Cluster Computing (2024) 27:1–26
https://doi.org/10.1007/s10586-023-04203-7

Foundation and large language models: fundamentals, challenges, opportunities, and social impacts

Devon Myers[1] · Rami Mohawesh[2] · Venkata Ishwarya Chellaboina[1] · Anantha Lakshmi Sathvik[1] · Praveen Venkatesh[1] · Yi-Hui Ho[1] · Hanna Henshaw[1] · Muna Alhawawreh[3] · David Berdik[1] · Yaser Jararweh[1]

Received: 13 May 2023 / Revised: 3 November 2023 / Accepted: 4 November 2023 / Published online: 27 November 2023

Abstract
Foundation and Large Language Models (FLLMs) are models that are trained using a massive amount of data with the intent to perform a variety of downstream tasks. FLLMs are very promising drivers for different domains, such as Natural Language Processing (NLP) and other AI-related applications. These models emerged as a result of the AI paradigm shift, involving the use of pre-trained language models (PLMs) and extensive data to train transformer models. FLLMs have also demonstrated impressive proficiency in addressing a wide range of NLP applications, including language generation, summarization, comprehension, complex reasoning, and question answering, among others. In recent years, there has been unprecedented interest in FLLMs-related research, driven by contributions from both academic institutions and industry players. Notably, the development of ChatGPT, a highly capable AI chatbot built around FLLMs concepts, has garnered considerable interest from various segments of society. The technological advancement of large language models (LLMs) has had a significant influence on the broader artificial intelligence (AI) community, potentially transforming the processes involved in the development and use of AI systems. Our study provides a comprehensive survey of existing resources related to the development of FLLMs and addresses current concerns, challenges and social impacts. Moreover, we emphasize on the current research gaps and potential future directions in this emerging and promising field.

Keywords Natural language processing · Foundation models · Large language models · Advanced pre-trained models · Artificial intelligence · Machine learning

1 Introduction

Natural Language Processing (NLP) is a subfield of Artificial Intelligence that has led to the development of technologies such as chatbots, voice assistants, translation tools, and a myriad of other commonplace tools [7]. NLP research began in the 1950s with the development of the first tools for performing automated translation from Russian to English. This research laid the foundation for all subsequent work including the recent advancements demonstrated by the release of OpenAI's ChatGPT. Alan Turing's "Imitation Game" concept–now commonly known as the Turing Test–was developed around the same time. Turing's concept proposed the idea that it may be possible to develop machines that are capable of thinking in a manner similar to how humans do [214]. This idea is very tightly linked to NLP research given that fundamentally, it involves the development of rules that computers can follow to parse the meaning of written and spoken languages. As it turns out, developing a set of rules for computers to follow for parsing meaning is not very straightforward. Nuances in language often cause miscommunications between people, and thus, it should not be a surprise that a similar issue occurs when developing NLP software [96, 147, 148, 149, 150].

As NLP models have improved, researchers have developed foundation models that are not only useful for performing many different NLP-related functions but can also be used as a starting point for developing more specialized tools [6, 20]. Authors at the Center for Research on Foundation Models at the Stanford Institute for Human-Centered Artificial Intelligence emphasize that foundation

Extended author information available on the last page of the article

models constitute a paradigm shift for AI and the areas in which they are applied, both in positive and potentially negative ways [20]. One characteristic of foundation models is that they can be "fine-tuned" by other developers to enable them to be used for different use cases [221]. Although these technologies have the potential to be incredibly useful, they are not without risks. For example, despite the authors' claims of possible benefits for students, there is a risk that tools such as ChatGPT will aid in cheating [176].

A "Large Language Model" is a large-scale, sophisticated neural network that has been trained with a large number of data samples and computer resources. Instead of training new models from scratch, we may use a foundation model to quickly and easily create new models to tackle a range of downstream tasks. To do this, we should add a rudimentary neural network head (often two fully connected layers) to the foundation model for a given task and then fine-tune it for a few epochs using a small number of task-related data samples. Then, we can construct a model that accomplishes the intended task more effectively [2, 99].

There are numerous downstream tasks in the realm of Natural Language Processing. For example, a sentence classification task (e.g., sentiment analysis) seeks to predict the label of a given sentence; a sequence tagging task can assign a class or label to each token in a given input sequence (e.g., name entry recognition). These downstream tasks have had significant research gaps in the past, necessitating task-specific structures and training methodologies. These various downstream tasks can be solved coherently and efficiently using large language NLP foundation models such as ELMo or BERT. These pre-trained models demonstrated a wide range of linguistic talents and adaptability to a wide range of linguistic contexts, indicating a shift toward a more generalist approach and goal [99].

Domain Adaptation (DA) research has focused chiefly on supervised domain adaptation [2]. A modest amount of labelled target domain data and an enormous amount of labelled source domain data are accessible in a traditional supervised DA setup. In light of limited target domain data, the aim is to adapt the source to the specific target domain. Annotation, on the other hand, is a time-consuming and costly manual process. While annotation directly addresses the absence of labelled data, it is not easy to scale to additional application targets. On the other hand, DA approaches try to move models' abilities to extrapolate to cases beyond the original training distribution from typical interpolation of similar examples [184].

Unsupervised Domain Adaptation (UDA) solves the domain shift problem by learning exclusively from unlabeled target data that is typically available in both the source and destination domains. UDA is more suited to the traditional real-world case in which labelled data in the target domain is scarce, but unlabeled data is plentiful. As a result, UDA provides a slick and scalable solution. These advancements in UDA, we believe, will aid in out-of-distribution generalisation.

This article presents in-depth literature that may support future research and advancement in this research area. This survey is the first comprehensive review covering the foundation model in NLP tasks to the best of our knowledge. The main contributions of this article can be summarised as follows:

- We present foundation models and theory, development, and applications of NLP.
- We also conduct a critical review and summarisation of the current methods to identify significant shortcomings.
- We present research gaps and potential future directions in this field.
- We provide the Social impact of the foundation models in the NLP area.

This paper is organised as follows: Sect. 2 describes NLP applications. Section 3 describes the NLP's theory and the development of the foundation models. Sections 4 and 5 provide the applications of NLP Foundation Models. Section 6 presents the Societal Impact of NLP Foundation Models. Section 7 provides the current limitations in this research area and some future directions. Section 8 provides the conclusion of this survey.

2 Natural Language Processing

Natural Language Processing is concerned with how computers communicate with human language, particularly how to train computers to manage and analyze vast volumes of text and other natural language data. NLP use artificial intelligence to absorb real-world data, integrate it, and interpret it in a way that a computer can comprehend, regardless of the language. Computers have input devices to read and microphones to collect sounds, similar to how humans have sensors such as ears and eyes to hear and see. Similar to how humans have a brain to process sensor data, computers have programmers to process their inputs. At some point during the method, the input is turned into computer-manageable and ultimately interpretable code. There are two main phases of NLP: data pre-processing and algorithm development. Data pre-processing is the task of accepting a large amount of data to be analysed and cleaned into an appropriate text format for the developed algorithm can handle [103].

2.1 Applications of NLP

- Text Generation is an area of study within natural language processing. It utilises computational languages and artificial intelligence to automatically generate natural language documents that satisfy certain communicative needs [2]. The language model (LM) is neither constrained nor driven by any particular rules or expectations when generating text. The only hope is to develop content that is believable, clear, and easily understood. Controllable Text Generation falls under the scope of text generation. Controllable text creation provides genuine sentences with controllable characteristics. For instance, we can define properties of the text to be generated, such as sentiment and happiness [99].
- Text Categorisation is the automatic assignment of natural language texts based on their content to predetermined categories. Its uses include categorizing texts to facilitate document retrieval and data extraction from texts[2]. Text categorisation tends to produce high accuracy using suitable statistical methods. Text categorisation also provides expert tag assignments, leading to document conceptual views. Overall, text categorisation has an enormous scope of applications [85].
- Text Summarisation is used to minimise a large document to a short document, listing important and highlighted points. Automatic summarization means an automatically summarised output is given when an input is applied [60]. Manually summarising documents could be complicated if the supporting documentation is massive, so utilising a model would be prudent as long as the model's summary is comprehensive and accurate. One challenge of text summarisation for a model scraping from the Internet is that the Internet can have too much information. Therefore, it is crucial to separate relevant documents from the web of documents available. Text summarisation methods can be classified into extractive and abstractive summarisation. Abstractive summarisation is a method for novel phrasing describing the content of the text, which requires heavy machinery from natural language processing [101]. Extractive summarization is a method for determining salient text units (typically sentences) by looking at the text unit's lexical and statistical relevance or matching phrasal patterns. An extension of text summarisation is the question-answering problem [12]. The goal of solving the question-answering problem is to provide an input (question) such that the output is a summary that answers the input question [126].
- Machine Translation is the task of processing one natural language input to another natural language keeping the original context meaning. For example, machine translating would take the natural input data of an English language greeting "Hello" to provide output data of a Telugu language greeting "Namaste". The two main methods used are statistical machine translation (SMT) and example-based machine translation (EBMT); these two use a bilingual corpus, or collections of translations, as a basis for new translations. Both methods take a passive approach to finding patterns and consistencies in translation examples to learn how to translate new words[112].
- Spelling Correction is one of the applications in NLP that is used to correct spelling as needed, not only for the author who may have had a typographical error but also for the model to be able to clean noisy data. Large datasets may contain many type errors, spaces, and special characters, leading to inaccurate search and summarisation results. To eradicate these issues, we use spelling correction methods. Many online tools and extensions are available to correct typographical errors giving suggestions to make the user experience easier [165].
- Sentiment analysis is a technique that employs analytical, large language models, and machine learning techniques to identify the emotional significance of information. One of the significant applications where sentiment analysis is used is on social media networks. Social networks like Facebook and Twitter have been hubs for data that is primarily of interest for product marketing [141]. Twitter specifically is a focal point for sentiment analysis since the purpose of Twitter is to share instantaneous thoughts and feelings through tweets. Sentiment analysis can parse tweets into positive and negative attributes, which product marketers can use to discern reviews of a product or advertise a product to solve an issue [141].

3 Theory and Development of Foundation Models

Foundation models train large amounts of unlabeled data by self-supervised learning that can be used for downstream tasks [20]. The emerging capabilities of machine learning come from the ability of models to learn 'how' a task should be completed through training examples. In the latter of the two ways, foundation models allow for the amalgamation of algorithmic learning methodologies, dubbed homogenization. Neural networks (i.e., deep learning) are used to train many recent NLP foundation models. Deep learning allows foundational models to be adaptable enough to be specialized for myriad downstream

tasks. These models, however, require an enormous amount of computational power for training and fine-tuning model parameters [191]. GPT-3 alone has 175 billion parameters that can be trained for various NLP tasks [134]. Newly developed large language models continue to grow in size.

3.1 How are foundation models trained?

Foundational models can be easily modified to a wide variety of downstream tasks by training on a large, diverse dataset (parencite: FM). Pre-training and other simple transfer learning concepts form the basis of the models. When given access to large amounts of data, machine learning and deep learning can display promising new abilities that can boost performance on many downstream tasks. There has been an unusual amount of standardization because of foundation models. These days, the most cutting-edge NLP models all derive from the same handful of base models like BERT, Roberta, BART, T5, etc[20]. All AI systems may inherit the same problematic biases of a few foundation models, which is a liability despite the fact that this homogenization produces exceptionally high leverage (any changes in the foundation models might lead to instantaneous benefits across all of NLP) (Fig. 1).

3.2 Pre-trained models

Due to the rise of deep learning, the variety of available model parameters has expanded significantly. In order to properly train model parameters and forestall overfitting, a significantly larger dataset is required. However due to the relatively expensive annotation costs, particularly for syntactic and semantically related jobs, creating large-scale labelled datasets is a huge problem for most NLP tasks [167]. Learning the parameters of deep neural networks has traditionally been accomplished by pre-training, with subsequent tasks serving as fine-tuning. Using greedy layer-wise unsupervised pre-training followed by supervised fine-tuning, deep learning made its first major advancements in 2006[167]. Pre-training has been around for decades, but pre-trained models (PTMs) have only recently piqued the interest of scientists [77]. As with much of natural language processing, the early trials of pre-trained models were based on trying to mimic human thought processes, such as applying prior solutions to a new problem [77].

For an example of pre-training, let us focus on BERT (Bidirectional Encoder Representations from Transformers), a popular foundation model used by Google. BERT is regarded as being eloquently simple since the pre-training involves analysing context to the left and right of specific

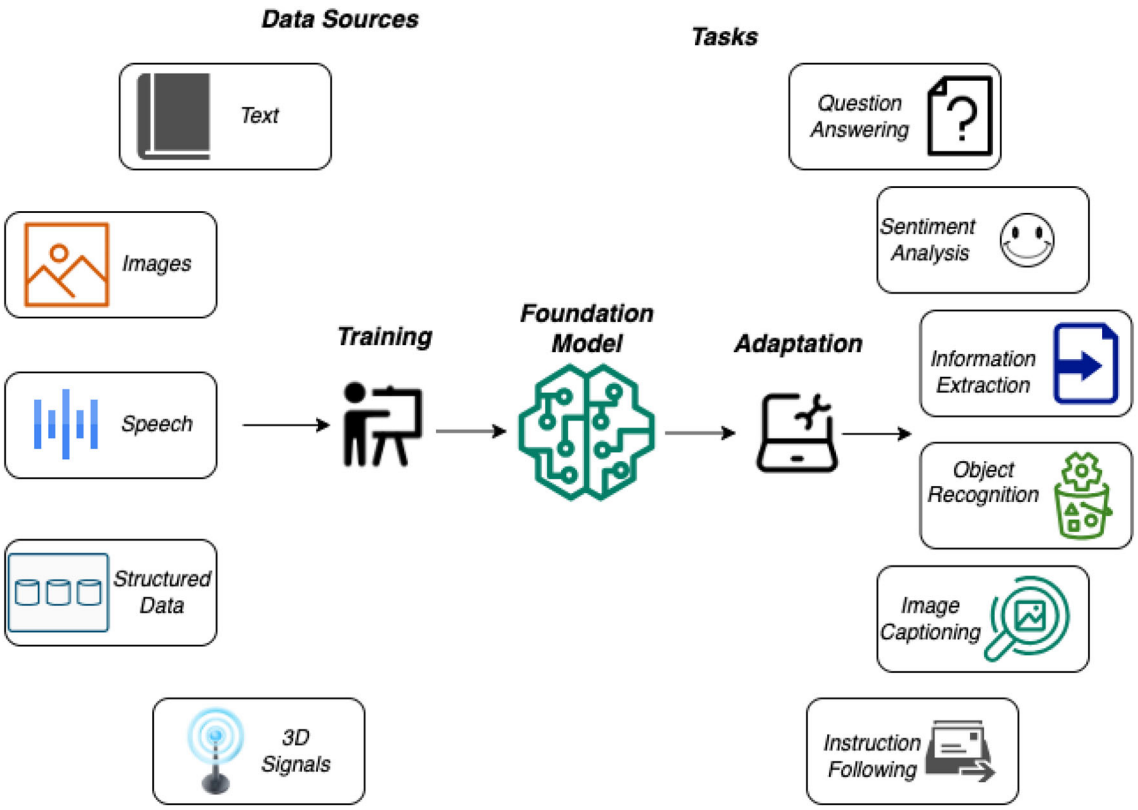


Fig. 1 Foundation model utilizes a large amount of data to complete a variety of downstream tasks

text. From there, BERT can handle tasks such as question answering or language inference after tuning the model [84].

3.2.1 Re-purposing(reusing) a pre-trained model

Re-purposing a pre-trained model is a process of removing the original classifier, then adding a new classifier that fits our purposes and fine-tuning our model according to our needs and strategies. Fine-tuning a pre-trained model includes deep transfer learning(DTL) training a pre-trained model in a smaller, labelled dataset. There are two generations of pre-trained models, First-Generation PTMs: Pre-trained Word Embeddings and Second-Generation PTMs: Pre-trained Contextual Encoders[167].

3.2.2 Transfer learning

Transfer learning is used to build accurate models while also cutting time costs because we can use a trained model that solves one problem to another. As one can sense from the similarities in definitions, pre-training is a crucial component of transfer learning. The basic idea is to use a model that solves a similar problem to the task at hand as a starting point to be altered and tuned for the specific task [167]. Deep transfer learning has the same concept as transfer learning but includes many labelled data for supervised learning or unlabeled data for unsupervised learning. Deep transfer learning has applications within NLP.

There are two widely used pre-training approaches to transfer learning: feature transfer and parameter transfer [77]. Feature transfer methods focus on retraining the model's features or attributes to be applied across tasks and domains. [77]. Parameter transfer methods have an assumption that one model's task and the resulting pre-trained model will share parameters or hyper-parameter distributions [77].

3.3 OpenAI Technologies

OpenAI is an AI research Laboratory whose stated goal is to promote and develop friendly AI that can benefit humanity in a way that humans can perform tasks independently [60]. One of the OpenAI technologies GPT-3 (Generative Pre-trained Transformer), is a third-generation, autoregressive language model that uses deep learning to produce human-like text [60]. Davinci and Ada are similar models, but GPT-3 is the most capable model.

GPT-3 is a computational system designed to receive input and produce sequences of words, code, or other data [60]. For example, machine translation predicts word sequences statistically. The system requires a large amount of data to produce relevant results; basically, the language model we trained is an unlabelled dataset made up of texts such as Wikipedia and many other website sources, primarily in English but may include other languages too. Recently, To answer user enquiries in a comprehensive and natural way, ChatGPT makes use of dynamic forms [127]. ChatGPT can generate high-quality text outputs in response to a prompt. ChatGPT plus, which is based on GPT-4, also accepts photos as input. ChatGPT succeeds at text-to-text tasks beyond its primary function as a chatbot, including, completion [53], summarization [53], classification [101], translation [37], and more.

In addition, OpenAI conducts studies on reinforcement learning, an AI subfield that tries to apply machines to make decisions based on the potential outcomes of those decisions rather than on their perceived value to the system. OpenAI has developed a variety of reinforcement learning algorithms, including Soft Actor-Critic (SAC) [106], Proximal Policy Optimisation (PPO) [211], and Trust Area Policy Optimisation (TRPO) [199]. These algorithms have been used to teach agents to do everything from playing video games to operating robots. The OpenAI Gym [76] is one of the many software tools developed by OpenAI so far to aid in its research efforts. It is a set of tools for developing and comparing various reinforcement learning algorithms. OpenAI has invested in a number of high-performance processing systems, including NVIDIA's DGX-1 and DGX-2 systems [150]. The computing capacity required to construct complex AI models is available on these platforms since they were designed with deep learning in mind. DALL-E [141], Whisper [135], and Codex [25] are a few of OpenAI's other well-known technologies outside of ChatGPT.

3.4 Spark NLP

To facilitate the training of NLP models that can be seamlessly integrated into downstream machine learning (ML) or deep learning (DL) pipelines, a good NLP library must accurately turn free text into structured features [206]. Spark NLP is the only library that can scale up for training and inference in any Spark cluster, take advantage of transfer learning, apply the latest and greatest algorithms and models in NLP research, and still provide a mission-critical, enterprise-grade solution [206]. It is an Apache Spark and Spark ML-based open-source NLP library. There are two variants of the Spark NLP library available: open-source and commercial. The latest deep learning (DL) frameworks and research trends are fully included in the

open-source version, along with all the features and components one would expect from an NLP library. The enterprise library is a paid upgrade to the original open-source library, with the intention of addressing real-world issues in the healthcare sector. Figure 2 is a flowchart illustrating the Spark NLP processing chain.

Spark NLP's annotators utilise rule-based algorithms. Machine learning and deep learning models implemented using TensorFlow are known for their accuracy, speed, scalability and memory utilisation. The Spark NLP is written in Scala and provides various APIs in Python, Java, Scala and R

4 NLP Foundation Models

Foundational models offer a wide range of uses and have been demonstrated to be beneficial for acquiring universal language representations. In the last few years, the combination of foundational models with natural language processing has become more powerful and efficient. The foundation model has achieved astounding results in various language comprehension tasks, including text classification or phrase categorisation, the semantic similarity between pairs of sentences, question-answering tasks involving paragraphs, text summarization, etc. Figure 2 shows the flow diagram of the spark NLP pipeline.

5 Applications of NLP Foundation Models

The foundation model has a wide range of applications and has proven useful in learning universal language representations. In recent years, the combination of the foundation model and natural language processing has gotten closer and closer. The foundation model has achieved outstanding results in many language understanding tasks such as text classification or sentence classification, the semantic similarity between pairs of sentences, question-answering tasks with paragraphs, text summarization, etc. The table below summarises the recent advanced Transformer-based models used in natural language processing tasks.

5.1 NLP Foundation Models that We Interact with Daily

Artificial intelligence (AI) is going through a paradigm shift with the rise of models trained at a massive scale. An emerging paradigm for building AI systems is defined, which is called foundation models [20]. Foundation models are not something new. They have been with us for several years. These models like GPT-1 [171], BERT [45], XLNet [223], Roberta [124], ELECTRA [34], T5 [174], ALBERT [108], BART [109] and PEGASUS [231] have achieved tremendous success in NLP because of their ability to learn universal language representations from large volumes of unlabeled text data and then transfer this knowledge to downstream tasks. Moreover, [61] introduced GPT-3, a third-generation, autoregressive language model that uses deep learning to produce human-like texts and uses the previous distinction to analyse them. With the progress that deep learning has provided, a new framework is proposed: generative adversarial nets (GAN) for estimating generative models via an adversarial process, which shows the potential of GAN through qualitative and quantitative evaluation of the generated samples [69]. BERT, a creation of Google, stands as another fundamental model in AI. This model emphasizes comprehending language bidirectionally. It undergoes pre-training through predicting omitted words in sentences and can later be fine-tuned for diverse NLP tasks. This process enables BERT to achieve leading-edge outcomes in tasks such as sentiment analysis, identifying named entities, and others [45]. GPT-1, created by OpenAI, marked an early milestone in foundational AI models. This transformer-based language model was pre-trained using extensive text data and has the ability to produce coherent text in response to prompts. Although its impact was noteworthy, subsequent iterations such as GPT-2 and GPT-3 demonstrated even more remarkable advancements. The BART model employs a bidirectional encoder to encode a corrupted input sequence, followed by a left-to-right decoder to reconstruct the original text. The

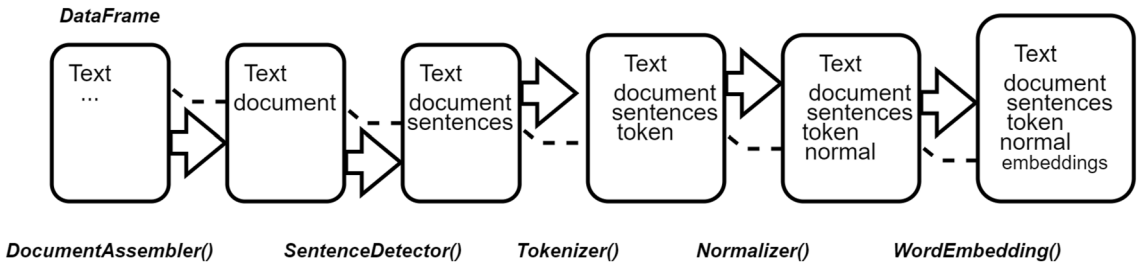


Fig. 2 The flow diagram of Spark NLP Pipeline

researchers involved in the development of BART conducted a series of experiments to explore different corruption methodologies. Ultimately, they trained the model using phrases that had been corrupted by the techniques of sentence permutation and text infilling. Later, The concept of PALM (Parallel Additive Layer Manufacturing) has been introduced in recent years [19]. For instance, models such as MASS (Masked Sequence-to-Sequence) and BART (Bidirectional and AutoRegressive Transformer) employ a bidirectional encoder to process corrupted text and a left-to-right autoregressive decoder to restore the original text. The PaLM model [154] has undergone pre-training using a large corpus consisting of 780 billion tokens. This corpus encompasses a wide range of natural language tasks and use cases, ensuring a high level of quality in the training process. The dataset encompasses a range of textual sources, such as webpages, books, Wikipedia articles, news items, source code derived from open-source repositories on GitHub, and social media interactions, which have been subjected to a filtering process. The Large Language Model Meta AI (LLaMA) is a collection of expansive language models (LLMs) that were introduced by Meta AI, commencing in February 2023 [201]. The initial iteration of LLaMa involved training four different model sizes, specifically 7, 13, 33, and 65 billion parameters. According to the developers of LLaMA, the performance of the 13B parameter model outperformed the far bigger GPT-3 model (with 175B parameters) on various natural language processing benchmarks. Additionally, they noted that the largest LLaMA model demonstrated competitiveness with state-of-the-art models like PaLM. Bard is an artificial intelligence chatbot created by Google, primarily utilising the LaMDA and PaLM large language models (LLMs) for conversational generation [21]. The development of the system was initiated in direct response to the emergence of OpenAI's ChatGPT.

5.2 Text Categorization Foundation Models

Text categorisation, also known as text classification, is a classical problem in natural language processing (NLP), which aims to assign labels or tags to textual units such as sentences, queries, paragraphs, and documents. In the past few decades, numerous models have been proposed for text classification, from traditional models to foundation models [115, 144]. The sparse feature information causes difficulties in text classification. An improved BERT pre-training model combined with a Wide &Deep-BERT model obtained a more effective and practicable method for news classification and user preference recommendation [89]. Similarly, foundation models can be applied for Turkish text classification with a large corpus, which obtained less effort and high performance on time and hardware cost [160]. Moreover, in another work, the goal was to classify the Tamil news articles to their related topics (Sports, Cinema, Politics) and then compare the existing TFIDF feature learning along with Pre-Trained embeddings given to Convolutional Neural Networks (CNN) [177]. The empirical result shows that CNN with pre-trained embeddings outperforms the performances of recent state-of-the-art models.

Managing a large volume of textual data by assigning each text to one or more categories is arduous and time-consuming. For Arabic text, this task becomes more challenging due to the complex morphology and structure of the Arabic language. To address this issue of multi-label classification, a classification system based on AraBERT, a pre-trained BERT model for the Arabic language, was proposed [66]. This system used AraBERT on the input articles and applied a fully connected neural network on the output of the AraBERT model to classify the articles according to their topics. Also, understanding the complex nature of unstructured data and the relations of features that exist in textual documents is a challenging task. Deep learning methods have been explored with pre-trained word embeddings, and pre-trained transformers such as DistilBERT and X-Transformer for extreme multi-label text classification [28, 57, 182].

Because a foundation model is meant to be adapted, one study assessed different fine-tuning methods of BERT and found a general solution for text classification utilising BERT; the solution obtained new state-of-the-art results [193].

The mobile application under consideration is specifically developed for the purpose of aggregating news content. Concurrently, the application aspires to personalize article recommendations in alignment with the unique preferences of individual users. To realize this ambitious goal, the application harnesses an enhanced BERT pre-training model, synergistically coupled with a Wide & Deep-BERT model. This symbiotic integration enhances the application's cognitive grasp of both article textual content and user interaction patterns. Consequently, the application evolves into a proficient tool for precisely categorizing news articles and furnishing users with contextually fitting content recommendations. This augmentation significantly amplifies user engagement levels and augments their overall content consumption experience within the application [193].

5.3 Text Generation Foundation Models

Text generation is one of the most important yet challenging tasks of the various applications of NLP foundation models. Text generation aims to produce plausible and readable text in human language from input data. With the

development of foundation models, text-generation technology has undergone tremendous changes and provided many kinds of services for human beings, such as conversation, summarisation, and translation. There are various aspects of applying pre-trained language models (PLMs) to text generation [110, 111]. One specific way is to perform Controllable Text Generation using transformer-based PLMs [230]. Another approach is to perform conditional text generation (CTG), including emotional text generation, personalised text generation, and so on [72, 133, 216]. Several general learning models for CTG, such as BERT, GAN, CoBERT, and CG-BERT, effectively leverage a large pre-trained language model to generate text conditioned on the intent label [72, 133, 216]. While PLMs could store specific knowledge/facts from the training corpus, their knowledge awareness is still far from satisfactory. Integrating knowledge into PLMs has recently become a very effective research area to address this issue, and various approaches have been developed. Several works and studies have provided a comprehensive approach analysis on knowledge-enhanced pre-trained language Models (KE-PLMs), knowledge-enhanced text generation, and knowledge-enhanced multimodal BART (KM-BART) as solutions [212, 217, 224].

Most large-scale models such as T5 and GPT-3 are trained in an auto-regressive way. This traditional fine-tuning approach demonstrates relatively weak performance when solving downstream language understanding tasks. In order to solve the above problems, a unified framework named ERNIE 3.0 was proposed [194]. ERNIE 3.0 fuses auto-regressive and auto-encoding networks for pre-training large-scale knowledge-enhanced models for language understanding and generation [194].

Generating text for video games helps players immerse themselves in virtual worlds and take on the role of virtual characters. GPT2 was fine-tuned for generating NPC (non-player character) quests in RPG (role-playing game) [204]. Another possible solution is to use question generation as a pre-training method, which is better at understanding the essence of the input and is a better language model for the target task [155].

Giving starting words to make a sentence or filling in sentences is a crucial task in NLP, which reflects whether the machine can have human thinking and creativity. The OpenAI GPT-2, GPT-3 and BERT models are currently widely used language models for text generation, and prediction [30, 121, 132]. Other researchers applied two new corpora to train the OpenAI GPT-2 model, used it to generate long sentences and articles, and finally performed a comparative analysis [169]. Meanwhile, it exploited the BERT model to predict intermediate words based on the context. The provided examples offer a clear portrayal of the inputs, outputs, and syntax associated with a range of text generation tasks discussed in the given text. For instance, Controllable Text Generation employs pre-trained language models (PLMs) to generate text aligned with specific attributes, such as sentiment or style, given a prompt. Similarly, Conditional Text Generation delves into conveying emotions by inputting a sentence or context alongside an emotional label. Moreover, Knowledge-Enhanced Text Generation integrates relevant facts with input text to enhance coherence and context. The ERNIE 3.0 Framework combines auto-regressive and auto-encoding networks to pre-train models that excel in comprehending and generating text. In gaming contexts, Non-Player Character (NPC) Quest Generation constructs quests within role-playing games, while Question Generation pre-training enhances language understanding. Lastly, initiating sentences or completing phrases illustrates creativity and human-like thinking. It's noteworthy that the provided syntax simplifies the input format, with real-world execution potentially involving more complex requirements. The generated text varies in length, dependent on the task and the model's response, thereby showcasing the multifaceted nature of text generation processes.

5.4 Question-Answering Foundation Models

The tasks related to question answering and conversational search encompass a variety of inputs, outputs, and syntax that drive their functionalities. For instance, in Extractive Question Answering using BERT, a passage of text and a question lead to an output span of text containing the answer. The innovative approach of Structured Pruning and Distillation involves modifying pre-trained BERT-based models like BERT or Roberta by selectively pruning parameters to create more compact, task-specific models without the need for pre-training distillation. Multi-passage BERT facilitates the normalization of answer scores across passages for a given question, while Fine-Tuning BERT for Answer Span Identification refines its ability to predict answer spans within passages. In the context of domain-specific questions, BERT is harnessed to provide health-related answers by incorporating medical knowledge. For semantic question retrieval, a modified BERT embedding integrates latent Dirichlet allocation (LDA)-derived topic information to retrieve related historical questions. BERT's applicability extends to predicting community question-answer quality and enhancing Conversational QA by embedding conversation history for context-aware answers. Furthermore, in Conversational Recommender Systems, BERT-driven insights into user preferences and interactions offer refined recommendations for books, movies, and music. These tasks collectively illustrate the versatility and potential of BERT in addressing diverse NLP challenges with distinct inputs, outputs, and syntax, albeit recognizing

the complexity of actual implementations. Question answering (QA) is a critical NLP problem. It allows a user to express a question in natural language and get an immediate and brief response. The extractive QA task is transformed into the spans prediction task by BERT [45]. In industry-setting Natural Language Processing (NLP) research, it needs to operate large pre-trained language models like BERT under strict computational limits. However, most model compression work has focused on "distilling" a general-purpose language representation using expensive pre-training distillation. A new study is BERT and Roberta-based question-answering systems by structured pruning of parameters from the underlying transformer model, which can build an inexpensive combination of task-specific structured pruning and task-specific distillation without the expense of pre-training distillation. However, obtaining highly-performing models [140]. Training BERT by viewing passages corresponding to the same question as independent training instances may cause incomparable scores for answers from different passages. To tackle this issue, utilising a multi-passage BERT model to globally normalise answer scores across all passages of the same question enables the QA model to find better answers by utilising more passages [209]. Furthermore, fine-tuning pre-trained BERT is sufficient to achieve high accuracy in identifying answer spans [221, 222]. While pre-trained language models have shown success in capturing syntactic, semantic, and world knowledge from text, they can also be further complemented by specific information like knowledge of symptoms, diagnoses, and treatments for health questions answering [78].

Community question-answering (CQA) services satisfy users' quest to seek answers from experts and share knowledge in any domain. To eliminate duplicate questions and avoid indefinite waiting time to get responses, retrieving semantically related historical questions for a new query is crucial in CQA. One challenge in question retrieval is the lexical gap between the new question and the question in the archive that is already answered. One possible solution is a modified BERT embedding using the topic information obtained by latent Dirichlet allocation (LDA) on top of questions pre-processed using the RAKE keyword extraction algorithm [195]. Moreover, an integrated question-answer quality model for community question-answering websites is inadequate. Using BERT to tackle the quality Q &A modelling problems can boost the performance of automated quality prediction [103, 186].

Conversational search is an emerging topic in the information retrieval community. However, one of the significant challenges to multi-turn conversational search is to model the conversation history to answer the current question. Conversational Question Answering systems are often criticised for not understanding or utilising the previous context of the conversation when answering the questions. A highly effective history answer embedding approach that seamlessly integrates conversation history into a conversational question answering (ConvQA) model built on BERT could help remedy the challenge [18, 95, 168, 225]. In addition, BERT can also be applied in Conversational Recommender Systems (CSR) on books, movies, and music and obtain better performance by analysing the knowledge stored in BERT's parameters [163].

5.5 Machine Translation Foundation Models

Machine Translation (MT) is a pivotal application of NLP foundation models, with the encoder-decoder framework being crucial for Neural Machine Translation (NMT) models. This involves encoding input tokens through the encoder and decoding output tokens in the target language via the decoder. The Transformer architecture has significantly enhanced NMT, prompting researchers to adapt and refine its modules to amplify performance. Leveraging foundation models for diverse NMT tasks has gained substantial attention. Challenges remain for languages beyond English due to data and computational constraints. Strategies include integrating pre-trained models such as BERT, ELMo, and mBERT with NMT models to address these limitations. Performance improvements have been noted by reframing lexical embeddings without altering transformer layers. Additionally, unsupervised pre-training initiates encoder and decoder weights using two languages, subsequently fine-tuning the model on parallel data for remarkable unsupervised translation outcomes. Incorporating multilingual pre-trained models like M-BERT and XLM-T into encoders and decoders has demonstrated significant enhancements across various MT scenarios. Novel methods, like integrating adapter modules in NMT and harnessing BERT for contextual information encoding, have shown promise. Nonetheless, a research gap exists in maximizing BERT's potential for NMT tasks. Approaches like BERT-JAM fusion and the BERT-fused model offer innovative ways to integrate BERT within NMT. Moreover, a novel technique involving joint pre-training of the encoder and decoder has exhibited potential superiority over traditional methods in unsupervised and supervised MT settings [9, 112, 166]. While large generative language models have succeeded in English, other languages lag due to data and computational limitations. One approach is integrating pre-trained models (e.g., ELMo, BERT, mBERT, BiBERT, GPT-2) with the NMT model to overcome these problems [35, 41, 54, 218]. This approach demonstrates performance improvements on English-Turkish, German, Italian and Dutch NMT models by retraining lexical embeddings without tuning the

transformer layers. Another approach is unsupervised pre-training for neural machine translation, which initialises the weights of the encoder and decoder with two languages and then fine-tunes the model on parallel data using Elastic Weight Consolidation (EWC) to avoid forgetting the original language modelling tasks [205]. Additionally, using GPT -3's zero-shot translation capability of large pre-trained language models achieves a new state-of-the-art unsupervised translation on the English-French benchmark [76].

A multilingual pre-trained model, such as M-BERT and XLM-T, can be used to initialise the entire encoder and decoder [29, 107, 129, 190]. They show a significant improvement in unsupervised MT and English-Romanian and Nepali-supervised MT. The bilingual masked language model pre-training is enhanced with lexical-level information using type-level cross-lingual subword embeddings [33]. Training neural machine translation (NMT) systems with a dataset of more than 40 billion bilingual sentence pairs are arduous. Practical solutions are proposed to handle these issues and demonstrate that large-scale pre-training significantly improves performance boost of +3.2 over existing state-of-the-art results in WMT17 Chinese-English database [143]. However, effectively and efficiently incorporating foundation models into natural language generation tasks remains challenging. There is an efficient way to apply BERT in NMT [73]. This reference integrates two pre-trained BERT models by introducing lightweight adapter modules inserted between BERT layers and tuned on downstream tasks. Apart from that, BERT can be leveraged to encode the contextual information for NMT [112]. These methods further achieve a noticeable improvement.

There still exists a research gap in studying how to maximise the utilisation of BERT for NMT tasks. BERT-JAM with fusion modules is proposed for composing BERT's multi-layer representations to be integrated with the encoder/decoder representations that can be leveraged by the NMT model [233]. The results show that BERT-JAM achieves state-of-the-art scores on multiple translation tasks. While BERT is more commonly used as fine-tuning instead of contextual embedding for downstream language understanding tasks, in NMT, the preliminary exploration of using BERT as contextual embedding is better than for fine-tuning. A new algorithm named the BERT-fused model is introduced, which uses BERT to extract representations for an input sequence, and then the representations are fused with each layer of the encoder and decoder of the NMT model through attention mechanisms [234].

Unlike only pre-training the encoder, one method is a pre-trained sequence-to-sequence model that utilises pre-training the encoder and decoder jointly [114, 210]. In the experiment, this approach can surpass the BERT-style pre-training for unsupervised MT and supervised MT [107].

5.6 Sentiment Analysis and Opinion Mining NLP Foundation Models

Foundation Models have emerged as a powerful machine learning technique and produce superior prediction results. Along with the success of foundation models in many application domains, foundation models are also used in opinion mining and sentiment analysis, which deals with the computational treatment of opinion, sentiment, and subjectivity in text. In [139, 232], it discusses various word embedding methods used for sentiment analysis, followed by an overview of state-of-the-art pre-trained models. BERT has been widely used in sentiment analysis, outperforming previous state-of-the-art models by simply fine-tuning [40]. [16] utilised BERT with transfer learning-based approaches and achieved an outstanding result in Japanese SA. While various techniques for sentiment analysis, recently, word embedding methods such as Word2Vec and GloVe have been widely used in sentiment classification tasks. However, word embeddings ignore the sentiment information of texts and need a considerable corpus of texts for training and generating exact vectors. To overcome this problem, [180] propose a novel method to Improve Word Vectors (IWV), which increases the accuracy of pre-trained word embeddings in sentiment analysis.

Although applying BERT in simple sentiment classification is successful, it shows less considerable improvement in leveraging BERT to aspect-based sentiment analysis (ABSA), which is a fine-grained SA task [219]. To better exploit the powerful representation of BERT, [192] constructed an auxiliary sentence by converting ABSA from a single sentence classification task to a sentence pair classification task. In addition, enhancing the performance of fine-tuning BERT is critical. [93, 117, 189] utilised BERT intermediate layers and exploited BERT for end-to-end aspect-based sentiment analysis, which can outperform state-of-the-art works. [118] introduced a new method named GBCN, which uses a gating mechanism with context-aware aspect embeddings to enhance and control the BERT representation for aspect-based sentiment analysis. Collecting labelled data for ABSA to help neural networks generalise better can be laborious and time-consuming. They were fine-tuning the general-purpose BERT and

domaining specific post-trained BERT (BERT-PT) by adversarial training to produce artificially similar data to real-world examples. [94] proposed a novel architecture called BERT Adversarial Training (BAT) to utilise adversarial training in SA, which outperforms the general BERT and the in-domain post-trained BERT in both tasks. Furthermore, [117] used the BERT Base Uncased model, which is a powerful deep learning model to elucidate the issue of sentiment analysis, and the BERT model obtained an improved performance with a good prediction and high accuracy compared to the other methods of Machine Learning in the experimental evaluation.

While GloVe and Word2vec embedding models have been widely used for feature extractions, these techniques generate vectors for just those words that are included in their vocabulary and ignore Out of Vocabulary Words (OOV), which can lead to information loss. [197] employed the BERT-based Convolution Bi-directional Recurrent Neural Network (CBRNN) model for exploring the syntactic and semantic information along with the sentimental and contextual analysis of the social media data. Instead of classical convolution to extract contextual semantic features, the authors proposed Bi-directional Long Short-Term Memory (Bi-LSTM) [170] which is used for the entire sequencing of the sentences. Experiments show that CBRNN can be efficiently used for performing SA tasks on social media reviews without any information loss. Moreover, [75] shows the effect of BERT, ELECTRA, and ALBERT language models on sentiment analysis for Turkish Product Reviews and.[46] applied the Albert model on sentiment analysis of movie reviews. Compared with the traditional LSTM and GRU, the accuracy of the Albert model is improved by 3

Opinion mining has arisen as a method for mining people's views on several issues using foundational Models. It extracted the Indian people's opinions on vaccines by leveraging the BERT and Roberta models and produced promising results compared to other models on the collected Twitter dataset. Investigating opinion mining for app reviews helps make strategic decisions to improve app quality. In [11], it compares different techniques for the textual representation of reviews from traditional Bag-of-Words to the most recent state-of-the-art Neural Language models such as BERT, DistilBERT, and LIME. Despite the fact that BERT has been widely overcome by the existing methods in several NLP tasks, upgrading the performance of most opinion mining tasks is still a big challenge. [1] introduced a novel model for fine-grained aspect-based opinion mining named FGAOM. It used adjusted BERT as an embedding layer and proposed Multi-head Self-Attention (MSHA) to effectively fuse internal semantic text representation and take advantage of convolutional layers to model aspect term interaction with surrounding sentiment features. Results show that the proposed model's performance outperforms recent models' performances.

5.7 Some Additional NLP Foundation Model Tasks

Sections 5.2 through 5.6 of this paper summarized several vital NLP tasks where foundation models proved to be useful tools for solving them. This section aims to give readers insight into other NLP tasks where foundation models are helping researchers gain traction.

First, part-of-speech tagging (POS-tagging) is a popular NLP task that aims at labeling words in a text corpus by their syntactic class [146, 185]. That is, POS tagging aims to label words in a corpus as nouns, verbs, conjunctions, prepositions, adverbs, adjectives, and so on based on the context in which the word is used in a sentence. POS-tagging can be challenging when the target dataset (the problem of interest to be solved) does not syntactically match the data that the model was trained upon [142]. For instance, suppose we wish to use a pre-trained model for POS-tagging in the plays of William Shakespeare. If the model we wish to use were trained on Tweets, which have vastly different verbiage and symbols than that of Shakespeare's plays, we could expect the performance of the model to suffer. Therefore, one should always be cautious and consider the data that a model is trained on and the degree of similarity the training data has to the target data. In any case, POS tagging can be used for pre-processing and filtering textual data, such as when one needs only one part of speech (i.e., nouns, verbs) from a corpus.

In addition to the previously mentioned tasks, another task on which NLP foundation models have performed well is entity recognition (NER). The main goal of NER is automating a system that is able to flag named entities, such as proper nouns, in a piece of text [151, 153]. NER is one of the cornerstones of information extraction because named entities are generally the most important piece of information in a sentence. Being able to identify named entities in text corpora can lead to improved question-answering systems, recommendation applications, and so on [67].

One area where NER systems are particularly powerful is in biomedical NER (BioNER). The challenge for BioNER stems from multiple sources, namely the use of many domain-specific acronyms and context-dependent named entities in biomedical documentation, but an NLP foundation model called BioALBERT has been constructed using an ALBERT model [156]. BioNER can be useful for automating a system to determine things like interactions between drugs [119] and drug-treatment interactions [181], which can make the work of clinicians easier. Foundation models can make constructed BioNER

systems simple to construct, provided that a corpus of biomedical documents is available.

Paraphrase generation is another crucial NLP task where foundation models have performed well. The goal of paraphrase generation is to summarize an input sentence or document so that the meaning of that input is preserved. One example of an NLP foundation model used for paraphrasing is that which was constructed by Hegde and Patil [79] using Open AI's pre-trained model GPT-2 [172]. Paraphrasing can be useful for a myriad of purposes, such as augmentation and the curation of data, generating synthetic training data for small datasets, making a textual dataset more robust by adding linguistic variability and enhancing the results of downstream classification tasks [79]. Additionally, paraphrasing can improve the performance of question-answering systems by presenting a question and its paraphrases to the QA system [23, 50]

The final NLP task we discuss here is natural language inference (NLI). NLI seeks to determine if a hypothesis h can be inferred from a natural language statement p through logical reasoning as well as semantic knowledge and contextual knowledge [22, 130]. Similar to paraphrasing, NLI can be used to increase performance for downstream NLP tasks such as QA, SA, and text summarization. The Stanford Natural Language Inference corpus consists of 570K labelled sentence pairs, one of the largest corpora specifically for NLI, and was compiled by Bowman et al. in 2015 [22]. Interestingly enough, a model capable of NLI, called InferLite, that does not rely on context, word order, or positioning, was constructed by Kiros and Chan in 2018 [98].

5.8 Multilingual and Non-English NLP Foundation Models

Much of the discourse in this survey has revolved around NLP foundation models geared toward tasks in the English language. However, there are many foundation models capable of performing NLP tasks in Non-English languages, such as the Arabic model AraBERT model [10, 48, 58, 66] mentioned in section 5.2, and even models that can do NLP tasks in several different languages, such as M-BERT and XLM-T [29, 107, 129, 190]. As discussed in section 5.5, foundation models are able to translate languages, however, tasks such as POS-tagging, NER, QA, and SA are all applicable to non-English languages. Table 1 summarizes a few NLP foundation models, monolingual and multilingual, for a myriad of languages.

Multilingual language models have the ability to perform NLP tasks in two or more languages and can be constructed in a variety of ways. One way to construct these models is to exhaustively train them using corpora for each language, which requires large amounts of both data and computing power. On the other hand, it is possible to train a deep-learning model on a large corpus containing a concatenation of many different languages. The performance of models constructed in the latter manner can suffer, especially when the vocabulary available in the corpus for each of the languages is small [131, 208]. The work of Madyastha et al. sought to rectify the decrease in performance of multilingual language models by employing mixture mappings in a BERT-based model [131].

The main challenge that must be overcome to train some non-English and multilingual NLP foundation models is the scarcity of labeled language data. For English, this is not an issue due to the overwhelming amount of data publicly available on the Web. However, for many other languages, such as Spanish and Vietnamese, labelled data can be very scarce [26, 158]. One of the great aspects of foundation models and transfer-learning, though, is that once a portion of labelled data has been collected, one can design a system to automatically label data to extend the original data set [190, 209] (Table 1).

6 Societal Impact of NLP Foundation Models

Foundational models for NLP can aid in constructing specialized models for downstream tasks. These models make training AI specialising in human language quick and efficient [20]. AI, in general, is hypothesised to have a transformative effect on society [38, 71], specifically AI geared toward NLP. NLP foundation models can be utilized to implement chatbots, interactive story-writing AI, question-answering systems, and even improve search engine capabilities.

However, certain biases can be perpetuated downstream if the foundation model has an intrinsic bias, which can lead to harmful results [20]. Here, we define intrinsic bias as any aspect of a foundation model that can contribute to inequality in downstream applications. Bias in any language model arises from the data-driven training process [80]. This can be especially harmful to marginalised groups (i.e., racial and religious minorities in America) An example of this harmful nature is found in the paper by Sweeney [196] where Google's search engine was found to deliver discriminatory ads when searching for racially associated names. Similarly, in the paper by Abid et al., the religious bias of GPT-3 was explored using prompt completion and story generation. Some of the results of this study were that GPT-3 associated the word "Muslim" with violence 66% of the time, and the term "Jewish" was associated with money in 5% of test cases [4].

Table 1 A selection of Non-English NLP models

Model Name	Article	Mono- or Multilingual	Language(s)	Achievable NLP Tasks
PhoBERT	[158]	Monolingual	Vietnamese	POS-Tagging, NER, Dependency Parsing, NLI
IndoBERT	[100]	Monolingual	Indonesian	POS-tagging, NER, SA, Summarization
Spanish-BERT	[26]	Monolingual	Spanish	NER, Paraphrasing, NLI, POS-tagging, QA
M6	[120]	Monolingual	Chinese	Text-to-Image Generation, QA, Image Captioning, Text Generation
Est-BERT	[198]	Monolingual	Estonian	SA, NER, POS-tagging
Finnish BERT	[207]	Monolingual	Finnish	POS-tagging, NER, Dependency Parsing
LitLat BERT	[202]	Multilingual	Lithuanian and Latvian	NER, Dependency Parsing, POS-tagging
RuBERT	[104]	Monolingual	Russian	QA, SA
AlBERTo	[164]	Monolingual	Italian	QA, SA
CamemBERT	[136]	Monolingual	French	POS-tagging, Dependency Parsing, NLI, NER
RobBERT	[43]	Monolingual	Dutch	SA, POS-tagging, NER
AraBERT	[58]	Monolingual	Arabic	SA
M-BERT	[45]	Multilingual	Over 100 languages	QA, SA, POS-tagging, NER, NLI, and more
mT5	[220]	Multilingual	101 languages	NER, POS-tagging, Text Classification

6.1 NLP Demographic, geographic and temporal bias

Lately, there has been a growing concern about demographic-based biases within natural language processing (NLP) applications. Many recent studies have focused on elucidating bias and presenting a broader perspective on its implications [82]. They identified five primary origins of bias that can manifest in NLP systems: (1) data quality, (2) the process of annotation, (3) the representation of input, (4) the models employed, and (5) the structure of research design. [162] introduces a novel method that employs sentiment analysis to explore how the emotions of tourists are affected by temporal and spatiotemporal factors when they visit various tourist destinations within a city. The proposed approach is structured into four main phases: data collection and preprocessing from social media platforms, identification of visitor sources, sentiment analysis during visits, and an examination of temporal and spatiotemporal dynamics. Additionally, it delves into spatiotemporal elements including the sentiment evolution across different locations, enjoyment assessment, and sentiment trends across multiple sites. The framework is applied to the context of Chicago, using a substantial dataset comprising over eight million tweets. The outcomes reveal that tourists' emotional states are influenced by seasonal weather patterns and special events like concerts. Furthermore, the study uncovers that visitors tend to report heightened levels of enjoyment when visiting specific locations such as observatories, in contrast to zoos. Interestingly, both local and international tourists tend to express predominantly negative sentiments when engaging in multiple sightseeing activities within a single day, while out-of-state visitors display a divergent sentiment pattern in such scenarios.

[145] Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. Later, [161] suggested that fairness evaluation in NLP frequently involves employing fairness metrics to gauge a model's behavior disparities across various demographic groups. They provided deeper insights into the fairness metrics utilized within the NLP domain. Initially, they categorized a diverse array of established metrics into three overarching fairness metrics, uncovering the interrelationships among them. Subsequently, [70] conducted an extensive empirical analysis of these existing metrics and showcased that the variations in bias assessment can be methodically made through disparities in parameter selections for our generalized metrics.

6.2 Social Impact

Given the relatively recent increases in AI, it should be expected that the general public would be sceptical of AI implementation in society, such as in public transportation and medical AI, workplace automation, online chatbots,

Table 2 CO_2 emissions for common consumption compared to emissions of NLP model training [191]

Consumption	CO_2 Emission (lbs)
Air travel, 1 passenger, NY to SF	1984
Human life, avg., 1 year	11,023
American life, avg. 1 year	36,156
Car, avg. including fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
NLP pipeline with tuning and experimentation	78,468
Transformer (big)	192
Transformer with neural architecture search	626,155

and more. In the 2019 study by Zhang and Defoe, it was found that only 41% of the Americans surveyed were in favor of the development of AI [227]. Regardless, it is commonly agreed upon (roughly 82% of those surveyed) that robots and AI must be carefully managed and monitored to ensure privacy and civil liberties are upheld.

Many downstream tasks implemented using NLP foundation models, like machine translation and text generation, are specifically aimed at helping people toward some goal. Therefore, it is essential to merge technical and social methodologies when creating foundation models to mitigate bias and rectify harmful circumstances from intrinsically biased foundation models [3, 90, 226]. Combining the expertise of sociologists and the technical prowess of AI engineers can identify sources of bias in our culture and help foundation models to avoid perpetuating these biases. [125].

Many NLP models are trained using neural networks (i.e., deep learning). However, these models require enormous computational power for training and fine-tuning model parameters, see Table 1 [74, 191]. GPT-3 alone has 175 billion parameters that can be trained for various NLP tasks [134]. Though profitable in the short-term, the storage of machine learning results and copious computation requirements is unsustainable from an energy-use standpoint [74]. Being those foundation models have the prospect to grow and be applied in many different areas, there is a concern for higher model quality leading to energy and environmental implications [215].

Due to the sheer computing necessary to train foundation models, model parameter weights for specific downstream tasks can download online. This makes foundation models more accessible to interested parties but also begs the question of the integrity of these parameters. Specifically, how can one ensure that the parameters have not been maliciously tampered with to cause a model to tend toward incorrect or harmful results?

While foundation models, in theory, are not inherently biased, if the training data that the model is learning from demonstrating biases, stereotypes, and prejudices, the resulting foundation model will carry those negative attributes [32]. For example, foundation models as a tool to apply NLP could be used when assessing essays among students. The model may be tuned to reward essays that use a vast vocabulary, have correct spelling and grammar, and highlight specific words relevant to the topic. While this may seem like a fair rubric, if a machine is assessing based on these qualities, a student whose first language is not English or a student with an intellectual disability may not be appropriately assessed, provided their abilities to write in English [32]. Similarly, "Criminal" host Phoebe Judge discussed the concerns about using computers for lie detection with guest and forensic psychologist Andy Morgan. Morgan proposed that a person telling the truth would use unique and colourful words when retelling an account, while a person lying will often repeat themself and use simple words. Based on these criteria, a foundation model could certainly be used as a lie detection service. However, people with an intellectual disability and people with English as a secondary language could be consistently detected as lying because their English vocabulary may not be as vast as a native English speaker without an intellectual disability. [91].

A possible area for foundation models to apply NLP is decision-making within the law, known as algorithmic governance [49]. One practicality of foundation models, in general, is the need for clean data. When considering algorithmic governance, data collection, storage, and usage regulations would be necessary to provide an ethical basis for automated decision-making [49]. While algorithmic governance aims to remove human biases, many automated systems still "favour social groups of privilege" [49].

Using fill-in-the-blank analysis, BERT was found to associate more negative sentiment with phrases referencing people with disabilities [83]. Since BERT is a foundation model for many NLP applications, having a bias at the base of BERT creates a trickle-down impact on applications [83]. In contrast, foundation models have been created with an implicit bias against people with disabilities, foundation models to apply NLP can also be used to support people with disabilities. Google's Android Speech Recognition (ASR) library can perform speech to text in over 120 languages, an essential resource for people with visual or tactile disabilities [56].

6.3 Ethical Impact

One of the critical features of foundation models is the homogenisation of machine learning methodologies [20]. However, this homogenisation can amplify and standardise biases and lead to the exclusion of specific individuals or groups [36]. Moreover, foundation models for NLP can be used for various things, such as improving online search functionality and text generation for apps and video games [213].

Training data can impact the performance of a machine learning model, and this holds true for foundation models as well. From data scraped from the World Wide Web, Caliskan et al. show that text corpora contain accurate impressions of historical human biases toward morally-neutral subjects like bugs and more problematic biases like gender and race discrimination [25]. Similarly, inherently biased labels can lead to disparate outcomes for marginalised groups compared to majority groups [15, 17]. These biases can be incredibly harmful through the homogenisation of foundation models since discriminatory bias will be transferred to all subsequent downstream models constructed using an inherently biased foundation model.

Recently, Google Search began to utilise BERT-trained models to increase the quality of search results returned by a user query. These BERT-trained models allow the Google Search to determine the intent (i.e., context) of a query and then return results that most closely match this intent [157]. This updated Search functionality allows one to easily query Google in a somewhat conversational manner instead of using keywords strings to get relevant results.

7 Discussion

Foundation models are ever-evolving. While many challenges to optimisation and efficiency have been remedied or at least alleviated to some degree, some concerns and room for improvement still exist. We also

7.1 Vulnerabilities and Limitations

Similar to computer networks, foundation models are susceptible to cyber-attacks. Poisoning attacks occur when an attacker manipulates data the model is trained on to influence the results of a predictive model [64, 86]. Model poisoning refers to the intentional corruption of model parameter weights [105, 113]. For NLP foundation models, poisoning can lower model accuracy and even force a model to spread false information or bias and hate speech when prompted by the attacker [51, 137]. Model poisoning is effective for two main reasons: first, poisoning can be challenging to detect because the performance of the model only degrades for samples that the attacker has targeted and second, a tiny portion of the training dataset needs to be manipulated to cause the model to misclassify the attacker's targeted samples [27]. Currently, there is much in the way of preventative research due to the severity of the consequences of these kinds of attacks.

Most foundation models for NLP tasks are trained on textual data scraped from the Internet. Since anyone has the ability to post things online regardless of the truth value of the post, this can make it relatively easy for foundation models to pick up harmful information and spread misinformation or even allow an attacker to infiltrate a foundation model. Backdoor attacks, where an attacker embeds a predefined trigger word or phrase to cause the model to influence the model's prediction capabilities, are an example of model poisoning. To embed a backdoor in a model, an attacker maliciously tweaks the training data to make the model misclassify samples containing the predefined trigger [64, 116]. These attacks can damage any system but can be especially detrimental for pre-trained models. If an attacker embeds a backdoor into the pre-trained foundation model, all downstream models will inherit this backdoor as well [30, 209]. Generally, backdoor attacks do not affect the model's performance when the trigger is not present in a sample, which can make backdoor attacks very difficult to detect [39].

8 Conclusion, Moving Forward and Future Directions

This survey presents foundation models and NLP's theory, development, and applications. First, we describe the NLP applications. Then, we analyse the NLP's theory and the development of the foundation models. Then, we provide the applications of NLP Foundation Models and the Societal Impact of NLP Foundation Models. Finally, we provide the current limitations in this research area and some future directions. We believe this survey will give diverse academics a thorough understanding of this field's most important aspects.

Moving forward, foundation models through emergence will have the ability to continue to be a potent tool. By improving and maintaining foundation models, natural language processing tasks will be more effective and efficient. This progress will come with positive outcomes for society, such as quicker and more accurate machine translation tools for those accessing an unfamiliar language and better text-reading machines for people with disabilities. Subsequently, we proceed to provide a concise overview of the deliberations held throughout this survey, while also presenting the obstacles and future directions for

Language Learning Machines (LLMs) in the following points:

1. The concepts of theory and principle. In order to comprehend the fundamental operational principles of Language Models (LLMs), a significant enigma lies in the manner by which information is disseminated, structured, and harnessed within the extensive and intricate neural network. The identification of fundamental concepts or aspects that form the basis of the capabilities of LLMs holds significant importance. The impact of scaling appears to be significant in enhancing the capability of large language models (LLMs) [173]. Previous research has demonstrated that certain unforeseen capabilities, characterised by sudden improvements in performance, can emerge when the scale of language models reaches a critical threshold, such as 10 billion parameters. These capabilities often involve in-context learning, following instructions, and engaging in step-by-step reasoning. The acquisition of these emergent abilities by LLMs remains intriguing and enigmatic, as the precise timing and mechanisms via which they are acquired have yet to be fully elucidated. Recent research has been dedicated to conducting comprehensive experiments in order to investigate the impact of emergent skills and the elements that contribute to their development [345, 680, 681]. Alternatively, other investigations have sought to elucidate specific talents by applying established theoretical frameworks. A scholarly technical post also specifically addresses this subject, focusing on the GPT-series models. Nevertheless, there is a notable absence of formal theories and principles that can be used to comprehend, define, and elucidate the capabilities or behaviours of LLMs. Given the close resemblance between emergent abilities and phase transitions observed in natural phenomena [31, 58], it may be beneficial to explore cross-disciplinary theories or principles in order to elucidate and comprehend the behaviours of LLMs. For instance, investigating whether LLMs may be regarded as complex systems could provide valuable insights. The exploration of these fundamental questions holds significant value [201].

2. The research community plays a crucial role in the advancement of next-generation Language Learning Models (LLMs). The proposed framework for the system's structure. The Transformer design, which incorporates stacked multi-head self-attention layers, has emerged as the preferred choice for constructing Language Models (LLMs) because of its scalability and effectiveness. Several solutions have been suggested to enhance the efficiency of this design, including neural network setup and scalable parallel training. In order to augment the model's capability, such as its ability to engage in multi-turn conversations, current Language Learning Models (LLMs) commonly employ a strategy of maintaining a lengthy context window. For instance, GPT is characterised by an exceptionally extensive context duration of 32,768 tokens. Hence, a pertinent factor to address is the reduction of time complexity, namely the quadratic costs associated with the conventional self-attention technique. Conducting research on the impact of enhanced Transformer variations in constructing Language Models (LLMs) is of utmost significance. For instance, the utilisation of sparse attention has been observed in the development of GPT-3. In addition, the phenomenon of catastrophic forgetting has posed a persistent obstacle for neural networks, hence exerting a detrimental influence on language model models (LLMs). When doing fine-tuning on Language Models (LLMs) using fresh data, it is probable that the previously acquired information may be negatively impacted. For instance, fine-tuning an LLM for certain tasks might have adverse effects on the overall proficiency of LLMs. A comparable scenario arises when LLMs exhibit alignment with human values, which is sometimes referred to as alignment tax. Therefore, it is imperative to contemplate the expansion of current designs by incorporating more adaptable mechanisms or modules that can efficiently facilitate data updates and task specialisation [201].

3. The process of training a model. In practical applications, the pre-training of proficient Language Learning Models (LLMs) poses significant challenges. These challenges arise from the extensive computational resources required, as well as the susceptibility of the models to the quality of the data and the intricacies of the training techniques [69, 84]. Therefore, it is crucial to focus on the development of systematic and cost-effective pre-training methods for enhancing the performance, efficiency, and stability of language model models (LLMs). There is a need for further development of model checking and performance diagnosis techniques, such as the exploration of predictable scaling in GPT-4. These advancements are necessary to effectively identify and address aberrant issues that may arise during the training process at an early stage. In addition, it also necessitates the implementation of more adaptable hardware support mechanisms or resource scheduling techniques in order to effectively manage and optimise the utilisation of resources inside a computing cluster. Given the high expense associated with the initial training of a Language Model (LLM) from the ground up, it is imperative to develop appropriate approaches for consistently pre-training or fine-tuning the LLM using publically accessible model checkpoints such as LLaMA and Flan-T5. In order to achieve this objective, several technical challenges need to be addressed, including the

concerns of catastrophic forgetting and task specialisation. Nevertheless, as of now, there remains a dearth of open-source model checkpoints for Language Model (LM) Learning Models (LLMs) that include comprehensive pre-processing and training logs. These logs would encompass essential information such as the scripts used to prepare the pre-training data, enabling the reproducibility of the experiments. It is believed that the inclusion of more technical details in open-source models pertaining to the investigation of LLMs will yield significant benefits. Moreover, it is crucial to devise additional enhancement tuning procedures that effectively evoke the capabilities of the model. The utilisation of models. Due to the high expense associated with fine-tuning in practical scenarios, the utilisation of prompting has emerged as the prevailing method for using Language Models (LLMs). Incorporating task descriptions and demonstration examples into prompts facilitates in-context learning, a specialised form of education.

4. The utilisation of contrastive learning, a specific sort of prompting, empowers LLMs (Language Model Models) to exhibit strong performance on novel tasks, occasionally surpassing the performance of models that have been fine-tuned using complete data. Moreover, in order to augment the capacity for intricate thinking, scholars have put forth improved prompting techniques, such as the chain-of-thought (CoT) method, which incorporates intermediate reasoning processes within prompts. However, the current methods used for prompting still exhibit a number of flaws, which are outlined as follows. Firstly, the process entails significant human exertion in the formulation of prompts. The automatic generation of effective prompts for solving diverse tasks would be highly advantageous. Furthermore, certain intricate activities, such as formal proof and numerical computation, necessitate specialised knowledge or logical principles that may not be effectively conveyed in natural language or exemplification. Therefore, the development of more informative and adaptable task structuring approaches for prompts is of utmost importance. Furthermore, it is important to note that current prompting systems predominantly prioritise the evaluation of single-turn performance. Developing interactive prompting mechanisms, such as those facilitated by natural language discussions, has proven to be highly advantageous in addressing intricate tasks. This has been exemplified by the notable efficacy of ChatGPT [31].

5. The topics of safety and alignment should be discussed. Although LLMs provide certain capabilities, they present comparable safety concerns to those of tiny language models. As an illustration, it has been observed that LLMs have a proclivity for producing hallucinations, which refer to texts that appear to be credible but may contain factual inaccuracies. LLMs may be triggered by deliberate directives to generate detrimental, prejudiced, or noxious texts for malevolent systems, hence engendering the potential hazards of misapplication. For a comprehensive examination of the safety concerns associated with Language Model Models (LLMs), such as privacy, excessive dependence, dissemination of false information, and manipulation of influence, interested individuals are encouraged to consult the technical papers of GPT-3. Reinforcement learning from human feedback (RLHF) has emerged as a prominent strategy for addressing these challenges. This approach involves integrating human input into the training process to enhance the alignment of language models (LLMs). In order to enhance the safety of the model, it is imperative to incorporate cues that are pertinent to safety during Reinforcement Learning from Human Feedback (RLHF), as demonstrated by the findings of GPT-4. Nevertheless, Reinforcement Learning from Human Feedback (RLHF) is strongly dependent on the availability of high-quality human feedback data provided by experienced labellers. This reliance poses significant challenges for its practical implementation. Hence, there is a need to enhance the RLHF framework in order to minimise the workload of human labellers and explore a more effective annotation method that ensures high-quality data. For instance, the utilisation of Language Models (LLMs) can be considered to support the labelling process. It is also significant to build an appropriate learning mechanism for Language Models (LLMs) to acquire human feedback through chat interactions and effectively utilise it for self-enhancement. Furthermore, it is crucial to take into account privacy considerations while refining Language Model Models (LLMs) with domain-specific data. In situations where privacy is prohibited, the utilisation of federated learning libraries can prove to be advantageous [102].

Author Contributions All authors contributed equally to the paper.

Funding No Funding

Data Availability All datasets are open-source, and the sources are cited.

Declarations

Conflict of interest No competing interest.

Ethical approval and consent to participate: No ethical issue is involved.

References

1. Abas, A.R., El-Henawy, I., Mohamed, H., Abdellatif, A.: Deep learning model for fine-grained aspect-based opinion mining. IEEE Access **8**, 128845–128855 (2020)
2. Abdullah, M., Madain, A., Jararweh, Y.: Chatgpt: Fundamentals, applications and social impacts. In: 2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS), pp. 1–8. IEEE, (2022)
3. Abebe, R., Barocas, S., Kleinberg, J., Levy, K., Raghavan, M., Robinson, D.G.: Roles for computing in social change. In: Proceedings of the 2020 COnference on Fairness, Accountability, and Transparency, (2020)
4. Abid, A., Farooqi, M., Zou, J.: Persistent anti-muslim bias in large language models. arXiv preprint arXiv:2101.05783 (2021)
5. Akhila, N. et al.: Comparative study of bert models and roberta in transformer based question answering. In: 2023 3rd International Conference on Intelligent Technologies (CONIT), pp. 1–5. IEEE, (2023)
6. Al-Hawawreh, M., Aljuhani, A., Jararweh, Y.: Chatgpt for cybersecurity: practical applications, challenges, and future directions. Clust. Comput. pp. 1–16 (2023)
7. Alan Ramponi, B.P.: Neural unsupervised domain adaptation in nlp–a survey, (2020)
8. Alkhurayyif, Y., Rahaman Wahab Sait, A.: Developing an open domain arabic question answering system using a deep learning technique. In: IEEE Access (2023)
9. An, T., Song, J., Liu, W.: Incorporating pre-trained model into neural machine translation. In: 2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD), pp. 212–216 (2021)
10. Antoun, W., Baly, F., Hajj, H.: Ara bert: Transformer-based model for arabic language understanding. arXiv preprint arXiv:2003.00104 (2021)
11. Araujo, A.F., Gôlo, M.P.S., Marcacini, R.M.: Opinion mining for app reviews: an analysis of textual representation and predictive models. Autom. Softw. Eng. **29**, 1–30 (2022)
12. Arumae, K., Liu, F.: Guiding extractive summarization with question-answering rewards. CoRR, abs/1904.02321 (2019)
13. Baldini, I., Wei, D., Ramamurthy, K.N., Yurochkin, M., Singh, M.: Your fairness may vary: Pretrained language model fairness in toxic text classification. arXiv preprint arXiv:2108.01250 (2021)
14. Bani-Almarjeh, M., Kurdy, M.-B.: Arabic abstractive text summarization using rnn-based and transformer-based architectures. Inf. Process. Manag. **60**(2), 103227 (2023)
15. Bartlett, Robert: Morse, Adair, Stanton, Richard. Wallace. Discrimination in the FinTech Era, National Bureau of Economic Research, Nancy (2019)
16. Bataa, E., Wu, J.: An investigation of transfer learning-based sentiment analysis in Japanese (2019)
17. Benjamin, Ruha: Assessing risk, automating racism. Science **366**, 421–422 (2019)
18. Bhattacharjee, S., Haque, R., de Buy Wenniger, G.M., Way, A.: Investigating query expansion and coreference resolution in question answering on bert. In Elisabeth Métais, Farid Meziane, Helmut Horacek, and Philipp Cimiano, editors, Natural Language Processing and Information Systems, pp. 47–59, Cham (2020). Springer International Publishing
19. Bi, B., Li, C., Wu, C., Yan, M., Wang, W., Huang, S., Huang, F. Si, Luo: P.: Pre-training an autoencoding &autoregressive language model for context-conditioned generation. arXiv preprint arXiv:2004.07159 (2020)
20. Bommasani, R., Hudson, D.A, Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)
21. Borji, A.: A categorical archive of chatgpt failures. arXiv preprint arXiv:2302.03494 (2023)
22. Bowman, S.R., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. arXiv preprint arXiv:1508.05326 (2015)
23. Buck, C., Bulian, J., Ciaramita, M., Gajewski, W., Gesmundo, A., Houlsby, N., Wang, W.: Ask the right questions: active question reformulation with reinforcement learning. arXiv preprint arXiv: 1705.07830, (2018)
24. Büyüköz, B., Hürriyetoglu, Ö.: Arzucan: Analyzing elmo and distilbert on socio-political news classification. Proceedings of AESPEN **2020**, 9–18 (2020)
25. Caliskan, Aylin, Bryson, Joanna J., Narayanan, Arvind: Semantics derived automatically from language corpora contain human-like biases. Science **356**, 183–186 (2017)
26. Canete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., Pérez, J.: Spanish pre-trained bert model and evaluation data. PML4DC at ICLR 2020 (2020)
27. Carlini, N., Terzis, A.: Poisoning and backdooring contrastive learning. arXiv preprint arXiv:2106.09667 (2022)
28. Chang, W.-C., Yu, H.-F., Zhong, K., Yang, Y., Dhillon, I.S.: Taming pretrained transformers for extreme multi-label text classification. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 3163–3171, New York, NY, USA, (2020). Association for Computing Machinery
29. Chen, G., Ma, S., Chen, Y., Dong, L., Zhang, D., Pan, J., Wang, W.W.: Zero-shot cross-lingual transfer of neural machine translation with multilingual pretrained encoders, Furu (2021)
30. Chen, K., Meng, Y., Sun, X., Guo, S., Zhang, T., Li, J., Fan, C: Badpre: Task-agnostic backdoor attacks to pre-trained NLP foundation models. arXiv (2021)
31. Chen, Q., Sun, H., Liu, H., Jiang, Y., Ran, T., Jin, X., Xiao, X., Lin, Z., Niu, Z., Chen, H.: A comprehensive benchmark study on biomedical text generation and mining with chatgpt. bioRxiv, pp. 2023–04 (2023)
32. Cheuk, Tina: Can AI be racist? Color-evasiveness in the application of machine learning to science assessments. Sci. Educ. **105**(5), 825–836 (2021)
33. Chronopoulou, A., Stojanovski, D., Fraser, A.: Improving the lexical ability of pretrained language models for unsupervised neural machine translation. arXiv preprint arXiv:2103.10531 (2021)
34. Clark, K., Luong, M.T., Le, Q.V., Manning, C.D.: Electra: pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555 (2020)
35. Clinchant, Stéphane.: Jung, Kweon Woo. Nikoulina. On the use of bert for neural machine translation, Vassilina (2019)
36. Creel, K., Hellman, D.: The algorithmic leviathan: arbitrariness, fairness, and opportunity in algorithmic decision making systems. In: Proceeding of the 2021 ACM Conference on Fairness, Accountability, and Transparency (2021)
37. Dabre, R., Chu, C., Kunchukuttan, A.: A survey of multilingual neural machine translation. ACM Comput. Surv. **53**(5), 1–38 (2020)
38. Dafoe, A.: AI governance: a research agenda. Governance of AI program, the Future of Humanity Institute, the University of Oxford, Oxford (2018)
39. Dai, J., Chen, C., Li, Y.: A backdoor attack against LSTM-based text classification systems. IEEE Access **7**, 138872–138878 (2019)
40. Dang, E., Hu, Z., Li, T.: Enhancing collaborative filtering recommender with prompt-based sentiment analysis. arXiv preprint arXiv:2207.12883, (2022)

41. de Vries, W., Nissim, M.: As good as new. how to successfully recycle english GPT-2 to make models for other languages. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics (2021)
42. Majd Saad Al Deen, M., Pielka, M., Hees, J., Soulef Abdou, B., Sifa, R.:Improving natural language inference in arabic using transformer models and linguistically informed pre-training. arXiv preprint arXiv:2307.14666 (2023)
43. Delobelle, P., Winters, T., Berendt, B.: Robbert: a dutch roberta-based language model. arXiv preprint arXiv:2001.0628 (2020)
44. Deng, X., Bashlovkina, V., Han, F., Baumgartner, S., Bendersky, M.: What do llms know about financial markets? a case study on reddit market sentiment analysis. In: Companion Proceedings of the ACM Web Conference 2023, pp. 107–110 (2023)
45. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018)
46. Ding, Z., Qi, Y., Lin, D.: Albert-based sentiment analysis of movie review. In: 2021 4th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), pp. 1243–1246 (2021)
47. Dinh, T.A., Niehues, J.: Perturbation-based qe: An explainable, unsupervised word-level quality estimation method for blackbox machine translation. arXiv preprint arXiv:2305.07457 (2023)
48. Djandji, M., Baly, F., Antoun, W., Hajj, H.: Multi-task learning using ara bert for offensive language detection. Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, pp. 97–101, (2020)
49. DoCarmo, T., Rea, S., Conaway, E., Emery, J., Raval, N.: The law in computation: What machine learning, artificial intelligence, and big data mean for law and society scholarship. Law & Policy **43**(2), 170–199 (2021)
50. Dong, L., Mallinson, J., Reddy, S., Lapata, M.: Learning to paraphrase for question answering. arXiv:1708.06022 (2017)
51. Du, Y., Bosselut, A., Manning, C.D.: Synthetic disinformation attacks on automated fact verification systems. arXiv preprint arXiv:2202.09381 (2022)
52. Duarte, J.M., Berton, L.: A review of semi-supervised learning for text classification. Artificial Intelligence Review, pp. 1–69 (2023)
53. Duong, D., Solomon, B.D: Analysis of large-language model versus human performance for genetics questions. medRxiv, pp. 2023–01 (2023)
54. Edunov, S., Baevski, A., Auli, M.: Pre-trained language model representations for language generation. arXiv preprint arXiv:1903.09722 (2019)
55. Eisenstein, J., Andor, D., Bohnet, B., Collins, M., Mimno, D.: Honest students from untrusted teachers: Learning an interpretable question-answering pipeline from a pretrained language model. arXiv preprint arXiv:2210.02498, (2022)
56. Emil, Z., Robbertz, A., Valente, R., Winsor, C:. Towards a more inclusive world: Enhanced augmentative and alternative communication for people with disabilities using ai and nlp. Worcester Polytechnic Institute, (2020)
57. Erciyes, Necdet Eren, Görür, Abdül Kadir: Deep learning methods with pre-trained word embeddings and pre-trained transformers for extreme multi-label text classification. In: 2021 6th International Conference on Computer Science and Engineering (UBMK), pp. 50–55, (2021)
58. Faraj, D., Abdullah, M.: Sarcasm det at sarcasm detection task 2021 in arabic using ara bert pretrained model. In: Proceedings of the Sixth Arabic Natural Language Processing Workshop, pp. 345–350 (2021)
59. Fernandes, P., Deutsch, D., Finkelstein, M., Riley, P., Martins, A.F., Neubig, G., Garg, A., Clark, J.H., Freitag, M., Firat, O.: The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. arXiv preprint arXiv:2308.07286, (2023)
60. Floridi, L., Chiriatti, M.: Gpt-3: Its nature, scope, limits, and consequences. Minds and Machines (2020)
61. Floridi, Luciano, Chiriatti, M.: Gpt-3: Its nature, scope, limits, and consequences. Minds Mach. **30**, 681–694 (2020)
62. Fuadi, M., Wibawa, A.D., Sumpeno, S.: idt5: Indonesian version of multilingual t5 transformer. arXiv preprint arXiv:2302.00856 (2023)
63. Fukumoto, D., Kashiwa, Y., Hirao, T., Fujiwara, K., Iida, H.: An empirical investigation on the performance of domain adaptation for t5 code completion. In: 2023 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER), pp. 693–697. IEEE (2023)
64. Gao, Y., Gia Doan, B., Zhang, Z., Ma, S, Zhang, J., Fu, A., Nepal, S., Kim, H.: Backdoor attacks and countermeasures on deep learning: a comprehensive review. arXiv preprint arXiv:2007.10760 (2020)
65. Geetha, M.P., Karthika Renuka, D.: Improving the performance of aspect based sentiment analysis using fine-tuned bert base uncased model. Int. J. Intell, Netw (2021)
66. Ghourabi, A.: A bert-based system for multi-topic labeling of arabic content. In: 2021 12th International Conference on Information and Communication Systems (ICICS), pp. 486–489 (2021)
67. Giorgi, John M., Wang, Xindi, Sahar, Nicola, Young Shin, Won, Bader, Gary D., Wang, Bo: End-to-end named entity recognition and relation extraction using pre-trained language models. arXiv preprint arXiv:1912.13415, (2019)
68. Giovannotti, P.: Evaluating machine translation quality with conformal predictive distributions. arXiv preprint arXiv:2306.01549 (2023)
69. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems—Volume 2, NIPS'14, pp. 2672–2680, Cambridge, MA, USA, (2014). MIT Press
70. Gore, Ross Joseph, Diallo, Saikou, Padilla, Jose: You are what you tweet: connecting the geographic variation in america's obesity rate to twitter content. PloS ONE **10**(9), e0133505 (2015)
71. Gruetzemacher, Ross, Whittlestone, J.: The transformative potential of artificial intelligence. Futures **135**, 102884 (2022)
72. Guo, B., Wang, H., Ding, Yasan, Wu, Wei, Hao, Shaoyang, Sun, Yueqi, Yu, Zhiwen: Conditional text generation for harmonious human-machine interaction. ACM Trans. Intell. Syst. Technol., 12(2), (apr 2021)
73. Guo, Junliang, Zhang, Zhirui, Xu, Linli, Chen, Boxing, Chen, Enhong: Adaptive adapters: An efficient way to incorporate bert into neural machine translation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, pp. 1740–1751, (2021)
74. Gupta, A., Lanteigne, C., Kingsley, S.: SECure: a social and environmental certificate for AI systems. arXiv preprint arXiv:2006.06217 (2020)
75. Guven, Z.A.: The effect of bert, electra and albert language models on sentiment analysis for turkish product reviews. In: 2021 6th International Conference on Computer Science and Engineering (UBMK), pp. 629–632 (2021)
76. Han, J.M., Babuschkin, I., Edwards, H., Neelakantan, A., Xu, T., Polu, S., Ray, A., Shyam, P., Ramesh, A., Radford, A.:

Sutskever. Unsupervised neural machine translation with generative language models only, Ilya (2021)
77. Han, Xu.: Zhang, Zhengyan, Ding, Ning, Gu, Yuxian, Liu, Xiao, Huo, Yuqi, Qiu, Jiezhong, Yao, Yuan, Zhang, Ao, Zhang, Liang, Han, Wentao, Huang, Minlie, Jin, Qin, Lan, Yanyan, Liu, Yang, Zhiyuan Liu, Zhiwu Lu, Qiu, Xipeng, Song, Ruihua, Tang, Jie, Wen, Ji-Rong, Yuan, Jinhui, Xin Zhao, Win, Zhu, Jun: Pre-trained model: Past, present, and future. Elsevier, Amsterdam (2021)
78. He, Y., Zhu, Z., Zhang, Y., Chen, Q., Caverlee, J.: Infusing disease knowledge into BERT for health question answering, medical inference and disease name recognition (2020). arXiv preprint arXiv:2010.03746
79. Hegde, C., Patil, S.: Unsupervised paraphrase generation using pre-trained language models. arXiv preprint arXiv:2006.05477 (2020)
80. Henderson, Peter, Sinha, Koustuv, Angelard-Gontier, Nicolas, Rosemary Ke, Nan, Fried, Genevieve, Lowe, Ryan, Pineau, Joelle: Ethical challenges in data-driven dialogue systems. arXiv preprint arXiv:1711.09050, (2017)
81. Hossain, Md Rajib, Hoque, Mohammed Moshiul, Siddique, Nazmul: Leveraging the meta-embedding for text classification in a resource-constrained language. Engineering Applications of Artificial Intelligence, 124:106586, (2023)
82. Hovy, D., Prabhumoye, S.: Five sources of bias in natural language processing. Lang. Linguistics Compass 15, 8 (2021), e12432 (2021)
83. Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., Denuyl, S.: Social biases in nlp models as barriers for persons with disabilities. Association for Computational Linguistics (2020)
84. Jacob, D., Chang, M.W., Kenton, L., Kristina, T.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv.org (2019)
85. Jacobs, P.S.: Joining statistics with nlp for text categorization . In: Third Conference on Applied Natural Language Processing, (1992)
86. Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., Li, B.: Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. arXiv preprint arXiv:1804.00308, (2021)
87. Jain, Praphula Kumar, Quamer, Waris, Pamula, Rajendra: Consumer sentiment analysis with aspect fusion and gan-bert aided adversarial learning. Expert Syst. **40**(4), e13247 (2023)
88. Jin, W., Mao, H., Li, Z., Jiang, H., Luo, C., Wen, H., Han, H., Lu, H., Wang, Z., Li, R., et al.: Amazon-m2: A multilingual multi-locale shopping session dataset for recommendation and text generation. arXiv preprint arXiv:2307.09688, (2023)
89. Jing, W., Bailong, Y.: News text classification and recommendation technology based on wide amp; deep-bert model. In: 2021 IEEE International Conference on Information Communication and Software Engineering (ICICSE), pp. 209–216 (2021)
90. Joyce, K., Smith-Doerr, L., Alegria, S., Bell, S., Cruz, T., Hoffman, S.G., Umoja Noble, S., Shestakofsky, B.: Towards a sociology of artificial intelligence: a call for research on inequalities and structural change. Socius (2021)
91. Phoebe Judge (Host). Pants on fire, February 14, (2014)
92. Kadaoui, Karima, Magdy, Samar M., Waheed, Abdul, Khondaker, Md Tawkat Islam, El-Shangiti, Ahmed Oumar, Nagoudi, El Moatez Billah, Abdul-Mageed, Muhammad: Tarjamat: Evaluation of bard and chatgpt on machine translation of ten arabic varieties. arXiv preprint arXiv:2308.03051, (2023)
93. Karimi, A., Rossi, L.: Prati. Improving bert performance for aspect-based sentiment analysis, Andrea (2020)
94. Karimi, A., Rossi, L., Prati, A.: Adversarial training for aspect-based sentiment analysis with bert. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 8797–8803, (2021)
95. Khan, Aisha Urooj, Mazaheri, Amir, da Vitoria Lobo, Niels, Shah, Mubarak: Mmft-bert: Multimodal fusion transformer with bert encodings for visual question answering, (2020)
96. Khan, Wahab, Daud, Ali, Nasir, Jamal A., Amjad, Tehmina: A survey on the state-of-the-art machine learning models in the context of nlp. Kuwait journal of Science, 43(4), (2016)
97. Kheiri, Kiana, Karimi, Hamid: Sentimentgpt: Exploiting gpt for advanced sentiment analysis and its departure from current machine learning. arXiv preprint arXiv:2307.10234, (2023)
98. Kiros, Jamie, Chan, William: Inferlite: Simple universal sentence representations from natural language inference data. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, (2018)
99. Kolides, A., Nawaz, A., Rathor, A., Beeman, D., Hashmi, M., Fatima, S., Berdik, D., Al-Ayyoub, J.Y.: Artificial intelligence foundation and pre-trained models: fundamentals, applications, opportunities, and social impacts. Simul. Model. Pract. Theory **126**, 102754 (2023)
100. Koto, F., Rahimi, A., Lau, J.H., Baldwin, T.: Indolem and indobert: A benchmark dataset and pre-trained language model for indoesian nlp. arXiv preprint arXiv:2011.00677 (2020)
101. Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., Brown, D.: Text classification algorithms: A survey. Information **10**(4), 150 (2019)
102. Kuang W, Qian B, Li Z, Chen D, Gao D, Pan X, Xie Y, Li Y, Ding B, Zhou J: A comprehensive package for fine-tuning large language models in federated learning. arXiv preprint arXiv:2309.00363, (2023)
103. Kumar, Shobhan, Chauhan, Arun: A finetuned language model for recommending cqa-qas for enriching textbooks. In Kamal Karlapalem, Hong Cheng, Naren Ramakrishnan, R. K. Agrawal, P. Krishna Reddy, Jaideep Srivastava, and Tanmoy Chakraborty, editors, Advances in Knowledge Discovery and Data Mining, pp. 423–435, Cham, (2021). Springer International Publishing
104. Kuratov, Y., Arkhipov, M.: Adaption of deep bidirectional multilingual transformers for russian language. arXiv preprint arXiv:1905.07213, (2019)
105. Kurita, K., Michel, P., Neubig, G,: Weight poisoning attacks on pre-trained models. arXiv preprint arXiv:2004.06660 (2020)
106. Lahire, T.: Actor loss of soft actor critic explained. arXiv preprint arXiv:2112.15568 (2021)
107. Lample, Guillaume: Conneau. Cross-lingual language model pretraining, Alexis (2019)
108. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: A lite bert for self-supervised learning of language representations, (2019)
109. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461. (2019)
110. Li, Junyi, Tang, Tianyi, Zhao, Wayne Xin, Nie, Jian-Yun. Wen, Ji-Rong. A survey of pretrained language models based text generation (2022)
111. Li, J., Tang, T., Zhao, W.X., Wen, J.-R.: Pretrained language models for text generation: A survey, (2021)
112. Li, L., Jiang, X.L.: Pretrained language models for document-level neural machine translation, Qun (2019)
113. Li, L., Song, D., Li, X., Zeng, J., Ma, R., Qiu, X.: Backdoor attacks on pre-trained models by layerwise weight poisoning. arXiv preprint arXiv:2108.13888, (2021)

114. Li, P., Li, L., Zhang, M., Wu, M., Liu, Q: Universal conditional masked language pre-training for neural machine translation. arXiv preprint arXiv:2203.09210 (2022)
115. Li, Qian, Peng, Hao, Li, Jianxin, Xia, Congying, Yang, Renyu, Sun, Lichao, Yu, Philip S., He, Lifang: A survey on text classification: From shallow to deep learning, (2020)
116. Li, S., Liu, H., Dong, T., Zi Hao Zhao, B., Xue, M., Zhu, H., Lu, J.: Hidden backdoors in human-centric language models. arXiv preprint arXiv:2105.00164, (2021)
117. Li, X., Bing, L., Zhang, W.L.: Exploiting bert for end-to-end aspect-based sentiment analysis, Wai (2019)
118. Li, X., Fu, X., Xu, G., Yang, Y., Wang, J., Jin, L., Liu, Q., Xiang, T.: Enhancing bert representation with context-aware embedding for aspect-based sentiment analysis. IEEE Access **8**, 46868–46876 (2020)
119. Lim, S., Lee, K., Kang, J.: Drug drug interaction extraction from the literature using a recursive neural network. PLoS ONE (2018)
120. Lin, Junyang, Men, Rui, Yang, An, Zhou, Chang, Ding, Ming, Zhang, Uichang, Wang, Peng, Wang, Ang, Jiang, Le, Jia, Xianyan, Zhang, Jie, Zhang, Jianwei, Zou, Xu, Li, Zhikang, Deng, Xiaodong, Xue, Jinbao, Zhou, Huiling, Ma, Jianxin, Yu, Jin, Li, Yong, Lin, Wei, Zhou, Jingren, Tang, Jie, Yang, Hongxia: M6: A chinese multimodal pretrainer. arXiv preprint arXiv:2103.00823, (2021)
121. Liu, Jiachang, Shen, Dinghan, Zhang, Yizhe, Dolan, Bill, Carin, Lawrence, Chen, Weizhu: What makes good in-context examples for gpt-3? (2021)
122. Liu, Shansong, Hussain, Atin Sakkeer, Sun, Chenshuo, Shan, Ying: Music understanding llama: Advancing text-to-music generation with question answering and captioning. arXiv preprint arXiv:2308.11276, (2023)
123. Liu, Wenbin, Wen, Bojian, Gao, Shang, Zheng, Jiesheng, Zheng, Yinlong: A multi-label text classification model based on elmo and attention. MATEC Web Conference, 309, (2020)
124. Liu, Yinhan, Ott, Myle, Goyal, Naman, Du, Jingfei, Joshi, Mandar, Chen, Danqi, Levy, Omer, Lewis, Mike, Zettlemoyer, Luke, Stoyanov, Veselin: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, (2019)
125. Liu, Zheng:. Sociological perspectives on artificial intelligence: A typological reading. Wiley Online Library, (2021)
126. Lloret, Elena: Llorens, Hector, Moreda, Paloma, Saquete, Estela, Palomar, Manuel: Text summarization contribution to semantic question answering: New approaches for finding answers on the web. International Journal of Intelligent Systems **26**(12), 1125–1152 (2011)
127. Lock, S.: What is ai chatbot phenomenon chatgpt and could it replace humans?):'Book What is AI chatbot phenomenon ChatGPT and could it replace humans, (2022)
128. Ma, Chunlan, ImaniGooghari, Ayyoob, Ye, Haotian, Asgari, Ehsaneddin, Schütze, Hinrich: Taxi1500: A multilingual dataset for text classification in 1500 languages. arXiv preprint arXiv:2305.08487, (2023)
129. Ma, Shuming, Yang, Jian, Huang, Haoyang, Chi, Zewen, Dong, Li, Zhang, Dongdong, Awadalla, Hany Hassan, Muzio, Alexandre, Eriguchi, Akiko, Singhal, Saksham, Song, Xia, Menezes, Arul, Wei, Furu: Xlm-t: Scaling up multilingual machine translation with pretrained cross-lingual transformer encoders, (2020)
130. MacCartney, Bill: Natural Language Inference. Stanford University ProQuest Dissertations Publishing, (2009)
131. Madhyastha, Pranava Swaroop, Bansal, Mohit, Gimpel, Kevin, Livescu, Karen: Mapping unseen words to task-trained embedding spaces. Proceedings of the 1st Workshop on Representation Learning for NLP, pp. 100–110, (2016)
132. Mager, Manuel, Astudillo, Ramon Fernandez, Naseem, Tahira, Sultan, Md Arafat, Lee, Young-Suk, Florian, Radu, Roukos, Salim: Gpt-too: A language-model-first approach for amr-to-text generation, (2020)
133. Mai, Florian, Pappas, Nikolaos, Montero, Ivan, Smith, Noah A.: Henderson. Plug and play autoencoders for conditional text generation, James (2020)
134. Maldonado, Abran, Pistunovich, Natalie: GPT-3 powers the next generation of apps, (2021)
135. Manias, George, Mavrogiorgou, Argyro, Kiourtis, Athanasios, Symvoulidis, Chrysostomos, Kyriazis, Dimosthenis: Multilingual text categorization and sentiment analysis: a comparative analysis of the utilization of multilingual approaches for classifying twitter data. Neural Computing and Applications, pp. 1–17, (2023)
136. Martin, Louis, Muller, Benjamin, Suárez, Pedro Javier Ortiz, Dupont, Yoann, Romary, Laurent, de la Clergie, Éric Villemonte, Seddah, Djamé, Sagot, Benoit: Camembert: a tasty french language model. arXiv preprint arXiv:1911.03894, (2020)
137. Marulli, Fiammetta: Verde, Laura, Campanile, Lelio: Exploring data and model poisoning attack to deep learning-based NLP systems. Procedica Computer Science **192**, 3570–3579 (2021)
138. Maslennikova, Elizaveta: Elmo word representations for news protection. CLEF (Working Notes, (2019)
139. Mathew, Leeja, Bindu, V. R.: A review of natural language processing techniques for sentiment analysis using pre-trained models. In: 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), pp. 340–345, (2020)
140. McCarley, J.S.: Chakravarti, Rishav. Sil. Structured pruning of a bert-based question answering model, Avirup (2019)
141. Arifuzzaman, M., Rakibul Hasan, Md., Maliha, Maisha: Sentiment analysis with nlp on twitter data. IEEE, (2019)
142. Meftah, Sara, Tamaazoust, Youssef, Semmar, Nasredine, Essafi, Hassane, Sadat, Faitha: Joint learning of pre-trained and random units for domain adaptation in part-of-speech tagging. arXiv preprint arXiv: 1904.03595, (2019)
143. Meng, Yuxian, Ren, Xiangyuan, Sun, Zijun, Li, Xiaoya, Yuan, Arianna, Wu, Fei, Li, Jiwei: Large-scale pretraining for neural machine translation with tens of billions of sentence pairs, (2019)
144. Minaee, Shervin, Kalchbrenner, Nal, Cambria, Erik, Nikzad, Narjes, Chenaghlu, Meysam, Gao, Jianfeng: Deep learning–based text classification: A comprehensive review. ACM Comput. Surv., 54(3), (April 2021)
145. Mitchell, Lewis, Frank, Morgan R., Harris, Kameron Decker, Dodds, Peter Sheridan, Danforth, Christopher M.: The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. PloS one **8**(5), e64417 (2013)
146. Mitkov, Ruslan: The upper Oxford Handbook of Computational Linguistics. Oxford University Press Inc., (2004)
147. Mohawesh, Rami, Al-Hawawreh, Muna, Maqsood, Sumbal, Alqudah, Omar: Factitious or fact? learning textual representations for fake online review detection. Cluster Computing, pp. 1–16, (2023)
148. Mohawesh, Rami: Liu, Xiao, Arini, Hilya Mudrika, Wu, Yutao, Yin, Hui: Semantic graph based topic modelling framework for multilingual fake news detection. AI Open **4**, 33–41 (2023)
149. Mohawesh, Rami, Xu, Shuxiang, Springer, Matthew, Al-Hawawreh, Muna, Maqsood, Sumbal: Fake or genuine? contextualised text representation for fake review detection. arXiv preprint arXiv:2112.14343, (2021)
150. Mohawesh, Rami: Xu, Shuxiang, Springer, Matthew, Jararweh, Yaser, Al-Hawawreh, Muna, Maqsood, Sumbal: An explainable

ensemble of multi-view deep learning model for fake review detection. Journal of King Saud University-Computer and Information Sciences **35**(8), 101644 (2023)
151. Mohit, Behrang: Natural Language Processing of Semitic Languages. Springer, Berlin, Heidelberg (2014)
152. Mumtarin, Maroa, Samiullah Chowdhury, Md., Wood, Jonathan: Large language models in analyzing crash narratives–a comparative study of chatgpt, bard and gpt-4. arXiv preprint arXiv:2308.13563, (2023)
153. Nadeau, David: Sekine, Satoshi: A survey of named entity recognition and classification. Lingvisticæ Investigationes **30**, 3–26 (2007)
154. Narang, Sharan, Chowdhery, Aakanksha: Pathways language model (palm): Scaling to 540 billion parameters for breakthrough performance. Google AI Blog, (2022)
155. Narayan, Shashi, Simoes, Gonçalo, Ma, Ji, Craighead, Hannah, Mcdonald, Ryan: Qurious: Question generation pretraining for text generation, (2020)
156. Usman Naseem, Matloob Khushi, Vinay Reddy, Sakthivel Rajendran, Imran Razzak, and Jinman Kim. Bio albert: A simple and effective pre-trained language model for biomedical named entity recognition. International Joint Conference on Neural Networks, 2021
157. Nayak, Pandu: Understanding searches better than ever before, (Oct 2019)
158. Nguyen, Dat Quoc, Nguyen, Anh Tuan: Phobert: Pre-trained language models for vietnamese. arXiv preprint arXiv:2003.00744, (2020)
159. Nguyen, Thanh Thi, Wilson, Campbell, Dalins, Janis: Fine-tuning llama 2 large language models for detecting online sexual predatory chats and abusive texts. arXiv preprint arXiv:2308.14683, (2023)
160. Okur, Halil Ibrahim, Sertbaş, Ahmet: Pretrained neural models for turkish text classification. In: 2021 6th International Conference on Computer Science and Engineering (UBMK), pp. 174–179, (2021)
161. Orgad, Hadas, Belinkov, Yonatan: Debiasing nlp models without demographic information. arXiv preprint arXiv:2212.10563, (2022)
162. Padilla, Jose J., Kavak, Hamdi, Lynch, Christopher J., Gore, Ross J., Diallo, Saikou Y.: Temporal and spatiotemporal investigation of tourist attraction visit sentiment on twitter. PloS one **13**(6), e0198857 (2018)
163. Penha, Gustavo, Hauff, Claudia: What does BERT know about books, movies and music? probing BERT for conversational recommendation. In: Fourteenth ACM Conference on Recommender Systems. ACM, (sep 2020)
164. Polignano, M., Basile, P., de Gemmis, M., Semeraro, G., Basile, V.: Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. CEUR Workshop Proceedings, 2481, (2019)
165. Etoori, Pravallika: Mamidi, Radhika: Chinnakotla. Automatic spelling correction for resource-scarce languages using deep learning. ACL Anthology, Manoj (2018)
166. Qi, Ye, Sachan, Devendra Singh, Felix, Matthieu, Padmanabhan, Sarguna Janani, Neubig, Graham: When and why are pre-trained word embeddings useful for neural machine translation?, (2018)
167. Qiu, Xipeng, Sun, Tianxiang, Yige, Xu., Shao, Yunfan, Dai, Ning, Huang, Xuanjing: Pre-trained models for natural language processing: A survey. Science China Technological Sciences **63**(10), 1872–1897 (2020)
168. Qu, Chen, Yang, Liu, Qiu, Minghui, Bruce Croft, W., Zhang, Yongfeng, Iyyer, Mohit: BERT with history answer embedding for conversational question answering. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, (jul 2019)
169. Qu, Yuanbin, Liu, Peihan, Song, Wei, Liu, Lizhen, Cheng, Miaomiao: A text generation and prediction system: Pre-training on new corpora using bert and gpt-2. In: 2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC), pp. 323–326, (2020)
170. Quan, Wei, Zhang, Jinli, Hu, Xiaohua Tony: End-to-end joint opinion role labeling with bert. In: 2019 IEEE International Conference on Big Data (Big Data), pp. 2438–2446, (2019)
171. Radford, Alec: Narasimhan. Improving language understanding by generative pre-training. OpenAI, Karthik (2018)
172. Radford, Alec, Wu, Jeffrey, Child, Rewon, Luan, David, AModei, Dario, Sutskever, Ilya: Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. ACL Anthology, (2019)
173. Rae, Jack W., Borgeaud, Sebastian, Cai, Trevor, Millican, Katie, Hoffmann, Jordan, Song, Francis, Aslanides, John, Henderson, Sarah, Ring, Roman, Young, Susannah, et al.: Scaling language models: Methods, analysis & insights from training gopher. arXiv preprint arXiv:2112.11446, (2021)
174. Raffel, Colin: Shazeer, Noam, Roberts, Adam, Lee, Katherine, Narang, Sharan, Matena, Michael, Zhou, Yanqi, Li, Wei, Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, Peter J. (2019)
175. Rahsepar, Amir Ali, Tavakoli, Neda, Kim, Grace Hyun J., Hassani, Cameron, Abtin, Fereidoun, Bedayat, Arash: How ai responds to common lung cancer questions: Chatgpt vs google bard. Radiology, 307(5):e230922, (2023)
176. Ramponi, Alan, Plank, Barbara: Neural unsupervised domain adaptation in nlp—a survey. arXiv preprint arXiv:2006.00632, (2020)
177. Ramraj, S., Arthi, R., Murugan, Solai, Julie, M.S.: Topic categorization of tamil news articles using pretrained word2vec embeddings with convolutional neural network. In: 2020 International Conference on Computational Intelligence for Smart Power System and Sustainable Energy (CISPSSE), pp. 1–4, (2020)
178. Rehman, Abdul, Abbasi, Rabeeh Ayaz, Khattak, Akmal Saeed, et al.: Classifying text-based conspiracy tweets related to covid-19 using contextualized word embeddings. arXiv preprint arXiv:2303.03706, (2023)
179. Reimers, Nils, Schiller, Benjamin, Beck, Tilmann, Daxenberger, Johannes, Stab, Christian, Gurevych, Iryna: Classification and clustering of arguments with contextualized word embeddings. arXiv preprint arXiv:1906.09821, (2019)
180. Rezaeinia, Seyed Mahdi, Rahmani, Rouhollah, Ghodsi, Ali, Veisi, Hadi: Sentiment analysis based on improved pre-trained word embeddings. Expert Systems with Applications **117**, 139–147 (2019)
181. Rosario, Barbara, Hearst, Marti A.: Classifying semantic relations in bioscience texts. Proceedings of the 42nd Annual meeting of the association for computational linguistics, (2004)
182. Roudsari, Arousha Haghighian, Afshar, Jafar, Lee, Charles Cheolgi, Lee, Wookey: Multi-label patent classification using attention-aware deep learning model. In: 2020 IEEE International Conference on Big Data and Smart Computing (BigComp), pp. 558–559, (2020)
183. Sarkar, Sagnik, Singh, Pardeep: Combining the knowledge graph and t5 in question answering in nlp. In: Sentiment Analysis and Deep Learning: Proceedings of ICSADL 2022, pp. 405–409. Springer, (2023)
184. Saunders, Danielle: Domain adaptation and multi-domain adaptation for neural machine translation: A survey. arXiv preprint arXiv:2104.06951, (2021)

185. Schmid, Helmut: Part-of-speech tagging with neural networks. arXiv preprint arXiv:9410018, (1994)
186. Sen, Bhaskar, Gopal, Nikhil, Xue, Xinwei: Support-bert: Predicting quality of question-answer pairs in msdn using deep bidirectional transformer, (2020)
187. Shi, Yucheng, Ma, Hehuan, Zhong, Wenliang, Mai, Gengchen, Li, Xiang, Liu, Tianming, Huang, Junzhou: Chatgraph: Interpretable text classification by converting chatgpt knowledge to graphs. arXiv preprint arXiv:2305.03513, (2023)
188. Singhal, Karan, Tu, Tao, Gottweis, Juraj, Sayres, Rory, Wulczyn, Ellery, Hou, Le, Clark, Kevin, Pfohl, Stephen, Cole-Lewis, Heather, Neal, Darlene, et al.: Towards expert-level medical question answering with large language models. arXiv preprint arXiv:2305.09617, (2023)
189. Song, Youwei, Wang, Jiahai, Liang, Zhiwei, Liu, Zhiyue, Jiang, Tao: Utilizing bert intermediate layers for aspect based sentiment analysis and natural language inference, (2020)
190. Stickland, Asa Cooper, Li, Xian: Ghazvininejad. Recipes for adapting pre-trained monolingual and multilingual models to machine translation, Marjan (2020)
191. Strubell, Emma, Ganesh, Ananya, McCallum, Andrew: Energy and policy considerations for deep learning in nlp. arXiv preprint arXiv:1906.02243, (2019)
192. Sun, Chi: Huang, Luyao. Qiu. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence, Xipeng (2019)
193. Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune bert for text classification?, 2019
194. Sun, Yu, Wang, Shuohuan, Feng, Shikun, Ding, Siyu, Pang, Chao, Shang, Junyuan, Liu, Jiaxiang, Chen, Xuyi, Zhao, Yanbin, Lu, Yuxiang, Liu, Weixin, Wu, Zhihua, Gong, Weibao, Liang, Jianzhong, Shang, Zhizhou, Sun, Peng, Liu, Wei, Ouyang, Xuan, Yu, Dianhai, Tian, Hao, Wu, Hua, Wang, Haifeng: Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation, (2021)
195. Suneera, C. M., Prakash, Jay: A bert-based question representation for improved question retrieval in community question answering systems. In Srikanta Patnaik, Xin-She Yang, and Ishwar K. Sethi, editors, Advances in Machine Learning and Computational Intelligence. Springer Singapore, (2021)
196. Sweeney, Latanya: Discrimintion in online ad delivery. arXiv preprint arXiv:1301.6822, (2013)
197. Tabinda Kokab, Sayyida, Asghar, Sohail, Naz, Shehneela: Transformer-based deep learning models for the sentiment analysis of social media data. Array, page 100157, (2022)
198. Tanvir, Hasan, Kittask, Claudia, Eiche, Sandra, Sirts, Kairit: Estbert: a pretrained language-specific bert for estonian. arXiv preprint arXiv:2011.04784, (2021)
199. Terpin, Antonio: Lanzetti, Nicolas, Yardim, Batuhan, Dorfler, Florian, Ramponi, Giorgia: Trust region policy optimization with optimal transport discrepancies: Duality and algorithm for continuous actions. Advances in Neural Information Processing Systems **35**, 19786–19797 (2022)
200. Balaji, T.K., Annushree, Bablani, and Sreeja, S.R.: Opinion mining on covid-19 vaccines in india using deep and machine learning approaches. In: 2022 International Conference on Innovative Trends in Information Technology (ICITIIT), pp. 1–6, (2022)
201. Touvron, Hugo, Lavril, Thibaut, Izacard, Gautier, Martinet, Xavier, Lachaux, Marie-Anne, Lacroix, Timothée, Rozière, Baptiste, Goyal, Naman, Hambro, Eric, Azhar, Faisal, et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, (2023)
202. Ulcar, Matej: Robnik-Sikonja, Marko: Training dataset and dictionary sizes matter in bert models: the case of baltic languages. Analysis of Images, Social Networks and Texts (2021)
203. Uthus, David, Ontañón, Santiago, Ainslie, Joshua, Guo, Mandy: mlongt5: A multilingual and efficient text-to-text transformer for longer sequences. arXiv preprint arXiv:2305.11129, (2023)
204. van Stegeren, Judith, Myundefinedliwiec, Jakub: Fine-tuning gpt-2 on annotated rpg quests for npc dialogue generation. In: The 16th International Conference on the Foundations of Digital Games (FDG) 2021. Association for Computing Machinery, (2021)
205. Variš, Duš an, Bojar, Ondřej: Unsupervised pretraining for neural machine translation using elastic weight consolidation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. Association for Computational Linguistics, (2019)
206. Veysel Kocaman, David Talby: SparkNLP: Natural language understanding at scale. Elsevier, (2021)
207. Virtanen, Antti, Kanerva, Jenna, Ilo, Rami, Luoma, Jouni, Luotolahti, Juhani, Salakoski, Tapio, Ginter, Filip, Pyysalo, Sampo: Multilingual is not enough: Bert for finnish. arXiv preprint arXiv:1912.07076, (2019)
208. Wang, Hai, Yu, Dian, Sun, Kai, Chen, Jianshu, Yu, Dong: Improve pre-trained multilingual models with vocabulary expansion. arXiv preprint arXiv:1909.12440, (2019)
209. Wang, Shuo, Nepal, Surya, Rudolph, Carsten, Grobler, Marthie, Chen, Shangyu, Chen, Tianle: Backdoor attacks against transfer learning with pre-trained deep learning models. arXiv preprint arXiv:2001.03274, (2020)
210. Wang, Wenxuan: Jiao, Wenxiang, Hao, Yongchang, Wang, Xing, Shi, Shuming, Tu, Zhaopeng. Lyu. Understanding and improving sequence-to-sequence pretraining for neural machine translation, Michael (2022)
211. Wang, Yuhui, He, Hao, Tan, Xiaoyang: Truly proximal policy optimization. In: Uncertainty in Artificial Intelligence, pp. 113–122. PMLR, (2020)
212. Wei, Xiaokai, Wang, Shen, Zhang, Dejiao, Bhatia, Parminder, Arnold, Andrew: Knowledge enhanced pretrained language models: A compreshensive survey, (2021)
213. Wiggers, Kyle: (2021)
214. Wikipedia contributors. Turing test — Wikipedia, the free encyclopedia, (2022). [Online; accessed 26-April-2022]
215. Wu, Carole-Jean, Raghavendra, Ramya, Gupta, Udit, Acun, Bilge, Ardalani, Newsha, Maeng, Kiwan, Chang, Gloria, Behram, Fiona Aga, Huang, James, Bai, Charles, Gschwind, Michael, Gupta, Anurag, Ott, Myle, Melnikov, Anastasia, Candido, Salvatore, Brooks, David, Chauhan, Geeta, Lee, Benjamin, Lee, Hsien-Hsin S., Akyildiz, Bugra, Balandat, Maximilian, Spisak, Joe, Jain, Ravi, Rabbat, Mike, Hazelwood, Kim: Sustainable ai: Environmental implications, challenges and opportunities. arXiv, (2021)
216. Xia, Congying, Zhang, Chenwei, Nguyen, Hoang, Zhang, Jiawei, Yu, Philip: Cg-bert: Conditional text generation with bert for generalized few-shot intent detection, (2020)
217. Xing, Yiran, Shi, Zai, Meng, Zhao, Lakemeyer, Gerhard, Ma, Yunpu, Wattenhofer, Roger: Km-bart: Knowledge enhanced multimodal bart for visual commonsense generation, (2021)
218. Xu, Haoran, Van Durme, Benjamin, Murray, Kenton: Bert, mbert, or bibert? a study on contextualized embeddings for neural machine translation. ACL Anthology, (2021)
219. Xu, Hu., Shu, Lei, Yu, Philip S.: Liu. Understanding pre-trained bert for aspect-based sentiment analysis, Bing (2020)
220. Xue, Linting, COnstant, Noah, Roberts, Adam, Kale, Mihir, Al-Rfou, Rami, Siddhant, Aditya, Barua, Aditya, Raffel, Colin: mt5: A massively pre=trained text-to-text transformer. arXiv preprint arXiv:2010.11934, (2021)
221. Yang, Wei, Xie, Yuqing, Lin, Aileen, Li, Xingyu, Tan, Luchen, Xiong, Kun, Li, Ming, Lin, Jimmy: End-to-end open-domain question answering with. In: Proceedings of the 2019

Conference of the North. Association for Computational Linguistics, (2019)
222. Yang, Wei: Xie, Yuqing, Tan, Luchen, Xiong, Kun, Li, Ming. Lin. Data augmentation for bert fine-tuning in open-domain question answering, Jimmy (2019)
223. Yang, Zhilin, Dai, Zihang, Yang, Yiming, Carbonell, Jaime G., Salakhutdinov, Ruslan, Le, Quoc V.: Xlnet: Generalized autoregressive pretraining for language understanding. CoRR, abs/1906.08237, (2019)
224. Yu, Wenhao, Zhu, Chenguang, Li, Zaitang, Hu, Zhiting, Wang, Qingyun, Ji, Heng, Jiang, Meng: A survey of knowledge-enhanced text generation. ACM Comput. Surv., (jan 2022)
225. Zaib, Munazza, Tran, Dai Hoang, Sagar, Subhash, Mahmood, Adnan, Zhang, Wei E., Sheng, Quan Z.: Bert-coqac: Bert-based conversational question answering in context. In Li Ning, Vincent Chau, and Francis Lau, editors, Parallel Architectures, Algorithms and Programming, pp. 47–57, Singapore, (2021). Springer Singapore
226. Zajko, M.: Artificial intelligence, algorithms, and social inequality: Sociological contributions to contemporary debates. Sociology Compass, (2022)
227. Zhang, B., Dafoe, A.: Artificial intelligence: American attitudes and trends. Governance of AI program, the Future of Humanity Institute, the University of Oxford, Oxford, UK (2019)
228. Zhang, B., Yang, H., Liu, X.-Y.: Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models. arXiv preprint arXiv:2306.12659, (2023)
229. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858, (2023)
230. Zhang, H., Song, H., Li, S., Zhou, Ming, Song. A survey of controllable text generation using transformer-based pre-trained language models, Dawei (2022)
231. Zhang, J., Zhao, Y., Saleh, M., Liu, P.J.: Pegasus: Pre-training with extracted gap-sentences for abstractive summarization (2019)
232. Zhang, T., Xu, B., Thung, F., Haryono, S.A., Lo, D., Jiang, L.: Sentiment analysis for software engineering: How far can pre-trained transformer models go? In: 2020 IEEE International Conference on Software Maintenance and Evolution (ICSME), pp. 70–80, (2020)
233. Zhang, Z., Wu, S., Jiang, D., Chen, G.: BERT-JAM: Maximizing the utilization of BERT for neural machine translation. Neurocomputing **14**(460), 84–94 (2021)
234. Zhu, Jinhua: Xia, Yingce, Wu, Lijun, He, Di, Qin, Tao, Zhou, Wengang, Li, Houqiang. Liu. Incorporating bert into neural machine translation, Tie-Yan (2020)

Devon Myers is a Master of Science in Computer Science student at Duquesne University, USA. His research interest includes Foundation Models and Large Language Models.

Rami Mohawesh got his B.S. in computer science from AlAl-bayt University, his M.S. in computer science from the Jordan University of Science and Technology, and his PhD in computer science from the University of Tasmania in Tasmania, Australia. In his Ph.D. research, he was the first researcher to investigate the concept drift in fake review detection. He worked as a post-doctoral research fellow at Deakin University, Australia. He is currently an assistant professor at Al-Ain University, UAE.

Venkata Ishwarya Chellaboina is a Master of Science in Computer Science student at Duquesne University, USA. Her research interest includes Foundation Models and Large Language Models.

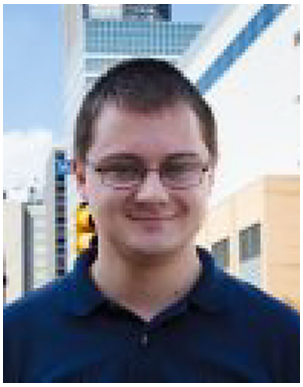
Anantha Lakshmi Sathvik is a Master of Science in Computer Science student at Duquesne University, USA. His research interest includes Foundation Models and Large Language Models.

Praveen Venkatesh is a Master of Science in Computer Science student at Duquesne University, USA. His research interest includes Foundation Models and Large Language Models.

Yi-Hui Ho is a Master of Science in Computer Science student at Duquesne University, USA. Her research interest includes Foundation Models and Large Language Models.

Hanna Henshaw is a Master of Science in Computer Science student at Duquesne University, USA. Her research interest includes Foundation Models and Large Language Models.

Muna Al-Hawawreh received the B.E. and M.E. degrees in computer science from Mutah University, Jordan. She is currently pursuing the Ph.D. degree with the University of New South Wales (UNSW), Canberra, Australia. She works as a Research Assistant at UNSW Canberra Cyber. In her Ph.D. degree, she developed the world's first ransomware framework targeting IIoT edge gateway in the critical infrastructure. Her research interests include cloud computing, industrial control systems, the Internet of Things, cybersecurity, and deep learning. She is a program committee member and a reviewer for several cybersecurity conferences. She was awarded the First Prize for high impact publications in the School of Engineering and Information Technology (SEIT), UNSW, in 2019, and the Dr. K. W. Wang Best Paper Award (2018–2020). She is a Reviewer of high-impact factor journals, such as the IEEE internet of things journal, IEEE transcations on industrial informatics, and IEEE transaction on dependable and secure computing.



David Berdik is a two-time graduate of Duquesne University. He received his Bachelor of Science in Computer Science in December 2019 followed by his Master of Science in Computer Science in December 2021. Professionally, he has three years of industry experience as a software engineer split between two years at General Dynamics Mission Systems and one year at Amazon. As of November 2023, David works as a full-time instructor of computer science at Duquesne University.

Yaser Jararweh received the Ph.D. degree in computer engineering from the University of Arizona in 2010. He was also an Associate Professor of computer science, at Jordan University of Science and Technology, Jordan. He is currently a Professor of computer science with the Department of Mathematics and Computer Science, at Duquesne University, and a Visiting Research Associate Professor, at Carnegie Mellon University, USA. He has coauthored many technical articles in established journals and conferences in fields related to cloud and edge computing, software-defined systems, the Internet of Things, and big data. He is also a Chair of many IEEE events, such as ICICS, SNAMS, BDSN, IoTSMS, BCCA, and many others. He served as a Guest Editor for many special issues in different established journals. He is also the Co-Editor Chief of Cluster Computing Journal (Springer), an Editor for Information Processing and Management (Elsevier), and many others.

Authors and Affiliations

Devon Myers[1] · Rami Mohawesh[2] · Venkata Ishwarya Chellaboina[1] · Anantha Lakshmi Sathvik[1] · Praveen Venkatesh[1] · Yi-Hui Ho[1] · Hanna Henshaw[1] · Muna Alhawawreh[3] · David Berdik[1] · Yaser Jararweh[1]

✉ Rami Mohawesh
rami.mohawesh@aau.ac.ae

Devon Myers
myersd2@duq.edu

Venkata Ishwarya Chellaboina
chellaboinav@duq.edu

Anantha Lakshmi Sathvik
tirukkovalluria@duq.edu

Praveen Venkatesh
venkateshp@duq.edu

Yi-Hui Ho
hoy@duq.edu

Hanna Henshaw
henshawh@duq.edu

Muna Alhawawreh
muna.alhawawreh@deakin.edu.au

David Berdik
dgberdik@gmail.com

Yaser Jararweh
yaser@email.arizona.edu

[1] Duquesne University, Pittsburgh, Pennsylvania, USA

[2] Al Ain University, Al Ain, Abu Dhabi, United Arab Emirates

[3] Deakin University, Geelong, Melbourne, Australia