popularity and adoption in NLP (Amatriain et al., 2023; Dong et al., 2023; Vaswani et al., 2017). Some examples of state-of-the-art models include BART (Bidirectional and Auto-Regressive Transformers), trained to reconstruct texts that have been corrupted into coherent text, and GPT (Generative Pretrained Transformer), trained on large datasets to predict the next word in a sequence in a human-readable manner (Amatriain et al., 2023).

3.2.2 Machine Translation

The field of machine translation employs computational techniques to convert written or spoken content from one language to another in an efficient manner and at a lower cost (Babhulgaonkar & Sonavane, 2020; Chaman Kumar et al., 2020; Maruf et al., 2022; Phan & Jannesari, 2020; Qin, 2022). The primary methodologies employed in machine translation include Rule-Based Machine Translation (RBMT), Statistical Machine Translation (SMT), and Neural Machine Translation (NMT) (Maruf et al., 2022; Sethi et al., 2022).

3.2.2.1 Rule-Based Machine Translation

RBMT is a computational approach that examines the grammatical structure of the source text and employs linguistic rules to generate a translation in the target text (Chauhan et al., 2023; Sethi et al., 2022). According to current research findings, the utilization of RBMT as a standalone approach has diminished significantly (Maruf et al., 2022). However, it demonstrates its efficacy as a means of enhancing various systems (Aulamo et al., 2021; Hu et al., 2007; J.-X. Huang et al., 2020; Maruf et al., 2022).

In their study, Hu et al. (2007) propose a method where they utilized a pre-existing RBMT model to generate a synthetic bilingual corpus. This corpus was subsequently employed to train an SMT system. The utilization of the RBMT model for converting a monolingual

dataset into a bilingual dataset resulted in enhanced performance of the SMT system, particularly for languages with restricted data.

RBMT has been found to be a valuable technique for data augmentation due to its ability to create general-domain knowledge, which can be difficult to achieve through training alone (Aulamo et al., 2021; Hu et al., 2007; J.-X. Huang et al., 2020). In their study, Aulamo et al. (2021) employed an RBMT model to conduct back translations, generating a dataset for their NMT model. This improved the results of their NMT model.

3.2.2.2 Statistical Machine Translation

SMT employs statistical models to effectively ascertain the most probable translation for a given input (Kessikbayeva & Cicekli, 2020; Martin et al., 2011; Phan & Jannesari, 2020; Sebastian & G., 2023). These models are trained using extensive datasets to learn language patterns and identify probabilities of translation (Phan & Jannesari, 2020; Sebastian & G., 2023). Their translation process involves selecting the most probable translation based on the patterns of the input (Haffari et al., 2009; Lopez, 2008; Martin et al., 2011; Phan & Jannesari, 2020; Stasimioti et al., 2020).

A big limitation of SMT models is their difficulty acquiring knowledge of infrequent lexical items or idiomatic expressions (Maruf et al., 2022). This can be mitigated by the implementation of active learning techniques, which involve repetitive training of systems on larger datasets (Haffari et al., 2009; Phan & Jannesari, 2020).

Stasimioti et al. (2020) conducted a comparative analysis of the performance exhibited by an SMT model, a generic NMT model, and a customized NMT model. The researchers discovered that the NMT models demonstrate superior performance compared to the SMT model in terms of human evaluation criteria.

Phan & Jannesari (2020) employed and enhanced an SMT and an NMT model to address an issue known as Prefix Mapping (the challenge of learning correct sequences). The authors found that the SMT model achieved an accuracy rate of 60% to 90%, outperforming the NMT model (59% to 83%). They state that the SMT model was able to produce more accurate translations due to analyzing and identifying patterns in the data corpus quicker.

3.2.2.3 Neural Machine Translation

NMT models are trained on extensive datasets and employ deep neural networks to acquire the ability to conduct end-to-end translations, hence producing translations of superior quality when compared to other approaches (Aulamo et al., 2021; Banar et al., 2020; Belinkov et al., 2020; Maruf et al., 2022; Shi et al., 2023; Stasimioti et al., 2020). NMT differs from SMT because NMT learns implicit representations of the translation process through learning, while SMT relies on explicit rules and probabilities it has learned from the training dataset (Martin et al., 2011; Maruf et al., 2022; Phan & Jannesari, 2020).

The utilization of neural networks in these models enables the acquisition of intricate connections between source and target languages, leading to translations that are characterized by enhanced accuracy and fluency (Belinkov et al., 2020; Maruf et al., 2022). Nevertheless, NMT models have limitations. In order to attain satisfactory performance, a substantial quantity of training data is necessary, which may pose challenges for less prevalent languages (Ranathunga et al., 2021). While these models are capable of generating coherent translations, their accuracy may be compromised due to the machine's limited explicit comprehension of the underlying meaning of the translated sentence (Belinkov et al., 2020; Ranathunga et al., 2021;

Shi et al., 2023).

According to Maruf et al. (2022), both SMT and NMT models commonly engage in sentence-based translation, wherein the models translate text on a sentence-by-sentence basis. This approach may result in a limited comprehension of context and potential mistakes in the translation process. The utilization of the document-based translation strategy has the potential to alleviate these mistakes by considering the sentences preceding and following the desired one, hence enhancing the comprehension of the contextual nuances.

Table 2 summarizes the different approaches to machine translation. Before determining which approach to use, it is important to understand the context of which machine translation will be needed.

Approach	Pros	Cons
RBMT	Produces high-quality translations for languages with well-defined grammatical rules.	Difficult to maintain due to the need for extensive linguistic knowledge and manual rule creation.
	Can handle language-specific phenomena that do not follow regular grammatical rules.	Potential of increased errors when translating between languages with different/complex grammatical structures.
	It is transparent and predictable.	Unable to handle large amounts of data.
SMT	Able to handle large amounts of data and learn from the data.	Potential for increased errors when translating idiomatic expressions and language-specific phenomena that do not follow regular patterns.
	Able to produce fluent translations.	Needs large amounts of data to produce optimal results.
	Highly flexible, as the models can be updated and improved with data.	Potential for increased errors when translating languages that have very different grammatical structures.
NMT	Able to produce fluent and human-readable translations. Specifically, for longer sentences.	Needs large amounts of high-quality data.
	Able to learn implicit representations of translations through datasets, allowing it to handle idiomatic expressions and language-specific phenomena.	Need for significant computational resources to train the neural network and generate translations.
	Able to handle large amounts of data.	Potential for errors when translating low-frequency words or phrases.

Table 2. Machine Translation Approaches

3.2.3 Sentiment Analysis

Sentiment analysis is a classification NLP task that is specifically performed on text data to determine if the sentiment is positive, negative, or neutral (Raja Subramanian et al., 2021a). A review by Subramanian and et. (2021) provide a general workflow of sentiment analysis, depicted in *Figure 8*, that involves the following five stages: data collection, pre-processing, feature extraction/ selection, sentiment classification, and error analysis.

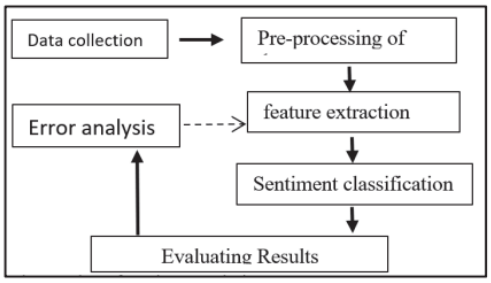

Figure 8. Sentiment Analysis Workflow (Ref: Subramanian et al. (2021))

Lighart et al. (2021) conducted an extensive review that offers a comprehensive analysis of the tasks, approaches, and challenges involved in sentiment analysis. In addition, they comprised a list of 112 academic articles on sentiment analysis that utilize state-of-the-art techniques. I will provide a brief overview of current tasks, and approaches for sentiment analysis, refer to Lighart et al. (2021) for more extensive results. *Figure 9* depicts an extensive overview of sentiment analysis, created by Lighart et al. (2021).

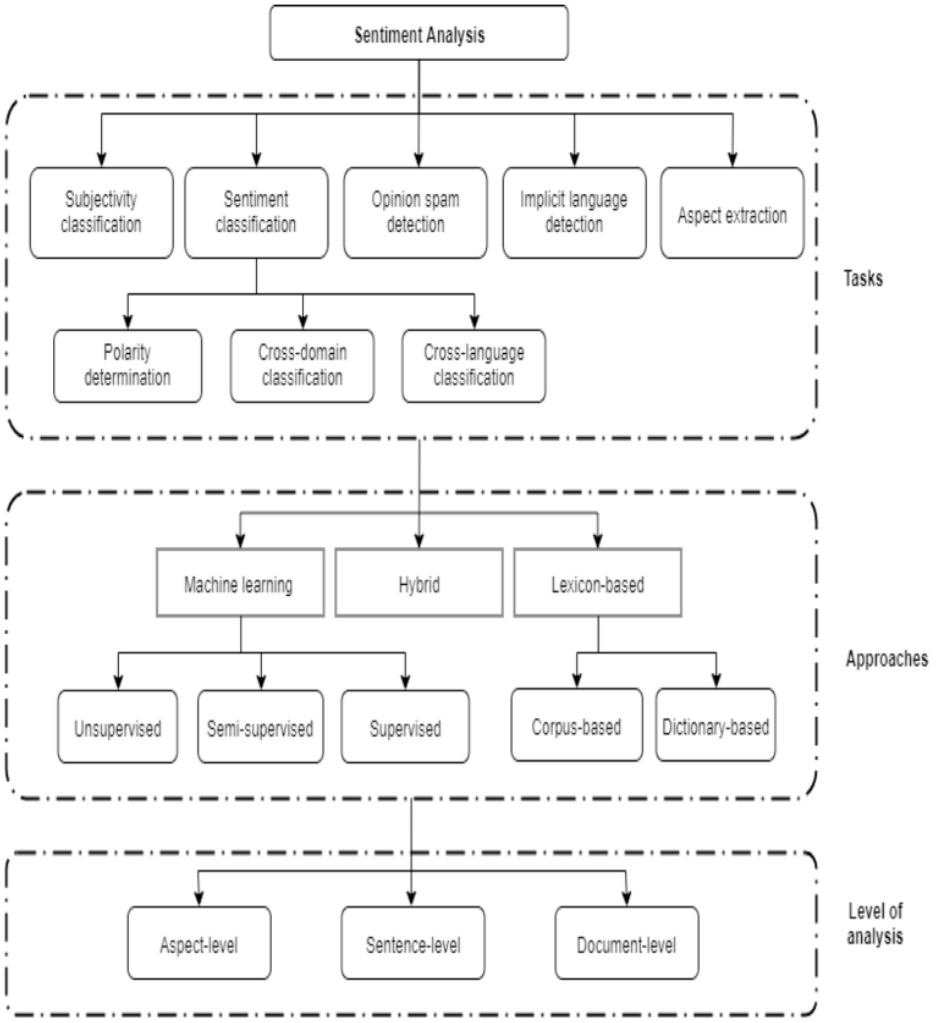

Figure 9. Sentiment Analysis Overview (Ref: Lighart et al. (2021))

3.2.3.1 Sentiment Analysis Tasks

Two of the most widely recognized sentiment analysis tasks include sentiment classification and subject classification (Ligthart et al., 2021; Thakur et al., 2022). While these are the two main tasks of sentiment analysis, others include opinion detection, implicit language detection, and aspect extraction (Ligthart et al., 2021).

Sentiment classification is the fundamental component of sentiment analysis, identifying the overall sentiment of a text (Ligthart et al., 2021). This can be done at three levels of granularity; document-level, sentence-level, or aspect-based (Behdenna et al., 2016; Chiha et al., 2022; Ligthart et al., 2021; Zhang et al., 2022).

Document-level sentiment analysis is the process of determining the overall sentiment of the entire document (Behdenna et al., 2016; Ligthart et al., 2021; Raja Subramanian et al., 2021a). This is typically done by analyzing the words and phrases within the document and scoring them based on how positive, negative, or neutral they are (Behdenna et al., 2016; Choi et al., 2021).

Alternatively, sentence-level sentiment analysis looks at each sentence and determines the sentiment of that specific sentence (Behdenna et al., 2016; Chiha et al., 2022). This can be useful for identifying specific parts of a text. Choi et al. (2021) proposed a deep-learning model based on the combination between document-level and sentence-level sentiment analysis. While it is primarily document-level based, it automatically considers the importance degrees of sentences within the document. They conducted experiments using sentiment datasets from movie, hotel, restaurant, and music reviews. They found that their model outperformed state-of-the-art document-level models and expressed the importance of sentences in document-level sentiment classification tasks.

Lastly, aspect-based sentiment analysis takes it a step further by identifying the sentiment of specific features or components of a product or service (Zhang et al., 2022). For example, if we look at a review for a movie, the aspect-based analysis could tell us that the atmosphere, cinematography, and actors are being talked about and can determine the sentiment of the reviews.

Subjective classification involves determining whether a given piece of text is subjective or objective in nature (Bing Liu, 2020; Ligthart et al., 2021). The subjective text expresses the author's opinions, feelings, and emotions (Ligthart et al., 2021). The objective text expresses the facts without any bias (Ligthart et al., 2021). Understanding the difference between the two can help to filter out objective text when conducting sentiment classification (Bing Liu, 2020; S. Li, 2013).

Li (2013) proposed a semi-supervised method for classifying sentiment behind web consumer reviews by using a large number of unlabeled examples. Each review had a subjective and objective view. The method is based on the co-training framework, which needs three basic sentiment classifiers to get the end sentiment classifier. In the suggested method, the common unigram features from customer reviews are used to build the first sentiment classifier. The second sentiment classifier is trained on the subjective views made up of opinion words taken from user reviews. The rest of the reviews' text features are used to get independent views that can be used to train the third classifier. The author state that the results from this method outperformed self-learning support vector models.

3.2.3.2 Sentiment Analysis Approaches

The three main sentiment analysis approaches include lexicon-based, machine learning, and a hybrid approach (Behdenna et al., 2016; Ligthart et al., 2021; Raja Subramanian et al., 2021). The lexicon-based approach uses a pre-defined dictionary of words and classifies them with a score or into a class (positive, negative, neutral), to determine the sentiment of a given text (Raja Subramanian et al., 2021). These scores are then added together mathematically to get an overall sentiment score (Behdenna et al., 2016; Raja Subramanian et al., 2021). For example, a common way to figure out the sentiment score is:

Sentiment Score = (number of positive words - number of negative words) / total number of words

If the text's sentiment number is negative, it is called negative. If the score is positive, the text is considered positive. If the number is zero, the text is considered neutral.

ML is often used to automatically classify the sentiment of text. Some common ML sentiment analysis approaches used are Naïve Bayes, linear regression, support vector machines (SVM), and CNNs (Behdenna et al., 2016). These methods use different ways to examine text and pull-out features that can be used to figure out what the text is trying to say. For example, Naive Bayes gives a probability to whether a given word or phrase should be seen as positive or negative, while SVM can be taught to classify text based on the features extracted from the text (Bin Sulaiman et al., 2016; Nandwani & Verma, 2021). Each method has its own pros and cons, and the best method to use relies on the problem at hand.

Hybrid approaches in sentiment analysis combine the best of lexicon-based and ML methods. For example, a hybrid method might use a ML algorithm to classify text based on features extracted from the text and a lexicon-based approach to find and label positive and negative words (Behdenna et al., 2016; Ligthart et al., 2021). It has been shown that hybrid methods reduce sentiment errors on training data that are getting more complicated (Behdenna et al., 2016; Ligthart et al., 2021; Nandwani & Verma, 2021; Raja Subramanian et al., 2021).

3.3 Challenges & Biases in NLP

NLP is a rapidly growing field; however, there are still many challenges and biases that exist. Datasets have biases and limitations, and models and approaches face challenges regarding accuracy, scalability, and interpretability. The richness, fluidity, and irregularity of human language is a significant barrier for NLP. The meaning of a word or phrase, for instance, can shift based on the surrounding material (Bansal, 2022). NLP also faces significant difficulties with bias. It is possible for unsupervised AI algorithms to automatically identify linguistic regularities that represent human prejudices like racism, sexism, and ableism by mining natural language datasets for hidden patterns (Bansal, 2022).

3.3.1 Machine Translation Challenges

The field of machine translation has made great strides in recent years, but it still faces a number of obstacles. Human language presents a significant obstacle due to its inherent complexity and irregularity. When employed in different contexts, the same words and phrases can have multiple meanings (Du, 2019; Robertson & Díaz, 2022; Sethi et al., 2022). Another difficulty is that machine translation tools are not yet sophisticated enough to understand

cultural nuance, context, and local slang, which can lead to translations that sound stilted, choppy, and misaligned with the target culture (Chaman Kumar et al., 2020; Gong, 2022; Maruf et al., 2022). Specifically, neural machine translation faces many obstacles, including domain mismatch, a lack of training data, uncommon words, lengthy sentences, improper word alignment, and beam search . These problems emphasize the importance of maintaining and expanding machine translation research.

The lack of datasets for low-resource languages (not commonly spoken languages) can lead to inaccuracies in translation models, resulting in limited training and limitations in tasks the models can accomplish (Bansal, 2022; Omar et al., 2022). Meta AI's No Language Left Behind (NLLB) underwent intensive research to ensure they had an adequate language dataset. This process included interviewing multiple native speakers and bringing in professionals to assess the translations their model produced. It is important to note that NLLB is a large project that has the funding and resources to generate a dataset for low-resource languages adequately. However, this is not the reality for many project groups.

3.3.2 Sentiment Analysis Challenges

Challenges in sentiment analysis stem from the intricacies and irregularities of human language. For instance, the meaning of a word or phrase can change based on the surrounding text. People also frequently employ upbeat language to convey unpleasant emotions, making sarcasm detection a difficult task. Furthermore, certain statements cannot be characterized as positive, negative, or neutral, making it difficult for algorithms to infer the polarity of a statement (Ligthart et al., 2021). Additionally, it becomes more difficult for the algorithm to distinguish the intended meaning when words have multiple meanings.

Sentiment analysis provides an overview of the intent behind the text and cannot detect specific emotions. Emotion detection was derived from sentiment analysis to address this issue (Nandwani & Verma, 2021; Seyeditabari et al., 2018). This has the capability to recognize emotions such as happiness, sadness, love, or even frustration (Nandwani & Verma, 2021; Seyeditabari et al., 2018). Detecting emotion from text is extremely challenging as most research uses physical attributes (heart rate, shivering, sweating, and voice pitch) to detect emotions (Nandwani & Verma, 2021). In order to convey emotion detection to text, researchers introduced emotion models that make up a set of labels used to annotate sentences or documents (Nandwani & Verma, 2021; Seyeditabari et al., 2018).

3.3.3 Bias in NLP

Language patterns that represent human prejudices like racism, sexism, and ableism can be captured by unsupervised AI algorithms that automatically find hidden patterns in natural language datasets, leading to harmful stereotypes and biases in NLP (Aylin Caliskan, 2021; Omar et al., 2022). Additionally, the lack of diverse and representative datasets limits the ability of NLP models to understand and perform tasks accurately. This can result in poor performance for marginalized groups, making AI harmful to those societies if these issues are not addressed appropriately. For example, Caliskan (2021) discusses the Amazon Resume NLP system, which resulted in the model favouring male candidates over females due to the underrepresentation of women in the training dataset.

A review by Bansal (2022) examines the history of biases, definitions of fairness, and approaches used by NLP to reduce bias. They explain why it is essential to investigate the societal impact of NLP models and how integrated they have become in people's lives to

understand the effects of bias within them. Despite these models' linguistic sophistication and improved performance on challenging downstream tasks, research has shown that they reinforce harmful gender, racial, and cultural prejudices and perpetuate the problem in many contexts. While researchers are making an effort to identify and mitigate bias in NLP, there is still more work to be done to ensure these biases are resolved.

3.4 Intersecting XAI & NLP

The necessity for openness and trust in the decision-making process of AI systems that employ NLP integrates XAI. There is a growing need to put faith in the decision-making process as AI gains traction in high-stakes industries. Especially in contexts where a human and a computer interact, like with chatbots, a lack of interpretability can lead to a loss of confidence.

Yu et al. (2022) argue that current XAI approaches in NLP only focus on delivering one explanation, failing to account for the diversity of human thoughts and language expression. To address this, they propose a generative XAI framework called INTERACTION for NLP interface explanations in two steps: “Explanation and Label Prediction” and “Diverse Evidence Generation.” Step one involves the framework explaining the model’s prediction and predicts the most appropriate label. The explanation is generated based on the model’s understanding of the input text and reasoning for making the predictions. The label predictions are made by the model based on its analysis of the input text and its understanding of the task at hand. Step two uses a generative model to generate multiple diverse explanations that are semantically similar to the explanation generated in step one but are expressed differently. This helps the user see multiple perspectives on the model’s reasoning and can help build trust in the predictions. To evaluate their model, the authors conduct intensive experiments with the Transformer architecture on a benchmark dataset. They show that their method achieves competitive or better performance against state-of-the-art baseline models on explanation generation and prediction.

A survey paper by Danilevsky et al. (2020) discusses the main ways to explain the decisions made by NLP models, as well as techniques currently available for generating explanations for these models. Explanations can be categorized into local or global predictions. Local explanations pertain to a particular prediction, whereas global explanations consider the entire model as a whole. For example, there is a model that predicts that a patient has a high risk of developing diabetes. A local explanation would provide information about the specific features and their values, such as the patient’s age, weight, or family history, that contributed to the prediction (David Lazaridis, 2021). Where a global explanation would predict the risk of diabetes and could show which features add to that risk across all patients (David Lazaridis, 2021).

3.4.1 Explainability Techniques

Within the data corpus, some XAI methods that can be used to explain NLP models include feature importance, surrogate models, and example-driven techniques (David Lazaridis, 2021; Yu et al., 2022).

3.4.1.1 Feature Importance

Feature importance derives explanations by investigating the importance of specific features and their given scores regarding the models output prediction (David Lazaridis, 2021; Husna

Sayedi, 2021; Yu et al., 2022). These methods might be based on lexical characteristics such as words or tokens, or on manual features created through feature engineering (David Lazaridis, 2021; J. Huang et al., 2022). In NLP, this is used to rank the importance of individual words and phrases in the model's prediction (Husna Sayedi, 2021).

A study by Huang et al. (2022) explores XAI and the development of reliable and fair AI models through the application of effective explanation approaches. The authors offer measures to evaluate how well feature importance order and saliency maps capture feature contribution explanation summaries. Motivated by these metrics, they create a XAI procedure based on the XAI criterion of feature importance, which systematically selects XAI methods and evaluates the coherence of explanations. The authors showcase the XAI techniques' contributions to the process development of three distinct areas: a search ranking system, code vulnerability detection, and picture categorization. Their contribution is a practical and systematic process with defined consistency metrics to produce rigorous feature contribution explanations.

3.4.1.2 Surrogate Model

The predictions of an ML model can be explained by a second model, known as a surrogate model, which can provide both local and global explanations (David Lazaridis, 2021; Husna Sayedi, 2021). One limitation of this approach is the potential disparity in methodology employed by the learnt surrogate models and the original model during the prediction process (David Lazaridis, 2021; Sun et al., 2020). This implies that the explanations offered by the surrogate model may not consistently and precisely depict the decision-making mechanism of the original model.

Sun et al. (2020) explored the use of a surrogate model to improve explanations in NLP by proposing a self-explanatory framework for DL models. Instead of developing a complete separate model, they wanted to build onto the one they already had. The suggested framework's central idea is to build an extra interpretation layer, on top of any preexisting NLP model. Each text span is given a weight in this layer, and the softmax function receives the combined weights for the best possible prediction. The authors state that the model's self-explanatory nature eliminates the need for a separate model to interpret results. Other contributions include the model's generalizability, and the weight associated with each text span provides direct importance scores for phrases and sentences.

3.4.1.3 Example Driven

As introduced in *Section 3.1.6*, this method employs semantically related instances, often drawn from labelled data to explain the model's prediction (Barredo Arrieta et al., 2020; Kenny et al., 2021). For example, if an NLP model is taught to assign texts to various buckets, an explanation based on examples could provide texts that are comparable to the input text and have been assigned to the same bucket. This has the potential to improve faith and understanding in NLP-based AI decision-making.

4. Discussion

The field of NLP is growing immensely with the development and improvement of state-of-the-art models and the integration of chat bots, such as ChatGPT or Bing Chat. With the expansion of NLP in day-to-day everyday lives, the need for interpretability increases along with it. As seen throughout the data corpus, the majority of research in NLP revolves around the models

and their algorithms rather than their usability. A big take away is that NLP models require evaluations that go beyond standard metrics and measures in order to truly understand their potential and challenges in day-to-day settings. There has been a development in Human-in-the-loop (HITL) NLP frameworks that integrate human feedback throughout the model design and evaluation process to improve the model further . Collecting diverse feedback from various people and applying different methods to learn from human feedback is a step towards interpretability and transparency in NLP models.

As mentioned in *Section 3.1*, the goal of XAI is to make opaque models more interpretable and transparent. This is done by providing explanations for the decisions made by the AI, thus increasing trust in these systems. However, there is still a long way to go before we can truly achieve this goal. While there are many approaches for XAI, not many have been evaluated in real world scenarios, therefore, begging the question "Are they truly explainable?". As seen in our data corpus, there is a lack of unity when it comes to evaluation metrics for XAI approaches. This makes it challenging to understand which approach to use for which situation and how the approaches compare against one another. Another interesting observation is that some XAI approaches lack interpretability themselves, especially for those who lack prior knowledge. For example, local explanation approaches such as SHAP or LIME, are hard to comprehend without prior knowledge or explanations. This makes these approaches less applicable to real world scenarios. Additionally, there is much work to be done to realize the potential of XAI and NLP and how to integrate them to produce trustworthy and understandable AI.

Lastly, as noticed in the data corpus, there is a lack of consistency in defining sentiment analysis terms. This makes it challenging to understand current research in the field and what is needed for future work. Some researchers merge the meaning of sentiment analysis with emotion detection and/or topic modelling. It is important to be able to differentiate between terms in order to ensure transparency and interpretability throughout the field.

5. Future Work

As XAI and NLP are a vastly growing field, there are many future directions to delve into. The goals for future work differ between the field of AI and Human-Computer Interaction (HCI).

Future works for AI include the following:

1. Developing large scale reliable and efficient models that have the capability of performing multiple NLP tasks.
2. Creating publicly available datasets and open-access NLP models.

Future works for HCI include the following:

1. Using XAI to conduct studies to make more interpretable, transparent, and understandable models. As well as understanding the limitations and benefits of using these XAI approaches with these NLP models.
2. Conducting more human-centric studies in regard to XAI and NLP.
3. Creating research protocol that has unified definitions for terms used in NLP, specifically sentiment analysis.

6. Conclusion

NLP and XAI are two fields that have the potential to change how we interact with intelligent

systems. NLP helps robots perceive, interpret, and output natural language. This enables language translation, sentiment analysis, question-answering systems, and information retrieval. XAI provides understandable and justified explanations for AI model decisions. This literature review demonstrates the potential for enhancing the interpretability of language-based systems through the integration of NLP and XAI. By emphasizing optimal methodologies, addressing ethical concerns, and facilitating collaboration between these two fields, I shed light on the need for explainability in AI. By combining these two disciplines, scholars can cultivate AI systems that prioritize human needs and exhibit enhanced efficacy, as well as increased transparency and reliability. This review has offered significant insights on the potential for future research in this particular domain and has established the necessary foundation for further investigation into the intersection between NLP and XAI.

Bibliography:

Abdelwahab, O., & Elmaghraby, A. S. (2018). Deep learning based vs. Markov chain based text generation for cross domain adaptation for sentiment classification. *Proceedings - 2018 IEEE 19th International Conference on Information Reuse and Integration for Data Science, IRI 2018*, 252–255. https://doi.org/10.1109/IRI.2018.00046

Amatriain, X., Sankar, A., Bing, J., Bodigutla, P. K., Hazen, T. J., & Kazi, M. (2023). *Transformer models: an introduction and catalog*. http://arxiv.org/abs/2302.07730

Aulamo, M., Virpioja, S., Scherrer, Y., & Tiedemann, J. (2021). *Boosting Neural Machine Translation from Finnish to Northern Sámi with Rule-Based Backtranslation*. https://giellatekno.uit.no/

Aylin Caliskan. (2021). *Detecting and mitigating bias in natural language processing*.

Babhulgaonkar, A., & Sonavane, S. (2020). Language Identification for Multilingual Machine Translation. *2020 International Conference on Communication and Signal Processing (ICCSP)*, 401–405. https://doi.org/10.1109/ICCSP48568.2020.9182184

Banar, N., Daelemans, W., & Kestemont, M. (2020). Character-Level Transformer-Based Neural Machine Translation. *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, 149–156. https://doi.org/10.1145/3443279.3443310

Bansal, R. (2022). *A Survey on Bias and Fairness in Natural Language Processing*. http://arxiv.org/abs/2204.09591

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Behdenna, S., Barigou, F., & Belalem, G. (2016). Sentiment analysis at document level. *Communications in Computer and Information Science*, *628 CCIS*, 159–168. https://doi.org/10.1007/978-981-10-3433-6_20

Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., & Glass, J. (2020). On the Linguistic Representational Power of Neural Machine Translation Models. *Computational Linguistics*, *46*(1), 1–52. https://doi.org/10.1162/coli_a_00367

Bennetot, A., Laurent, J.-L., Chatila, R., & Díaz-Rodríguez, N. (2019). *Towards Explainable Neural-Symbolic Visual Reasoning*. http://arxiv.org/abs/1909.09065

Bin Sulaiman, H. Asyrani., IEEE Computer Society. Malaysia Chapter, IEEE Malaysia Section, & Institute of Electrical and Electronics Engineers. (n.d.). *2014 I4CT : 1st International Conference on Computer, Communications, and Control Technology : proceedings : Fave Hotel, Langkawi, Kedah, Malaysia, 2-4 Sept 2014*.

Bing Liu. (2020). Sentence Subjectivity and Sentiment Classification. In *Sentiment Analysis* (pp. 89–114). Cambridge University Press. https://doi.org/10.1017/9781108639286.005

Chaman Kumar, K. M., Aswale, S., Shetgaonkar, P., Pawar, V., Kale, D., & Kamat, S. (2020). A Survey of Machine Translation Approaches for Konkani to English. *2020 International Conference on Emerging Trends in Information Technology and Engineering (Ic-ETITE)*, 1–6. https://doi.org/10.1109/ic-ETITE47903.2020.110

Chauhan, S., Shet, J. P., Beram, S. M., Jagota, V., Dighriri, M., Ahmad, M. W., Hossain, M. S., & Rizwan, A. (2023). Rule Based Fuzzy Computing Approach on Self-Supervised Sentiment Polarity Classification with Word Sense Disambiguation in Machine Translation for Hindi Language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *22*(5), 1–21. https://doi.org/10.1145/3574130

Chiha, R., Ayed, M. Ben, & Pereira, C. da C. (2022). A complete framework for aspect-level and sentence-level sentiment analysis. *Applied Intelligence*, *52*(15), 17845–17863. https://doi.org/10.1007/s10489-022-03279-9

Choi, G., Oh, S., & Kim, H. (2021). *Improving Document-Level Sentiment Classification Using Importance of Sentences*. www.mdpi.com/journal/entropy

Danilevsky, M., Dhanorkar, S., Li, Y., Popa, L., Qian, K., & Xu, A. (2021). Explainability for Natural Language Processing. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 4033–4034. https://doi.org/10.1145/3447548.3470808

David Lazaridis. (2021, April 12). *Explainable AI (XAI) and Interpretable Machine Learning (IML) models*.

Dong, C., Li, Y., Gong, H., Chen, M., Li, J., Shen, Y., & Yang, M. (2023). A Survey of Natural Language Generation. *ACM Computing Surveys*, *55*(8), 1–38. https://doi.org/10.1145/3554727

Du, P. (2019). Elimination of Machine Translation Errors in English Language Transformation. *2019 11th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, 642–647. https://doi.org/10.1109/ICMTMA.2019.00147

Feng, S., & Boyd-Graber, J. (2019). What can AI do for me? *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 229–239. https://doi.org/10.1145/3301275.3302265

Gohel, P., Singh, P., & Mohanty, M. (2021). *Explainable AI: current status and future directions*. http://arxiv.org/abs/2107.07045

Gong, Y. (2022). Study on Machine Translation Teaching Model Based on Translation Parallel Corpus and Exploitation for Multimedia Asian Information Processing. *ACM Transactions on Asian and Low-Resource Language Information Processing*. https://doi.org/10.1145/3523282

Haffari, G., Roy, M., & Sarkar, A. (2009). *Active Learning for Statistical Phrase-based Machine Translation **.

Hailesilassie, T. (2016). Rule Extraction Algorithm for Deep Neural Networks: A Review. In *IJCSIS) International Journal of Computer Science and Information Security* (Vol. 14, Issue 7). https://sites.google.com/site/ijcsis/

Hartung, T., Kloft, M., Golden, E., Tritscher, J., Krause, A., & Hotho, A. (n.d.). *Feature relevance XAI in anomaly detection: Reviewing approaches and challenges*. https://professor-x.de/feature-relevance-AD.

Hjorth, A. (2021). NaturalLanguageProcesing4All. *Proceedings of the 17th ACM Conference on International Computing Education Research*, 347–354. https://doi.org/10.1145/3446871.3469749

Hu, X., Wang, H., & Wu, H. (2007). *Using RBMT Systems to Produce Bilingual Corpus for SMT*. http://www.statmt.org/wmt06/shared-

Huang, J., Wang, Z., Li, D., & Liu, Y. (2022). *The Analysis and Development of an XAI Process on Feature Contribution Explanation*.

Huang, J.-X., Lee, K.-S., & Kim, Y.-K. (2020). Hybrid Translation with Classification: Revisiting Rule-Based and Neural Machine Translation. *Electronics*, *9*(2), 201. https://doi.org/10.3390/electronics9020201

Hudon, A., Demazure, T., Karran, A., Léger, P. M., & Sénécal, S. (2021). Explainable Artificial Intelligence (XAI): How the Visualization of AI Predictions Affects User Cognitive Load and Confidence. *Lecture Notes in Information Systems and Organisation*, *52 LNISO*, 237–246. https://doi.org/10.1007/978-3-030-88900-5_27

Hughes, R., Edmond, C., Wells, L., Glencross, M., Zhu, L., & Bednarz, T. (2020). eXplainable AI (XAI). *SIGGRAPH Asia 2020 Courses*, 1–62. https://doi.org/10.1145/3415263.3419166

Husna Sayedi. (2021, November 4). *Explainable AI (XAI): NLP Edition*.

Iqbal, T., & Qureshi, S. (2022). The survey: Text generation models in deep learning. *Journal of King Saud University - Computer and Information Sciences*, *34*(6), 2515–2528. https://doi.org/10.1016/J.JKSUCI.2020.04.001

Jacovi, A. (2023). *Trends in Explainable AI (XAI) Literature*. http://arxiv.org/abs/2301.05433

Kenny, E. M., Delaney, E. D., Greene, D., & Keane, M. T. (2021). Post-hoc Explanation Options for XAI in Deep Learning: The Insight Centre for Data Analytics Perspective. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12663 LNCS*, 20–34. https://doi.org/10.1007/978-3-030-68796-0_2

Kessikbayeva, G., & Cicekli, I. (2020). Impact of Statistical Language Model on Example Based Machine Translation System between Kazakh and Turkish Languages. *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, 112–118. https://doi.org/10.1145/3443279.3443286

Le, T.-T.-H., Prihatno, A. T., Oktian, Y. E., Kang, H., & Kim, H. (2023). Exploring Local Explanation of Practical Industrial AI Applications: A Systematic Literature Review. *Applied Sciences*, *13*(9), 5809. https://doi.org/10.3390/app13095809

Li, J., Tang, T., Zhao, W. X., & Wen, J.-R. (2021). *Pretrained Language Models for Text Generation: A Survey*. http://arxiv.org/abs/2105.10311

Li, S. (2013). Sentiment Classification using Subjective and Objective Views. In *International Journal of Computer Applications* (Vol. 80). http://opennlp.apache.org/

Ligthart, A., Catal, C., & Tekinerdogan, B. (2021). Systematic reviews in sentiment analysis: a tertiary study. *Artificial Intelligence Review*, *54*(7), 4997–5053. https://doi.org/10.1007/s10462-021-09973-3

LIME vs. SHAP: Which is Better for Explaining Machine Learning Models? (2021, August 7). https://ernesto.net/lime-vs-shap-which-is-better-for-explaining-machine-learning-models/

Lopez, A. (2008). Statistical machine translation. *ACM Computing Surveys*, *40*(3), 1–49. https://doi.org/10.1145/1380584.1380586

Luekhong, P., Limkonchotiwat, P., & Ruangrajitpakorn, T. (2019). A Study on an Effect of Using Deep Learning in Thai-English Machine Translation Processes. *2019 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing (ISAI-NLP)*, 1–6. https://doi.org/10.1109/iSAI-NLP48611.2019.9045115

Martin, E., Kaski, S., Zheng, F., Webb, G. I., Zhu, X., Muslea, I., Ting, K. M., Vlachos, M., Miikkulainen, R., Fern, A., Osborne, M., Raedt, L. De, Kersting, K., Zeugmann, T., Zhang, X., Bain, M., Czumaj, A., Sohler, C., Sammut, C., … Kersting, K. (2011). Statistical Machine Translation. In *Encyclopedia of Machine Learning* (pp. 912–915). Springer US. https://doi.org/10.1007/978-0-387-30164-8_783

Maruf, S., Saleh, F., & Haffari, G. (2022). A Survey on Document-level Neural Machine Translation. *ACM Computing Surveys*, *54*(2), 1–36. https://doi.org/10.1145/3441691

Mohammadreza (Reza) Salehi. (2020, January 18). *A Review of Different Interpretation Methods (Part 1: Saliency Map, CAM, Grad-CAM)*. https://mrsalehi.medium.com/a-review-of-different-interpretation-methods-in-deep-learning-part-1-saliency-map-cam-grad-cam-3a34476bc24d

Nagarhalli, T. P., Mhatre, S., Patil, S., & Patil, P. (2022). The Review of Natural Language Processing Applications with Emphasis on Machine Learning Implementations. *2022 International Conference on Electronics and Renewable Systems (ICEARS)*, 1353–1358. https://doi.org/10.1109/ICEARS53579.2022.9752326

Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. In *Social Network Analysis and Mining* (Vol. 11, Issue 1). Springer. https://doi.org/10.1007/s13278-021-00776-6

Nasukawa, T., & Yi, J. (2003). Sentiment analysis. *Proceedings of the 2nd International Conference on Knowledge Capture*, 70–77. https://doi.org/10.1145/945645.945658

Okolo, C. T., Dell, N., & Vashistha, A. (2022). Making AI Explainable in the Global South: A Systematic Review. *ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS)*, 439–452. https://doi.org/10.1145/3530190.3534802

Omar, M., Choi, S., Nyang, D., & Mohaisen, D. (2022). *Robust Natural Language Processing: Recent Advances, Challenges, and Future Directions*. http://arxiv.org/abs/2201.00768

Owens, E., Sheehan, B., Mullins, M., Cunneen, M., Ressel, J., & Castignani, G. (2022). Explainable Artificial Intelligence (XAI) in Insurance. In *Risks* (Vol. 10, Issue 12). MDPI. https://doi.org/10.3390/risks10120230

Pafla, M. (2020). *Researching Human-AI Collaboration through the Design of Language-Based Query Assistance*.

Panati, C., Wagner, S., & Bruggenwirth, S. (2022). Feature Relevance Evaluation using Grad-CAM, LIME and SHAP for Deep Learning SAR Data Classification. *2022 23rd International Radar Symposium (IRS)*, 457–462. https://doi.org/10.23919/IRS54158.2022.9904989

Phan, H., & Jannesari, A. (2020). Statistical machine translation outperforms neural machine translation in software engineering: why and how. *Proceedings of the 1st ACM SIGSOFT International Workshop on Representation Learning for Software Engineering and Program Languages*, 3–12. https://doi.org/10.1145/3416506.3423576

Qin, M. (2022). Machine Translation Technology Based on Natural Language Processing. *2022 European Conference on Natural Language Processing and Information Retrieval (ECNLPIR)*, 10–13. https://doi.org/10.1109/ECNLPIR57021.2022.00014

Raja Subramanian, R., Akshith, N., Murthy, G. N., Vikas, M., Amara, S., & Balaji, K. (2021). A survey on sentiment analysis. *Proceedings of the Confluence 2021: 11th International Conference on Cloud Computing, Data Science and Engineering*, 70–75. https://doi.org/10.1109/Confluence51648.2021.9377136

Ranathunga, S., Lee, E.-S. A., Skenduli, M. P., Shekhar, R., Alam, M., & Kaur, R. (2021). *Neural Machine Translation for Low-Resource Languages: A Survey*. http://arxiv.org/abs/2106.15115

Robertson, S., & Díaz, M. (2022). Understanding and Being Understood: User Strategies for Identifying and Recovering From Mistranslations in Machine Translation-Mediated Chat. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2223–2238. https://doi.org/10.1145/3531146.3534638

Sebastian, M. P., & G., S. K. (2023). Malayalam Natural Language Processing: Challenges in Building a Phrase-Based Statistical Machine Translation System. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *22*(4), 1–51. https://doi.org/10.1145/3579163

Sethi, N., Dev, A., Bansal, P., Sharma, D. K., & Gupta, D. (2022). Hybridization Based Machine Translations for Low-Resource Language with Language Divergence. *ACM Transactions on Asian and Low-Resource Language Information Processing*. https://doi.org/10.1145/3571742

Seyeditabari, A., Tabari, N., & Zadrozny, W. (2018). *Emotion Detection in Text: a Review*. http://arxiv.org/abs/1806.00674

Shi, X., Huang, H., Jian, P., & Tang, Y.-K. (2023). Approximating to the Real Translation Quality for Neural Machine Translation via Causal Motivated Methods. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *22*(5), 1–26. https://doi.org/10.1145/3583684

Shivahare, B. D., Ranjan, S., Rao, A. M., Balaji, J., Dattattrey, D., & Arham, M. (2022). Survey Paper: Study of Sentiment Analysis and Machine Translation using Natural Language Processing and its Applications. *Proceedings of 3rd International Conference on Intelligent Engineering and Management, ICIEM 2022*, 652–656. https://doi.org/10.1109/ICIEM54221.2022.9853044

Stasimioti, M., Sosoni, V., Mouratidis, D., & Kermanidis, K. (2020). *Machine Translation Quality: A comparative evaluation of SMT, NMT and tailored-NMT outputs*. https://la-tools.lexile.com/free-analyze/

Sun, Z., Fan, C., Han, Q., Sun, X., Meng, Y., Wu, F., & Li, J. (2020). *Self-Explaining Structures Improve NLP Models*. http://arxiv.org/abs/2012.01786

Szymanski, G., & Ciota, Z. (2014). *Hidden Markov Models Suitable for Text Generation*. https://www.researchgate.net/publication/255626759

Thakur, O., Saritha, S. K., & Jain, S. (2022). Topic Modeling, Sentiment Analysis and Text Summarization for Analyzing News Headlines and Articles. *Communications in Computer and Information Science*, *1762 CCIS*, 220–239. https://doi.org/10.1007/978-3-031-24352-3_18

Tsvetkov, Y., Prabhakaran, V., & Voigt, R. (2019). Socially Responsible Natural Language Processing. *Companion Proceedings of The 2019 World Wide Web Conference*, 1326–1326. https://doi.org/10.1145/3308558.3320097

van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, *291*, 103404. https://doi.org/10.1016/j.artint.2020.103404

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*. http://arxiv.org/abs/1706.03762

Wahde, M., & Virgolin, M. (2022). *Conversational Agents: Theory and Applications*. https://doi.org/10.1142/9789811246050_0012

Yu, J., Cristea, A. I., Harit, A., Sun, Z., Aduragba, O. T., Shi, L., & Moubayed, N. Al. (2022). *INTERACTION: A Generative XAI Framework for Natural Language Inference Explanations*. http://arxiv.org/abs/2209.01061

Zhang, W., Li, X., Deng, Y., Bing, L., & Lam, W. (2022). *A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges*. http://arxiv.org/abs/2203.01054