IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 27 November 2023, accepted 10 December 2023, date of publication 4 January 2024,
date of current version 16 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3349952

SURVEY

A Survey of Text Classification With Transformers: How Wide? How Large? How Long? How Accurate? How Expensive? How Safe?

JOHN FIELDS[1,2], **(Graduate Student Member, IEEE), KEVIN CHOVANEC**[2], **AND PRAVEEN MADIRAJU**[2]
[1]Business Analytics, Concordia University Wisconsin-Ann Arbor, Mequon, WI 53097, USA
[2]Department of Computer Science, Marquette University, Milwaukee, WI 53233, USA

Corresponding author: John Fields (john.fields@cuw.edu)

This work was supported by the Northwestern Mutual Data Science Institute and NSF under Award 1950826.

ABSTRACT Text classification in natural language processing (NLP) is evolving rapidly, particularly with the surge in transformer-based models, including large language models (LLM). This paper presents an in-depth survey of text classification techniques across diverse benchmarks, addressing applications from sentiment analysis to chatbot-driven question-answering. Methodologically, it utilizes NLP-facilitated approaches such as co-citation and bibliographic coupling alongside traditional research techniques. Because new use cases continue to emerge in this dynamic field, the study proposes an expanded taxonomy of text classification applications, extending the focus beyond unimodal (text-only) inputs to explore the emerging field of multimodal classification. While offering a comprehensive review of text classification with LLMs, this review highlights novel questions that arise when approaching the task with transformers: It evaluates the use of multimodal data, including text, numeric, and columnar data, and discusses the evolution of text input lengths (tokens) for long text classification; it covers the historical development of transformer-based models, emphasizing recent advancements in LLMs; it evaluates model accuracy on 358 datasets across 20 applications, with results challenging the assumption that LLMs are universally superior, revealing unexpected findings related to accuracy, cost, and safety; and it explores issues related to cost and access as models become increasingly expensive. Finally, the survey discusses new social and ethical implications raised when using LLMs for text classification, including bias and copyright. Throughout, the review emphasizes the importance of a nuanced understanding of model performance and a holistic approach to deploying transformer-based models in real-world applications.

INDEX TERMS NLP, text classification, transformers, survey.

I. INTRODUCTION

In the past five years, large language models have revolutionized natural language processing (NLP), achieving state-of-the-art across several classic NLP tasks. One of these tasks, text classification, is a diverse and growing set of aims in academia and industry related to categorizing and organizing text. In text classification, the goal is to assign some label, category, or tag to a body of text (sentence, paragraph, document). Traditionally, text classification, like classification tasks more generally, can be divided into three types:

- Binary classification: classifying texts into one of two mutually exclusive categories (for example, Spam or Not Spam)
- Multiclass classification: dividing texts into one of three or more mutually exclusive categories (for example, classifying a text's genre)

The associate editor coordinating the review of this manuscript and approving it for publication was Maria Chiara Caschera.

- Multilabel classification: labeling texts with three or more potentially overlapping labels, in which each text can receive multiple labels (such as offensive comments labeling, in which a comment might be marked for both violence and hate speech)

However, as automated text classification has expanded, common aims and data sources have reappeared often. For example, researchers often work with social media, surveys, scraped web data, emails, user reviews or comments, and they often attempt similar kinds of classification: sentiment analysis, news classification, topic labeling, emotion detection, offensive language labeling, etc.

Text data offers rich information, but it has historically been difficult and expensive to process. LLMs, especially open-source LLMs such as BERT, offer generalized models that can greatly facilitate processing this data. An enormous amount of research over the last half decade has thus been invested in deploying, fine-tuning, and adapting LLMs to text classification tasks. In this literature survey, we aim to provide an overview of the various ways LLMs have been deployed for text-classification, along with a new taxonomy of the common subtypes and an explanation of the central methods that have appeared in the research.

Our paper builds on two recent, related literature surveys. [1] have published a survey of deep learning text classification methods in 2021, covering research through 2020. In the same year, [2] published a detailed account of pre-trained language models used in NLP tasks. We extend the work of these earlier papers, contributing to research in three primary ways:

1) First, we fill in the last two years, which, in this rapidly-developing field, has seen several new models, approaches, and benchmarks. Indeed, the cultural impact of ChatGPT spurred rapid development in the field, and many of the current benchmark models, such as GPT-4, Llama 2, and Titan, were not yet released when these surveys were published.
2) Next, our survey focuses specifically on text classification and LLMs, which allows us to offer a comprehensive overview of the work done in this area. [1], for example, offers only a relatively short section on LLMs, and [2] devote their survey to all uses of LLMs, only briefly discussing classification. Our focus allows us to explore how LLMs perform across several subtasks within text classification. Moreover, [2] offer an excellent technical overview of LLMs, covering model architecture and fine-tuning in detail, while our survey leans more toward the novel practical and ethical questions that researchers have addressed when deploying LLMs for text classification.
3) Finally, based on the articles we survey, we propose an expansion of existing taxonomies of text classification. While still rooted in traditional tasks such as sentiment analysis, news classification, or topic labeling, recent approaches, propelled by the success of LLMs, have seen increasingly granular subtasks that appear often enough in the literature to merit space in the taxonomy, such as fine-grained sentiment analysis or intent recognition. As multimodal approaches are perhaps the most exciting subfield of text classification, this survey also goes beyond previous surveys to include classification methods that pair text data with other kinds of data.

By large language model, we refer to the pre-trained, transformer-based architectures that have been widely adopted since [3] seminal work on attention. These language models, beginning with GPT [4] and BERT [5] in 2018, employ a neural network using a parallel multihead attention mechanism, with either an encoder or encoder-decoder structure, to transform tokens into embeddings, or word vectors, with billions or even trillions of parameters, which can then be used in downstream tasks. While many of the most famous gains from these LLMs have been in classic question-answer problems, they have also been adopted widely for text classification. As Figure 1 shows, LLMs have also recently exploded after the popularity of ChatGPT (GPT-3.5/4), offering new opportunities for integrating these recently released models into research on text classification.

Our survey is organized around the central questions raised by the use of transformer-based LLMs for text classification. Many of these relate to the training or fine-tuning of the model itself, while others relate to the social and ethical questions. In the next section, we present our methodology, which blends traditional survey methods with an approach using new NLP tools. Then, in the third and central section of the work, we offer our novel taxonomy of text-classification.

The rest of the survey's body is then organized around questions uniquely related to text classification with LLMs. First, we consider the type of data used in the model or the 'width' of the model; next, we explore questions related to the size of the model (in 'How Large'); then, in 'How Long,' we consider the length of the documents being classified; 'How Accurate' summarizes benchmarking across text classification subtasks; and 'How Safe' considers the ethical and legal issues related to text classification with LLM.

Finally, we offer a conclusion that summarizes our findings and suggests directions for future research in this rapidly-evolving field.

II. METHODOLOGY

We follow the guidelines for comprehensive literature reviews proposed by [6]. We first formulated the research questions above, and then systematically searched for primary studies through keyword searches focusing on text classification and large language models, including all LLMs listed in table 3: for example, we searched "BERT" and "Text Classification"; "Llama" and "Text Classification"; "GPT" and "Text Classification," etc. Using IEEE Explore and the ACM database, this returned 277 results, after removing duplicates. Initially, Google Scholar was used as a third database, but the results were almost all either redundant

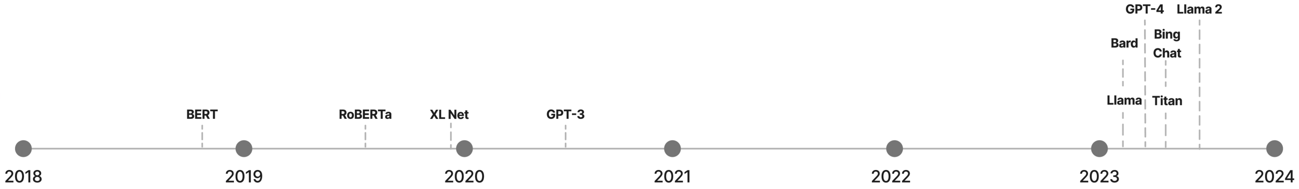

FIGURE 1. Timeline of major transformer-based large language model introductions.

or outside the scope of our survey, and we therefore narrowed the search to include only sources from the IEEE and ACM databases.

We then reviewed the primary studies for their quality and relevance to our topic, and extracted information about the methods, area, and focus of the work. We excluded works not primarily focused on text classification and works not using a transformer-based, pre-trained large language model. We also excluded works published before 2020, since previous surveys cover deep learning techniques through 2020. After removing articles using our exclusion criteria, we found 231 articles relevant to the topic. However, many of these papers overlapped significantly in focus and methodology, as the last three years have seen a proliferation of work using LLMs for text classification; to avoid redundancy, we chose the most representative and most cited works from each category.

Finally, we supplemented this traditional survey methodology in two ways. First, most simply, we used backward snowballing to identify frequently cited papers missed in our original search. We also integrated new NLP tools to enhance our search, namely, Connected Papers, a visual networking tool that explores interrelations between research papers [6]. Connected Papers uses similarity metrics (rather than citations alone) to link papers together, and for any paper searched, it returns a few dozens of the most frequently sighted similar papers. As in other studies that have employed this tool for the purposes of a literature review, we chose seed articles from our initial search for each section in our paper and fed these articles into Connected Papers. An example of the Connected Papers graph and prior/derivative works for [1] can be accessed at [7].

III. TAXONOMY OF TEXT CLASSIFICATION

Classification philosophies have existed for thousands of years to provide structure and order to the world around us [8]. Only in the 1950's and 1960's did we begin to utilize computers to aid in the classification of text-based documents, which was a time-consuming and laborious manually task [9]. These systems also relied on handcrafted, rule-based systems that had limited effectiveness and scalability. The developments from the 1960's to 1990's progressed slowly during the "AI winter" and significant progress was not made until the 1990's with new machine learning techniques and faster computers [10]. In recent years, new transformer-based algorithms were applied to classification problems in the late 2010's with great success and this expanded the potential use cases for text classification as shown in Table 1.

TABLE 1. Current taxonomy of text classification.

Type	Definition
Sentiment analysis	Classification based on the polarity (positive, negative, or neutral) or emotional tone of text.
News classification	Categorize news articles based on content and subject matter.
Topic labeling	Identify the main subject or topic in text.
Question-answering	Provide accurate answers to user queries.
Natural language inference	Determine the logical relationship (entailment, contradiction, or neutral) between sentences.
Dialog act classification	Identification of text intention such as requesting, questioning, or informing.
Named entity recognition	Categorize named entities such as people, organizations, and locations.
Syntactic parsing	Analyze the grammatical structure of text to understand the relationship between words and phrases.

Sources: [11]–[13]

The taxonomy in Table 1 has been sufficient for the early period of transformer-based text classification but there is a need for improved test datasets and an expanded taxonomy of applications [14]. With the advent of ChatGPT and other LLMs, the text classification capabilities and applications have expanded, and we propose evolving the above taxonomy as shown in Table 2 [15].

Some additional classification applications that are emerging are fine-grained sentiment analysis, aspect-based sentiment analysis, offensive language detection, intent recognition, document classifiers, fake news detection, cross-lingual classification, stance detection, emotion/mental health detection, malicious software detection, cause and effect classifiers, sentence classification, multilabel, multi-modal, and other applications. Some of these new areas, like emotion/mental health detection involve more complex social and ethical issues that will be described more fully in Section IX. However, as we develop systems with more "human like" capabilities, we should consider how these fit

TABLE 2. Additional taxonomy categories for text classification.

Type	Definition	Source
Fine-grained sentiment analysis	Capture more nuance of sentiment and work better with multiclass problems.	[16]
Aspect-based sentiment analysis	More detailed sentiment (e.g., not just like or dislike a product but by feature/aspect).	[17]
Offensive language detection	Identify text that contains abusive, offensive, or inappropriate content.	[18]
Intent recognition	Train on dialog to understand user intent.	[19]
Document classifier	Use new techniques for longer text documents.	[20]
Fake news classification	Identify patterns to predict true/fake news.	[21]
Cross-lingual classification	Work with multiple languages for a more global context.	[22]
Stance detection	Determine the perspective (support, opposition, or neutral) toward a piece of text.	[23]
Emotion/mental health detection	Assess the mental health status of a person from analyzing text.	[24]
Sentence classification	Determine the class of an entire sentence.	[25]
Multilabel	Assign multiple labels to a single input text.	[26]
Multimodal	Utilizing multiple types of input data such as text, video, audio, etc.	[1]

[1] See Section 4.2.

into the taxonomy and propose rules and guidelines for how we implement these safely, or not at all.

IV. HOW WIDE?

In addition to the applications described in Section III, we also consider "How wide?" in reference to the different data types and methods used for analysis. Text classification tasks with transformers can be categorized as unimodal or multimodal.

A. UNIMODAL

Unimodal text classification uses only textual information and applies the transformer model to make a classification prediction. Many survey papers have focused on traditional NLP methods and newer transformer-based methods applied to text classification [1], [13], [27], [28]. Since the launch of ChatGPT in late 2022, there has been a significant increase in the research and investment in LLMs. The primary application for these models is typically question-answer chat bots, explored more in Section VII. Table 3 below provides a summary of the major releases of LLMs that can perform unimodal text classification.

TABLE 3. Large language model examples and release dates.

Name	Organization	Date	Source
BERT	Google	Oct 2018	[5]
RoBERTa	Meta	Jul 2019	[29]
XL Net	Google	Dec 2019	[30]
GPT-3	OpenAI	Jun 2020	[31]
Llama	Meta	Feb 2023	[32]
Bard(LaMBDA)	Bloomberg	Mar 2023	[33]
GPT-4	OpenAI	Mar 2023	[34]
BloombergGPT	Bloomberg	Mar 2023	[35]
Dolly 2	Databricks	Apr 2023	[36]
StableLM	Stability AI	Apr 2023	[37]
Titan	Amazon	Apr 2023	[38]
Bing Chat	Microsoft	Apr 2023	[39]
Llama 2	Meta	Jul 2023	[40]

The primary differentiator for these LLMs is the size of the training data, and this aspect of LLMs is covered in detail in Section V. The technical differences between encoder-decoder/encoder only (BERT-method), or decoder-only (GPT-method) are covered in detail in the papers by [12] and [13].

These models have also been fine-tuned for domain-specific unimodal text classification. One example is the BloombergGPT model which is trained on financial text data for internal use by the company. This is an area for future research to determine if the investment in proprietary, domain specific LLM could be a competitive advantage for the companies who make these investments.

Multimethod approaches for unimodal text classification are also achieving best-in-class results for some applications such as aspect-based sentiment analysis. The combination of graph database methods with transformer-based text classification is another emerging area in NLP. Table 4 below shows two recent uses of this technique.

TABLE 4. Multimethod examples in aspect-based sentiment analysis.

Model	Paper
HGCN+BERT	*Learn from Structural Scope: Improving Aspect-Level Sentiment Analysis with Hybrid Graph Convolutional Networks* [41]
HAABSA	*A Hybrid Approach for Aspect-Based Sentiment Analysis Using a Lexicalized Domain Ontology and Attentional Neural Models* [42]

B. MULTIMODAL

Multimodal classification uses text, video, signal, image, audio, and columnar data for classification. A taxonomy for multimodal machine learning proposed by [43] included Representation, Translation, Alignment, Fusion, and Co-learning.

This is an emerging area of investment by large technology companies as evidenced by Elon Musk's goal of creating

TABLE 5. Estimated data types in businesses and other organizations.

Type	Proportion
Text	$\approx 60 - 80\%$
Numeric	$\approx 15\%$
Video	$\approx 5 - 20\%$
Voice	$\approx 1 - 5\%$

an "everything app" from his new company X.ai [44] and Google's refocusing their AI teams on "...a series of powerful, multimodal AI models" [45]. Some multimodal models, such as UniMSE [46] for sentiment analysis and SEMI-FND [47] for Fake News Detection, are already emerging as best-in-class as shown in Table 10 below.

As we consider multimodal classification, the prior research is primarily focused on the use of image data and text [43]. We hypothesize that the primary types of data in most companies and organizations are text and numeric/columnar data. We were unable to find academic research on the breakdown of data types most common in a business context, so we utilized ChatGPT [48] and Bard [49] to provide the following estimates.

Note: ChatGPT would not provide the sources for the estimates in Table 5. Bard provided the following sources:

- The Data & Analytics Association (D&AA) 2022 State of Data & Analytics report
- Gartner's 2022 Magic Quadrant for Data Integration Tools
- IDC's 2022 Worldwide DataSphere Market Forecast
- Statista's 2022 statistics on corporate data storage

If these estimates are accurate, it would indicate that the research on multimodal classification is too focused on video/voice and not enough on numeric or tabular data. For example, the paper by [50] includes a summary of 46 papers on Multimodal Classification with Deep Learning. However, only one of these papers ([51]) focuses on the use of text and tabular data which is commonly found in a variety of business applications such as health care and risk classification [52]. The more recent 2023 paper by [53] includes a similar list of 31 multimodal datasets with a mix of images, text, video, and audio but **none** with numeric/columnar data.

There are several recent surveys by [54], [55], [56], and [57] on the use of deep learning with columnar data but these focus on categorical/numeric data and not text. Since there is a gap in the research on multimodal text and columnar data applications, we will provide additional detail on this important but overlooked area of text classification. One example of a solution using text and tabular data is the Multimodal Toolkit package for Python that was developed by Georgian.io [58]. As mentioned previously, the focus on text and video may be of interest to academics due to the availability of datasets but corporations and organizations have access to text and columnar data with information that could provide valuable insights.

However, the benefits of solutions like Multimodal Toolkit have only shown limited gains by using transformers on text and numeric data [50]. Since this is only one example, we hope this paper will highlight the need to look at the emerging research on using transformers for columnar data and the opportunities to synthesize this research with the extensive effort and money that is focused on multimodal research by technology companies as described above. Future studies may delve into optimizing model architectures, data fusion strategies, and addressing ethical considerations to unlock the full capabilities of multimodal text classification in domains ranging from media analysis to content recommendation systems.

V. HOW LARGE?

> "Our results strongly suggest that larger models will continue to perform better, and will also be much more sample efficient than has been previously appreciated" [59].

The use of transformers for text classification began with BERT and GPT in 2018. Data scientists with an interest in NLP began to see the types of exponential improvements that the field of vision experienced from neural networks in 2012 [60]. The success of "small" transformers like BERT with 110-340 million parameters[1] and RoBERTa with 123-354 million parameters established many new benchmarks in text classification since 2018.

The subsequent creation of the HuggingFace platform provided a platform for data scientists to utilize these transformer-based models [61]. The launch of Google Colab and other cloud-based solutions then provided simpler and more affordable access to Graphical Processing Units (GPU) running these compute intensive applications [62].

The general public did not have extensive direct exposure to transformer-based NLP until the release of ChatGPT in late 2022. This application achieved a new record by reaching 100 million customers two months after the application launch [63]. The ensuing press, investment, and hype over ChatGPT has increased the interest in question-answer based NLP applications. However, the use of this technology for other text classification tasks has not received the same attention but should benefit from the increased focus. This paper will also survey the recent developments and explain how these can be leveraged to improve text classification using transformers. Table 6 below adds a size element to Table 3 to show the rapid increase in the scale of these LLMs.

Although the size of the various LLMs continues to grow, there are several other strategic approaches that should be considered when choosing the best option for text classification. Some of the considerations include:

- Amount of training data.
- Privacy and security.
- Complexity and uniqueness of the task.
- Scalability.
- Speed.
- Compute resources available.

[1] Parameters are the weights and biases that are adjusted during training and used to make predictions.

TABLE 6. Size of selected large language models released since ChatGPT (Nov 2022).

Name	Organization	Date	Size
Llama	Meta	Feb 2023	7B-65B
Bard(LaMBDA)	Bloomberg	Mar 2023	137B
GPT-4	OpenAI	Mar 2023	$\approx 1T$
BloombergGPT	Bloomberg	Mar 2023	50B
Dolly 2	Databricks	Apr 2023	12B
StableLM	Stability AI	Apr 2023	13B
Titan	Amazon	Apr 2023	$\approx 45B$
Bing Chat	Microsoft	Apr 2023	Unknown
Llama 2	Meta	Jul 2023	7B-70B

TABLE 7. Token length of selected large language models released since ChatGPT (Nov 2022).

Name	Organization	Date	Size
Llama	Meta	Feb 2023	2,048
Bard(LaMBDA)	Bloomberg	Mar 2023	$\approx$ 1000
GPT-4	OpenAI	Mar 2023	8,192
BloombergGPT	Bloomberg	Mar 2023	Unknown
Dolly 2	Databricks	Apr 2023	2,048
StableLM	Stability AI	Apr 2023	75
Titan	Amazon	Apr 2023	Unknown
Bing Chat	Microsoft	Apr 2023	Unknown
Llama 2	Meta	Jul 2023	4,096

As the amount of data available on the internet could eventually limit the growth of LLMs, these other factors may become more important as we continue to improve these models [64]. Additional research efforts, like the Pythia suite, are providing new tools to analyze LLMs and address this issue [65]. Other recent survey papers such as [66], seek to address the issue of how to apply more efficient transformer methods to NLP tasks. Approaches are grouped together by sparse, factorized attention, and architectural change. However, [66] concludes there are,

> "…no simple and universal guidelines regarding the current Transformer alternatives."

The cost to develop LLMs is another limiting factor where large technology companies and research institutions have the resources to afford investments in the millions of US dollars [67]. The ability to predict the gains in performance could be beneficial to better understand the value of these investments. However, the reality is that the "democratization" of this technology will not occur in the near future until Graphical Processing Units (GPU) are more accessible and affordable [68].

VI. HOW LONG?

Another factor to consider for text classification is the length of the input text. The original BERT model had a limit of 512 tokens or $\approx$ 400 words. Extensions of BERT such as Big Bird [69] and Longformer [70] extended the token limit to 4096 to handle longer text such as essays. Most of the new LLM's have implemented tokens $\leq$ 2048 as shown in Table 7. However, GPT-4 is the exception with a limit of 8,192 tokens. This new capability is already being tested in text classification challenges in health care [71] and will likely expand to many other domains. Additional research into hierarchical attention [72] and long-document summarization [73] techniques are emerging as there are many practical applications that would benefit from going beyond words and sentences to long documents.

One of the sensational headlines and subsequent journal article titles was "GPT-4 Passes the Bar Exam" [74]. One major factor attributed to the improved results from GPT-3 to GPT-4 was the increased context window allowing for more accurate processing of long sequences of text.

Another extension of long text capability is in multimodal applications. The significant research in multimodal applications should also lead to new longer document capabilities for this application. One of the authors of this paper recently modified the Multimodal Transformers package to add the capability to utilize Longformer for text classification with text up to 4096 tokens for numeric/categorical data [75]. This resulted in a .9% increase in the F1 score on the Women's E-Commerce Clothing Review data set compared to using only 512 tokens. The authors plan an additional paper to publish these results.

VII. HOW ACCURATE?

By utilizing attention mechanisms and self-attention layers, transformers have demonstrated state-of-the-art accuracy across a variety of text classification tasks. Pre-trained transformers like BERT and GPT have become the standard approach for many general text classification tasks such as sentiment analysis, document classification, and named entity recognition. Transfer learning with transformers has also pushed the boundaries of accuracy by using pre-trained models that can be fine-tuned on smaller datasets with excellent results.

To understand the accuracy of different models on text classification tasks, a review of Papers With Code [76] provided a summary of the best models from 358 datasets used in 20 different NLP classification tasks (as shown in Table 1, 2). Papers With Code was chosen since it is the most complete source to date on benchmarks in NLP and other machine learning tasks, although it should be noted that this source is not comprehensive as described by [77]. The methodology to determine if the model is "transformer-based" was to review the abstracts and search for the keywords "transformer", "attention", "BERT", or "GPT". If a model used attention or was "transformer-like", it was classified as a transformer-based model in Table 10 below. Overall, the transformer based models are $\approx$ 68% of the "best models".

To simplify the resulting analysis, any data set that had only 1 Best Model by application category was not included in the results shown in Table 10. This reduced the number of datasets from 358 to 151 and applications from 20 to 15. One unexpected observation is that the new LLM's are almost exclusively the top models in the

TABLE 8. Current research on the efficiency of large language models.

Author	Date	Paper
Subramanyam et al	2021	*AMMUS: A Survey of Transformer-based Pretrained Models in Natural Language Processing* [79]
Rae et al	2021	*Scaling Language Models: Methods, Analysis & Insights from Training Gopher* [80]
Artetxe et al	2021	*Efficient Large Scale Language Modeling with Mixtures of Experts* [81]
Hoffman et al	2022	*Training Compute-Optimal Large Language Models* [82]
Borgeaud et al	2022	*Improving Language Models by Retrieving from Trillions of Tokens* [83]

question-answer application. For the other 14 applications, many of the best models are transformer-based but the results also include non-transformer models like XGBoost and Spark NLP. This could be because the new LLM's work best on question-answer tasks or because they are so new that tests have not been conducted on a wide range of applications. Table 10 can be used as a guide for selecting models based on accuracy, but there are no one-size-fits-all models. The best choice depends on the specific characteristics of your dataset, the problem at hand, and the resources available to you.

VIII. HOW EXPENSIVE?

The expense of transformers used for text classification varies depending on several factors such as hardware, software, cost of personnel, and the type of model. The cloud computing cost estimates for early transformer models like BERT and GPT-2 were $2074 to $43,008 [78]. The cost for newer LLMs is an open research question. When OpenAI's ChatGPT was asked the cost of GPT-3, the response was "...it is widely believed that the training cost for GPT-3 is in the range of tens of millions of dollars." Google's Bard was asked a similar question and the response was "The exact cost to train Bard is not publicly known, but it is estimated to be in the millions of dollars."

The cost to train these LLMs and the economic/environmental concerns has emerged as a significant issue in recent years. However, most of the research to alleviate this issue has been focused on the efficiency of these models as illustrated in Table 8.

Creative solutions to the economic and environmental issues remains an open challenge for organizations and researchers. It is our hope that more attention will be focused on how to improve the access and sustainability.

IX. HOW SAFE?

The increasing popularity of LLMs across several NLP tasks has also raised novel questions about the safety and ethics of use, including issues related to privacy and security, fairness and equity, and copyright.

A primary concern of research into the security of large language models has been memorization, or the model's ability to repeat content from training verbatim [84]. Reference [85] have shown memorization becomes more frequent as the size of the model increases, or with a higher number of duplicates, and they predict that memorization will become more prevalent as models continue to expand.

As a technical concern, memorization may also lead to downstream data contamination, or when the pretrained corpus contains some of the material from the test set. Multiple studies have shown this to be the case in major training datasets, meaning LLM performance on benchmarking tasks may be misleading [31], [86]. Reference [87] attempt to measure the impact of this contamination on classification tasks.

Arguably more problematically, memorization can also open these models to adversarial attacks. Reference [88] have shown the use of web-scraped data has led to mining private information from LLMs, with duplication in the training data making the model more vulnerable. Reference [84] demonstrated GPT-2's vulnerability to adversarial attack, noting that public, personal information, including names, phone numbers, and email addresses can be extracted verbatim from the language model, and also finding that larger models were more vulnerable than smaller models. Reference [89] similarly found LLMs risked leaking personal information, though they distinguish between memorization and association, arguing that models do not pose significant risk because of their low association, since private information will only be leaked randomly rather than extracted. As these models are frequently used in email, text, and code auto-completion, they also risk revealing personal data when fine-tuned for sensitive tasks. Reference [90] have explored how fine-tuning a model impacts the risks to privacy, observing that fine-tuning the head of the model leaves it most susceptible to extraction attacks.

On the other hand, [91] attempted an adversarial attack on BERT to mine patient names and conditions from clinical notes in pre-training and found that the model did not meaningfully associate names and conditions.

Even if the risk of specific data leakage is low, several scholars have begun calling for privacy-preserving LLMs, either by adapting the training to ensure privacy or removing sensitive information from the training text [92], [93], [94], [95], [96], [97].

The memorization of demographic features also raises concerns over potential bias in downstream tasks. Other scholars, interested in fairness and equity rather than the privacy and security of these LLMs, have explored the extent to which LLMs have learned protected demographic features in training [98], [99], [100] and, relatedly, how this learned bias might impact downstream classification tasks. Perhaps most research in this area has been devoted to inequities in how the language models handle gender. For example,

TABLE 9. Transparency of selected large language models released since ChatGPT (Nov 2022).

Name	Organization	Date	Open/Closed
Llama	Meta	Feb 2023	Closed
Bard(LaMBDA)	Bloomberg	Mar 2023	Closed
GPT-4	OpenAI	Mar 2023	Closed
BloombergGPT	Bloomberg	Mar 2023	Closed
Dolly 2	Databricks	Apr 2023	Open
StableLM	Stability AI	Apr 2023	Open (but uses Llama)
Titan	Amazon	Apr 2023	Closed
Bing Chat	Microsoft	Apr 2023	Closed
Llama 2	Meta	Jul 2023	Open

[100] found gender bias impacted classification results when working with medical text. Similarly, several scholars have shown that large language models tend to classify text written by women as more emotional [98], [101], [102], [103]. In addition to gender, other scholarship notes bias when dealing with race [104]. Reference [105] recently attempted to quantify the extent to which BERT has incorporated demographic information, using BERT outputs in a logistic regression model to predict sensitive features.

Because of these known biases embedded in many pre-trained language models, scholars have also explored ways of reducing or mitigating this bias. Generally, this takes place during fine-tuning. Reference [102], for example, developed a method for identifying and removing semantic features which contained sensitive information. Reference [100] incorporated loss during training to minimize bias learned during fine tuning. Others attempted to modify the data used for fine tuning, [106], [107], [108], removing traits that indicate gender [109], [110], or race [103]. A third method uses ensemble methods [104], and, most recently, [105] use active sampling of protected-attribute-uninformative data, which was then used to fine-tune the model, also reducing downstream inequalities.

This bias stemmed in part from the kind of data used when building the language model ([109], [111], [112]) and research has recently highlighted the fact that almost all training data for these large language models remains closed as another security concern. When training data is known, as [86] have shown with the C4 dataset, the filtering applied often disproportionately removes texts from and about minoritized groups. Reference [113] demonstrate that GPT-2 was more likely to detect negative sentiment in texts written in African American English; [114], investigating toxic content produced by LLMs, argue that the cause of this toxic content can be found in the training datasets, exploring two corpora used to train LLMs including GPT-2, which both contain toxic content. Reference [115] note that these learned biases also impact downstream Question and Answering tasks.

As language models have grown larger and more profitable, they have also increasingly become proprietary, making potential issues in the data regarding personal privacy or general equity more difficult to explore. Many of even the ‘open’ large language models, shown in Table 9, have not released information about their training. Reference [116] have recently shown the lack of openness and transparency in ChatGPT, Llama, and other large language models, providing structure for discussing degrees of openness, noting for example that even if a model can be used, many crucial aspects of the model remain closed, such as the training data, processes for instruction tuning, and the code used to train the model.

Because full details on training are not available, some have also accused these models of incorporating copyright materials, potentially violating intellectual property laws. Reference [117], for example, have very recently shown that both BERT and GPT-4 know a wide range of copyright material. Further, [118] have argued that training data must be made open to assess sources of bias, thus dovetailing the concerns over fairness and equity with concerns over intellectual property.

Relating to the broader concerns over bias and fairness, significant research has gone into incorporating eXplainable Artificial Intelligence (XAI) methods into LLMs, especially when used in downstream classification tasks.

For example, there have been various approaches to modifying or highlighting BERT’s architecture to make outputs explainable. ExBERT, for example, offers a dashboard that overviews the model’s attention and internal representation [119]. Reference [120] similarly developed VisBERT, which tracks tokens as they are processed by BERT. They extract hidden states from each transformer block and apply Principal Component Analysis to map tokens to a 2d state where distance represents semantic similarity. Transformers Interpret, a Python package, uses Integrated Gradients to determine and visualize the significance of words in any task done with pre-trained language models [121].

Another approach is to use post hoc, model agnostic explainability methods in classification. Reference [122] use LIME and Anchors to explain “Fake News” classifications made with BERT, highlighting words that have the highest contribution to the classification result. Reference [123] apply “explanations-by-example” to a BERT-based model, using the twin-systems approach (that is, pairing a white-box model, Case-based reasoning (CBR), with the black-box BERT model).

Finally, several researchers have incorporated XAI into the text classification task, even if the main focus of their work was not on explainability approaches. These approaches have been deployed in almost every category of the novel taxonomy we propose above. For example, [124] apply gradientSHAP to a multimodal model using BERT for emotion classification. In news classification, [125] have added LIME to a BERT-based classifier detecting misinformation about COVID-19, which shows the users how the decision was reached and which data sources were used to make the classification, extracting sentences from relative

TABLE 10. Best models by application on 151 datasets with Count of Best Model ≥ 2 for text classification tasks from Papers with Code (24 Jul 2023).

Application	Model	Count	Transformer-Based
Aspect-Based Sentiment Analysis	InstructABSA [128]	4	N
	HGCN+BERT [41]	4	Y
	MvP(multitask) [129]	3	Y
	BERT-pair-QA-B [130]	2	Y
Document Classification	MPAD-path [131]	3	Y
	ApproxRepSet [132]	3	N
	RMDL(30 RDLs) [133]	2	N
	BilBOWA [134]	2	N
	KD-LSTMreg [135]	2	Y
Multilabel Text Classification	XGBoost [136]	3	N
	LAHA [137]	3	Y
	BERT [138], [139]	2	Y
	MAGNET [140]	2	Y
Multimodal Sentiment Analysis	UniMSE [46]	2	Y
Named Entity Recognition	Spark NLP [141]	4	N
	ACE + document-context [142]	4	Y
	BLSTM-CNN-Char(SparkNLP) [141]	3	N
	DeepStruct multitask w/finetune [143]	3	Y
	HGN [144]	2	Y
	ConNER [145]	2	N
	Ours: cross-sentence ALB [146]	2	Y
Natural Language Inference	Human Benchmark [147]	3	Y
	ERNIE 2.0 Large [148]	2	Y
	PaLM 540B(finetuned) [149]	2	Y
	NeuralLog [150]	2	N
Question-Answer	Bing Chat [151]	8	Y
	FLAN 137B zero-shot [152]	4	Y
	ChatGPT [151], [153]	4	Y
	monoT5-3B [154]	3	Y
	Human Benchmark [155]	3	Y
	PaLM 540B(Self) [156]	2	Y
	PaLM 2-L(one-shot) [149] [157]	2	Y
	BioLinkBERT(large) [158]	2	Y
	XLNet(single-model) [30]	2	Y
	PaLM 540B (finetuned) [149]	2	Y
	BigBird-etc [69]	2	Y
	UnitedQA [159]	2	Y
	Ma et al.-ELECTRA [160]	2	Y
	Fast Weight Memory [161]	2	N
	Longformer [162]	2	Y
Sentence Classification	SciBERT(SciVocab [163]	2	Y

TABLE 10. (Continued.) Best models by application on 151 datasets with Count of Best Model ≥ 2 for text classification tasks from Papers with Code (24 Jul 2023).

Application	Model	Count	Transformer-Based
Sentiment Analysis	XLNet [30]	3	Y
	AraBERTv1 [164]	3	Y
	RoBERTA-wwm-ext-large [165]	2	Y
	CNN-LSTM [166]	2	N
	xlmindic-base-uniscript [167]	2	Y
	BERT large [168]	2	Y
	LSTMs+CNNs ensemble [169]	2	N
	FinBERT [170]	2	Y
Syntactic Parsing	ACE [142]	4	Y
Topic Labeling	JoSH [171]	2	N
Fake News Detection	SEMI-FND [47]	2	Y
Cross-Lingual Document Classification	XLMft UDA [172]	5	N
	Biinclusion [173]	2	N
	MultiFiT, pseudo [174]	2	N
Stance Detection	MUSE + UMAP [175]	2	N

news articles to explain the classification. Similar approaches have been employed in the medical field [126] and analysis of online reviews [127].

X. CONCLUSION AND FURTHER RESEARCH

In conclusion, this paper has provided a unique exploration into the use of text classification with transformer-based models. The transformative impact of transformers on NLP tasks, particularly text classification, is evident through their ability to capture complex contextual relationships and semantic nuances. Throughout this study, we have examined the questions of How wide? How large? How long? How accurate? How expensive? and How safe? as it applies to this new technology.

The comparative analysis conducted on various transformer models across diverse datasets underscores their remarkable performance and versatility across many but not all text classification tasks. The current hype related to the use of LLMs for question-answering chat bots is understandable as this capability moves AI closer to human-level performance on this task. However, this is only 1 of 20 applications reviewed in this paper, and similar gains are likely possible in other areas.

We anticipate that the substantial volume of research, investment, and significant financial commitments directed towards multimodal applications will enhance the proficiency across various domains of text classification.

However, challenges such as computational requirements, model size, and potential biases present in pre-trained representations call for continued research and innovation. Efforts to address these challenges, alongside emerging techniques for model interpret-ability and explain-ability, pave the way for more responsible and ethically sound applications of transformer-based text classification.

As researchers and practitioners alike, we must strive to harness the full potential of transformers while staying vigilant to ethical considerations and the broader societal impact of our work. The insights gained from this paper serve as a foundation for future advancements in the field, guiding us towards a deeper understanding of transformer-based text classification and its implications for the ever-evolving landscape of natural language processing.

In the future, potential areas for further research related to transformer-based text classification include:

- Multimodal classification using text and columnar/numeric data.
- Multiclass classification applications.
- Hierarchical classification for multilabel and topic modeling.
- Changes over time (drift).
- Cost and access issues.
- Legal issues related to copyrighted data used in training.

REFERENCES

[1] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning-based text classification: A comprehensive review," *ACM Comput. Surv.*, vol. 54, no. 3, pp. 1–40, Apr. 2021.

[2] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, and D. Roth, "Recent advances in natural language processing via large pre-trained language models: A survey," *ACM Comput. Surv.*, vol. 56, no. 2, pp. 1–40, Sep. 2023.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5999–6009.

[4] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. *Improving Language Understanding by Generative Pre-Training*. Accessed: Jun. 29, 2023. [Online]. Available: https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[6] S. Keele, "Guidelines for performing systematic literature reviews in software engineering," Ver. 2.3, EBSE, Tech. Rep., 2007, vol. 5.

[7] *Connected Papers*. Accessed: Sep. 18, 2023. [Online]. Available: https://www.connectedpapers.com

[8] P. Studtmann, "Aristotle's categories," in *The Stanford Encyclopedia Philosophy*, E. N. Zalta, Ed. Stanford, CA, USA: Stanford Univ.-Metaphysics Research Lab, 2021.

[9] M. E. Maron, "Automatic indexing: An experimental inquiry," *J. ACM*, vol. 8, no. 3, pp. 404–417, Jul. 1961.

[10] T. Poibeau, "The 1966 ALPAC report and its consequences," in *Machine Translation*. Cambridge, MA, USA: MIT Press, 2017, pp. 75–89.

[11] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proc. Conf. Empirical Methods Natural Lang. Process., Syst. Demonstration*, 2020, pp. 38–45.

[12] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P. S. Yu, and L. He, "A survey on text classification: From shallow to deep learning," 2020, *arXiv:2008.00364*.

[13] A. Gasparetto, M. Marcuzzo, A. Zangari, and A. Albarelli, "A survey on text classification algorithms: From text to predictions," *Information*, vol. 13, no. 2, p. 83, Feb. 2022.

[14] D. Kiela, M. Bartolo, Y. Nie, D. Kaushik, A. Geiger, Z. Wu, B. Vidgen, G. Prasad, A. Singh, P. Ringshia, Z. Ma, T. Thrush, S. Riedel, Z. Waseem, P. Stenetorp, R. Jia, M. Bansal, C. Potts, and A. Williams, "Dynabench: Rethinking benchmarking in NLP," 2021, *arXiv:2104.14337*.

[15] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, Z. Wu, L. Zhao, D. Zhu, X. Li, N. Qiang, D. Shen, T. Liu, and B. Ge, "Summary of ChatGPT-related research and perspective towards the future of large language models," 2023, *arXiv:2304.01852*.

[16] M. Munikar, S. Shakya, and A. Shrestha, "Fine-grained sentiment classification using BERT," 2019, *arXiv:1910.03474*.

[17] H. H. Do, P. Prasad, A. Maag, and A. Alsadoon, "Deep learning for aspect-based sentiment analysis: A comparative review," *Expert Syst. Appl.*, vol. 118, pp. 272–299, Mar. 2019.

[18] A. Elmadany, C. Zhang, M. Abdul-Mageed, and A. Hashemi, "Leveraging affective bidirectional transformers for offensive language detection," 2020, *arXiv:2006.01266*.

[19] J. Zhang, K. Hashimoto, Y. Wan, Z. Liu, Y. Liu, C. Xiong, and P. S. Yu, "Are pretrained transformers robust in intent classification? A missing ingredient in evaluation of out-of-scope intent detection," 2021, *arXiv:2106.04564*.

[20] M. Yasunaga, J. Leskovec, and P. Liang, "LinkBERT: Pretraining language models with document links," 2022, *arXiv:2203.15827*.

[21] H. Jwa, D. Oh, K. Park, J. Kang, and H. Lim, "ExBAKE: Automatic fake news detection model based on bidirectional encoder representations from transformers (BERT)," *Appl. Sci.*, vol. 9, no. 19, p. 4062, Sep. 2019.

[22] C. Wang and M. Banko, "Practical transformer-based multilingual text classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., Ind. Papers*, 2021, pp. 121–129.

[23] B. Zhang, D. Ding, and L. Jing, "How would stance detection techniques evolve after the launch of ChatGPT?" 2022, *arXiv:2212.14548*.

[24] C. M. Greco, A. Simeri, A. Tagarelli, and E. Zumpano, "Transformer-based language models for mental health issues: A survey," *Pattern Recognit. Lett.*, vol. 167, pp. 204–211, Mar. 2023.

[25] A. Cohan, I. Beltagy, D. King, B. Dalvi, and D. S. Weld, "Pretrained language models for sequential sentence classification," 2019, *arXiv:1909.04054*.

[26] W.-C. Chang, H.-F. Yu, K. Zhong, Y. Yang, and I. S. Dhillon, "Taming pretrained transformers for extreme multi-label text classification," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, Aug. 2020, pp. 3163–3171.

[27] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, p. 150, Apr. 2019.

[28] Q. Xipeng, S. TianXiang, X. Yige, S. Yunfan, D. Ning, and H. Xuanjing, "Pre-trained models for natural language processing: A survey," *Sci. China Technol. Sci.*, vol. 63, pp. 1872–1897, Sep. 2020.

[29] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[30] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 5754–5764.

[31] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, and A. Askell, "Language models are few-shot learners," in *Proc. Adv. Neur. Inf. Process. Sys.*, vol. 33, 2020, pp. 1877–1901.

[32] *Introducing LLaMA: A Foundational, 65-Billion-Parameter Language Model*. Accessed: Jul. 5, 2023. [Online]. Available: https://ai.facebook.com/blog/large-language-model-llama-meta-ai/

[33] S. Pichai. (Feb. 2023). *An Important Next Step on Our AI Journey*. Accessed: Jul. 5, 2023. [Online]. Available: https://blog.google/technology/ai/bard-google-ai-search-updates/

[34] OpenAI, "GPT-4 technical report," 2023, *arXiv:2303.08774*.

[35] Introducing BloombergGPT. (Mar. 2023). *Bloomberg's 50-Billion Parameter Large Language Model, Purpose-Built From Scratch for Finance*. Accessed: Jul. 5, 2023. [Online]. Available: https://www.bloomberg.com/company/press/bloomberggpt-50-billion-parameter-llm-tuned-finance/

[36] (Apr. 2023). *Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM*. Accessed: Jul. 5, 2023. [Online]. Available: https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm

[37] *StableLM: Stability AI Language Models*. Github. Accessed: Jul. 5, 2023. [Online]. Available: https://github.com/Stability-AI/StableLM

[38] *Amazon Titan*. Accessed: Jul. 5, 2023. [Online]. Available: https://aws.amazon.com/bedrock/titan/

[39] *Bing Chat*. Accessed: Jul. 5, 2023. [Online]. Available: https://www.microsoft.com/en-us/edge/features/bing-chat?form=MT00D8

[40] *Llama 2*. Accessed: Jul. 5, 2023. [Online]. Available: https://ai.meta.com/llama/

[41] L. Xu, X. Pang, J. Wu, M. Cai, and J. Peng, "Learn from structural scope: Improving aspect-level sentiment analysis with hybrid graph convolutional networks," *Neurocomputing*, vol. 518, pp. 373–383, Jan. 2023.

[42] O. Wallaart and F. Frasincar, *A Hybrid Approach for Aspect-Based Sentiment Analysis Using a Lexicalized Domain Ontology and Attentional Neural Models*. New York, NY, USA: Springer, 2019.

[43] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[44] B. Jin and D. Seetharaman, "Elon Musk creates new artificial intelligence company X.AI," *Wall St. J.*, Apr. 2023. [Online]. Available: https://www.wsj.com/articles/elon-musks-new-artificialintelligence-business-x-ai-incorporates-in-nevada-962c7c2f

[45] S. Pichai. (Apr. 2023). *Google DeepMind: Bringing Together Two World-Class AI Teams*. Accessed: Jul. 6, 2023. [Online]. Available: https://blog.google/technology/ai/april-ai-update/

[46] G. Hu, T.-E. Lin, Y. Zhao, G. Lu, Y. Wu, and Y. Li, "UniMSE: Towards unified multimodal sentiment analysis and emotion recognition," 2022, *arXiv:2211.11256*.

[47] P. Singh, R. Srivastava, K. P. S. Rana, and V. Kumar, "SEMI-FND: Stacked ensemble based multimodal inference for faster fake news detection," 2022, *arXiv:2205.08159*.

[48] *ChatGPT*. Accessed: Jul. 31, 2023. [Online]. Available: https://chat.openai.com

[49] *Try Bard, an AI Experiment by Google*. Accessed: Jul. 31, 2023. [Online]. Available: https://bard.google.com

[50] W. C. Sleeman IV, R. Kapoor, and P. Ghosh, "Multimodal classification: Current landscape, taxonomy and future directions," 2021, *arXiv:2109.09020*.

[51] K. Xu, M. Lam, J. Pang, X. Gao, C. Band, P. Mathur, F. Papay, A. K. Khanna, J. B. Cywinski, K. Maheshwari, P. Xie, and E. P. Xing, "Multimodal machine learning for automated ICD coding," in *Proc. 4th Mach. Learn. Healthcare Conf.*, vol. 106. Maastricht, The Netherlands, 2019, pp. 197–215.

[52] N. Holtz and J. M. Gomez, "Multimodal transformer for risk classification: Analyzing the impact of different data modalities," in *Natural Language Processing and Machine Learning*. Zürich, Switzerland: Academy and Industry Research Collaboration Center (AIRCC), May 2023.

[53] X. Wang, G. Chen, G. Qian, P. Gao, X.-Y. Wei, Y. Wang, Y. Tian, and W. Gao, "Large-scale multi-modal pre-trained models: A comprehensive survey," *Mach. Intell. Res.*, vol. 20, no. 4, pp. 447–482, Jun. 2023.

[54] G. Badaro, M. Saeed, and P. Papotti, "Transformers for tabular data representation: A survey of models and applications," *Trans. Assoc. Comput. Linguistics*, vol. 11, pp. 227–249, Mar. 2023.

[55] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep neural networks and tabular data: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, doi: 10.1109/TNNLS.2022.3229161.

[56] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 18932–18943.

[57] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Inf. Fusion*, vol. 81, pp. 84–90, May 2022.

[58] K. Gu and A. Budhkar, "A package for learning on tabular and text data with transformers," in *Proc. 3rd Workshop Multimodal Artif. Intell.*, 2021, pp. 69–73.

[59] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," 2020, *arXiv:2001.08361*.

[60] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[61] S. M. Jain, *Introduction to Transformers for NLP: With the Hugging Face Library and Models to Solve Problems*. New York, NY, USA: Apress, Oct. 2022.

[62] E. Bisong, "Google colaboratory," in *Building Machine Learning and Deep Learning Models on Google Cloud Platform*. Berkeley, CA, USA: Apress, 2019, pp. 59–64.

[63] D. Milmo, "ChatGPT reaches 100 million users two months after launch," Guardian, Feb. 2023. [Online]. Available: https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app

[64] N. Muennighoff, A. M. Rush, B. Barak, T. Le Scao, A. Piktus, N. Tazi, S. Pyysalo, T. Wolf, and C. Raffel, "Scaling data-constrained language models," 2023, *arXiv:2305.16264*.

[65] S. Biderman, H. Schoelkopf, Q. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. Aflah Khan, S. Purohit, U. Sai Prashanth, E. Raff, A. Skowron, L. Sutawika, and O. van der Wal, "Pythia: A suite for analyzing large language models across training and scaling," 2023, *arXiv:2304.01373*.

[66] Q. Fournier, G. M. Caron, and D. Aloise, "A practical survey on faster and lighter transformers," *ACM Comput. Surv.*, vol. 55, no. 14s, pp. 1–40, Dec. 2023.

[67] C. Li. (Jun. 2020). *OpenAI's GPT-3 Language Model: A Technical Overview*. Accessed: Jul. 17, 2023. [Online]. Available: https://lambdalabs.com/blog/demystifying-gpt-3

[68] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, B. Yin, and X. Hu, "Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond," 2023, *arXiv:2304.13712*.

[69] M. Zaheer, G. Guruganesh, A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, and A. Ahmed, "Big bird: Transformers for longer sequences," 2020, *arXiv:2007.14062*.

[70] I. Beltagy, M. E. Peters, and A. Cohan, "LongFormer: The long-document transformer," 2020, *arXiv:2004.05150*.

[71] Z. Liu, Y. Huang, X. Yu, L. Zhang, Z. Wu, C. Cao, H. Dai, L. Zhao, Y. Li, P. Shu, F. Zeng, L. Sun, W. Liu, D. Shen, Q. Li, T. Liu, D. Zhu, and X. Li, "DeID-GPT: Zero-shot medical text de-identification by GPT-4," 2023, *arXiv:2303.11032*.

[72] X. Zhang, F. Wei, and M. Zhou, "HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization," 2019, *arXiv:1905.06566*.

[73] X. Dai, I. Chalkidis, S. Darkner, and D. Elliott, "Revisiting transformer-based models for long document classification," 2022, *arXiv:2204.06683*.

[74] D. M. Katz, M. J. Bommarito, S. Gao, and P. Arredondo, "GPT-4 passes the bar exam," 2023, doi: 10.2139/ssrn.4389233.

[75] *Multimodal-Toolkit: Multimodal Model for Text and Tabular Data With Huggingface Transformers as Building Block for Text Data*. Github. Accessed: Jul. 13, 2023. [Online]. Available: https://github.com/georgian-io/Multimodal-Toolkit

[76] *Papers With Code—The Latest in Machine Learning*. Accessed: Jul. 27, 2023. [Online]. Available: http://paperswithcode.com

[77] F. Martínez-Plumed, P. Barredo, S. Ó. H. Éigeartaigh, and J. Hernández-Orallo, "Research community dynamics behind popular AI benchmarks," *Nature Mach. Intell.*, vol. 3, no. 7, pp. 581–589, May 2021.

[78] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," 2019, *arXiv:1906.02243*.

[79] K. Subramanyam Kalyan, A. Rajasekharan, and S. Sangeetha, "AMMUS : A survey of transformer-based pretrained models in natural language processing," 2021, *arXiv:2108.05542*.

[80] J. W. Rae et al., "Scaling language models: Methods, analysis & insights from training gopher," 2021, *arXiv:2112.11446*.

[81] M. Artetxe et al., "Efficient large scale language modeling with mixtures of experts," 2021, *arXiv:2112.10684*.

[82] J. Hoffmann et al., "Training compute-optimal large language models," 2022, *arXiv:2203.15556*.

[83] S. Borgeaud et al., "Improving language models by retrieving from trillions of tokens," in *Proc. 39th Int. Conf. Mach. Learn.*, vol. 162, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds. Maastricht, The Netherlands, 2022, pp. 2206–2240.

[84] N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, Ú. Erlingsson, A. Oprea, and C. Raffel, "Extracting training data from large language models," in *Proc. 30th USENIX Secur. Symp.*, 2021, pp. 2633–2650.

[85] N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramer, and C. Zhang, "Quantifying memorization across neural language models," 2022, *arXiv:2202.07646*.

[86] J. Dodge, M. Sap, A. Marasovic, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, and M. Gardner, "Documenting large webtext corpora: A case study on the colossal clean crawled corpus," 2021, *arXiv:2104.08758*.

[87] I. Magar and R. Schwartz, "Data contamination: From memorization to exploitation," 2022, *arXiv:2203.08242*.

[88] N. Kandpal, E. Wallace, and C. Raffel, "Deduplicating training data mitigates privacy risks in language models," in *Proc. 39th Int. Conf. Mach. Learn.*, vol. 162, K. Chaudhuri, S. Jegelka, Le Song, C. Szepesvari, G. Niu, and S. Sabato, Eds. Maastricht, The Netherlands, 2022, pp. 10697–10707.

[89] J. Huang, H. Shao, and K. C.-C. Chang, "Are large pre-trained language models leaking your personal information?" 2022, *arXiv:2205.12628*.

[90] F. Mireshghallah, A. Uniyal, T. Wang, D. Evans, and T. Berg-Kirkpatrick, "An empirical analysis of memorization in fine-tuned autoregressive language models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Abu Dhabi, United Arab Emirates, 2022, pp. 1816–1826.

[91] E. Lehman, S. Jain, K. Pichotta, Y. Goldberg, and B. C. Wallace, "Does BERT pretrained on clinical notes reveal sensitive data?" 2021, *arXiv:2104.07762*.

[92] R. Anil, B. Ghazi, V. Gupta, R. Kumar, and P. Manurangsi, "Large-scale differentially private BERT," 2021, *arXiv:2108.01624*.

[93] X. Li, F. Tramèr, P. Liang, and T. Hashimoto, "Large language models can be strong differentially private learners," 2021, *arXiv:2110.05679*.

[94] D. Yu, S. Naik, A. Backurs, S. Gopi, H. A. Inan, G. Kamath, J. Kulkarni, Y. T. Lee, A. Manoel, L. Wutschitz, S. Yekhanin, and H. Zhang, "Differentially private fine-tuning of language models," 2021, *arXiv:2110.06500*.

[95] W. Shi, A. Cui, E. Li, R. Jia, and Z. Yu, "Selective differential privacy for language modeling," 2021, *arXiv:2108.12944*.

[96] S. Hoory, A. Feder, A. Tendler, A. Cohen, S. Erell, I. Laish, H. Nakhost, U. Stemmer, A. Benjamini, A. Hassidim, and Y. Matias, "Learning and evaluating a differentially private pre-trained language model," in *Proc. 3rd Workshop Privacy Natural Lang. Process.*, Punta Cana, Dominican Republic, 2021, pp. 1178–1189.

[97] H. Brown, K. Lee, F. Mireshghallah, R. Shokri, and F. Tramèr, "What does it mean for a language model to preserve privacy?" in *Proc. ACM Conf. Fairness, Accountability, Transparency*, New York, NY, USA, Jun. 2022, pp. 2280–2292.

[98] X. Jin, F. Barbieri, B. Kennedy, A. M. Davani, L. Neves, and X. Ren, "On transferability of bias mitigation effects in language model fine-tuning," 2020, *arXiv:2010.12864*.

[99] K. Lu, P. Mardziel, F. Wu, P. Amancharla, and A. Datta, "Gender bias in neural natural language processing," in *Logic, Language, and Security*, V. Nigam, T. B. Kirigin, C. Talcott, J. Guttman, S.Kuznetsov, B. T. Loo, M. Okada, Eds. Cham, Switzerland: Springer, 2020, pp. 189–202.

[100] A. Silva, P. Tambwekar, and M. Gombolay, "Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2021, pp. 2383–2389.

[101] S. Touileb, L. Øvrelid, and E. Velldal, "Using gender- and polarity-informed models to investigate bias," in *Proc. 3rd Workshop Gender Bias Natural Lang. Process.*, 2021, pp. 66–74.

[102] R. Bhardwaj, N. Majumder, and S. Poria, "Investigating gender bias in BERT," *Cognit. Comput.*, vol. 13, no. 4, pp. 1008–1018, Jul. 2021.

[103] M. Mozafari, R. Farahbakhsh, and N. Crespi, "Hate speech detection and racial bias mitigation in social media based on BERT model," *PLoS ONE*, vol. 15, no. 8, Aug. 2020, Art. no. e0237861.

[104] M. Halevy, C. Harris, A. Bruckman, D. Yang, and A. Howard, "Mitigating racial biases in toxic language detection with an equity-based ensemble framework," in *Proc. EAAMO*, New York, NY, USA, Nov. 2021, pp. 1–11.

[105] L. Sha, Y. Li, D. Gasevic, and G. Chen, "Bigger data or fairer data? Augmenting BERT via active sampling for educational text classification," in *Proc. 29th Int. Conf. Comput. Linguistics*, Oct. 2022, pp. 1275–1285.

[106] F. Prost, N. Thain, and T. Bolukbasi, "Debiasing embeddings for reduced gender bias in text classification," 2019, *arXiv:1908.02810*.

[107] R. Islam, K. N. Keya, Z. Zeng, S. Pan, and J. Foulds, "Debiasing career recommendations with neural fair collaborative filtering," in *Proc. Web Conf.*, New York, NY, USA, Jun. 2021, pp. 3779–3790.

[108] Y. Pruksachatkun, S. Krishna, J. Dhamala, R. Gupta, and K.-W. Chang, "Does robustness improve fairness? Approaching fairness with word substitution robustness methods for text classification," 2021, *arXiv:2106.10826*.

[109] D. de Vassimon Manela, D. Errington, T. Fisher, B. van Breugel, and P. Minervini, "Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models," in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics, Main Volume*, 2021, pp. 2232–2242.

[110] J. R. Minot, N. Cheney, M. Maier, D. C. Elbers, C. M. Danforth, and P. S. Dodds, "Interpretable bias mitigation for textual data: Reducing gender bias in patient notes while maintaining classification performance," 2021, *arXiv:2103.05841*.

[111] L. Lucy and D. Bamman, "Gender and representation bias in GPT-3 generated stories," in *Proc. 3rd Workshop Narrative Understand.*, 2021, pp. 48–55.

[112] M. Nadeem, A. Bethke, and S. Reddy, "StereoSet: Measuring stereotypical bias in pretrained language models," 2020, *arXiv:2004.09456*.

[113] S. Groenwold, L. Ou, A. Parekh, S. Honnavalli, S. Levy, D. Mirza, and W. Y. Wang, "Investigating African–American vernacular English in transformer-based text generation," 2020, *arXiv:2010.02510*.

[114] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "RealToxicityPrompts: Evaluating neural toxic degeneration in language models," 2020, *arXiv:2009.11462*.

[115] T. Li, T. Khot, D. Khashabi, A. Sabharwal, and V. Srikumar, "UnQovering stereotyping biases via underspecified questions," 2020, *arXiv:2010.02428*.

[116] A. Liesenfeld, A. Lopez, and M. Dingemanse, "Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators," in *Proc. 5th Int. Conf. Conversational User Interface*, New York, NY, USA, Jul. 2023, pp. 1–6.

[117] K. K. Chang, M. Cramer, S. Soni, and D. Bamman, "Speak, memory: An archaeology of books known to ChatGPT/GPT-4," 2023, *arXiv:2305.00118*.

[118] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford, "Datasheets for datasets," *Commun. ACM*, vol. 64, no. 12, pp. 86–92, Nov. 2021.

[119] B. Hoover, H. Strobelt, and S. Gehrmann, "ExBERT: A visual analysis tool to explore learned representations in transformers models," 2019, *arXiv:1910.05276*.

[120] B. V. Aken, B. Winter, A. Löser, and F. A. Gers, "VisBERT: Hidden-state visualizations for transformers," in *Proc. Companion Web Conf.*, New York, NY, USA, Apr. 2020, pp. 207–211.

[121] C. Pierse. (Feb. 2021) *Introducing Transformers Interpret—Explainable AI for Transformers*. Accessed: Aug. 24, 2023. [Online]. Available: https://towardsdatascience.com/introducing-transformers-interpret-explainable-ai-for-transformers-890a403a9470

[122] M. Szczepanski, M. Pawlicki, R. Kozik, and M. Choras, "New explainability method for BERT-based model in fake news detection," *Sci. Rep.*, vol. 11, no. 1, p. 23705, Dec. 2021.

[123] E. M. Kenny and M. T. Keane, "Explaining deep learning using examples: Optimal feature weighting methods for twin systems using post-hoc, explanation-by-example in XAI," *Knowl.-Based Syst.*, vol. 233, Dec. 2021, Art. no. 107530.

[124] T. Shaikh, A. Khalane, R. Makwana, and A. Ullah, "Evaluating significant features in context-aware multimodal emotion recognition with XAI methods," Authorea Preprints, 2023. [Online]. Available: https://doi.org/10.22541/au.167407909.97031004

[125] L. Moe, A. Kundu, and U. T. Nguyen. *A BERT-Based Explainable System for COVID-19 Misinformation Identification*. Accessed: Oct. 5, 2023. [Online]. Available: https://workshop-proceedings.icwsm.org/pdf/2023_46.pdf

[126] M. T. Rietberg, V. B. Nguyen, J. Geerdink, O. Vijlbrief, and C. Seifert, "Accurate and reliable classification of unstructured reports on their diagnostic goal using BERT models," *Diagnostics*, vol. 13, no. 7, p. 1251, Mar. 2023.

[127] F. Ahmed, S. Sultana, M. T. Reza, S. K. S. Joy, and Md. G. R. Alam, "Interpretable movie review analysis using machine learning and transformer models leveraging XAI," in *Proc. IEEE Asia–Pacific Conf. Comput. Sci. Data Eng. (CSDE)*, Dec. 2022, pp. 1–6.

[128] K. Scaria, H. Gupta, S. Goyal, S. Arjun Sawant, S. Mishra, and C. Baral, "InstructABSA: Instruction learning for aspect based sentiment analysis," 2023, *arXiv:2302.08624*.

[129] T. Tang, J. Li, W. Xin Zhao, and J.-R. Wen, "MVP: Multi-task supervised pre-training for natural language generation," 2022, *arXiv:2206.12131*.

[130] C. Sun, L. Huang, and X. Qiu, "Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence," 2019, *arXiv:1903.09588*.

[131] G. Nikolentzos, A. J. -P. Tixier, and M. Vazirgiannis, "Message passing attention networks for document understanding," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 5754–5764.

[132] K. Skianis, G. Nikolentzos, S. Limnios, and M. Vazirgiannis, "Rep the set: Neural networks for learning set representations," in *Proc. 23rd Int. Conf. Artif. Intell. Statist.*, vol. 108, S. Chiappa and R. Calandra, Eds. Maastricht, Netherlands, 2020, pp. 1410–1420.

[133] K. Kowsari, M. Heidarysafa, D. E. Brown, K. J. Meimandi, and L. E. Barnes, "RMDL: Random multimodel deep learning for classification," in *Proc. 2nd Int. Conf. Inf. Syst. Data Mining (ICISDM)*, Apr. 2018, pp. 19–28.

[134] S. Gouws, Y. Bengio, and G. Corrado, "BilBOWA: Fast bilingual distributed representations without word alignments," in *Proc. 32nd Int. Conf. Int. Conf. Mach. Learn.*, vol. 37, 2015, pp. 748–756.

[135] A. Adhikari, A. Ram, R. Tang, and J. Lin, "DocBERT: BERT for document classification," 2019, *arXiv:1904.08398*.

[136] P. H. L. de Araujo, T. E. de Campos, F. A. Braz, and N. C. da Silva, "VICTOR: A dataset for Brazilian legal documents classification," in *Proc. 12th Lang. Resour. Eval. Conf.*, Marseille, France, May 2020, pp. 1449–1458.

[137] X. Huang, B. Chen, L. Xiao, J. Yu, and L. Jing, "Label-aware document representation via hybrid attention for extreme multi-label text classification," *Neural Process. Lett.*, vol. 54, no. 5, pp. 3601–3617, Oct. 2022.

[138] J.-S. Lee and J. Hsiang, "PatentBERT: Patent classification with fine-tuning a pre-trained BERT model," 2019, *arXiv:1906.02124*.

[139] R.-C. Chang, C.-M. Lai, K.-L. Chang, and C.-H. Lin, "Dataset of propaganda techniques of the state-sponsored information operation of the people's republic of China," 2021, *arXiv:2106.07544*.

[140] A. Pal, M. Selvakumar, and M. Sankarasubbu, "Multi-label text classification using attention-based graph neural network," 2020, *arXiv:2003.11644*.

[141] V. Kocaman and D. Talby, "Biomedical named entity recognition at scale," in *Proc. Int. Conf. Pattern Recognit.* New York, NY, USA: Springer, 2021, pp. 635–646.

[142] X. Wang, Y. Jiang, N. Bach, T. Wang, Z. Huang, F. Huang, and K. Tu, "Automated concatenation of embeddings for structured prediction," 2020, *arXiv:2010.05006*.

[143] C. Wang, X. Liu, Z. Chen, H. Hong, J. Tang, and D. Song, "DeepStruct: Pretraining of language models for structure prediction," 2022, *arXiv:2205.10475*.

[144] J. Hu, Y. Shen, Y. Liu, X. Wan, and T.-H. Chang, "Hero-gang neural model for named entity recognition," 2022, *arXiv:2205.07177*.

[145] M. Jeong and J. Kang, "Enhancing label consistency on document-level named entity recognition," 2022, *arXiv:2210.12949*.

[146] Z. Zhong and D. Chen, "A frustratingly easy approach for entity and relation extraction," 2020, *arXiv:2010.12812*.

[147] T. Shavrina, A. Fenogenova, A. Emelyanov, D. Shevelev, E. Artemova, V. Malykh, V. Mikhailov, M. Tikhonova, A. Chertok, and A. Evlampiev, "RussianSuperGLUE: A Russian language understanding evaluation benchmark," 2020, *arXiv:2010.15925*.

[148] Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, and H. Wang, "ERNIE 2.0: A continual pre-training framework for language understanding," in *Proc. AAAI*, Apr. 2020, vol. 34, no. 5, pp. 8968–8975.

[149] A. Chowdhery et al., "PaLM: Scaling language modeling with pathways," 2022, *arXiv:2204.02311*.

[150] Z. Chen, Q. Gao, and L. S. Moss, "NeuralLog: Natural language inference with joint neural and logical reasoning," 2021, *arXiv:2105.14167*.

[151] X.-Q. Dao, N.-B. Le, T.-D. Vo, X.-D. Phan, B.-B. Ngo, V.-T. Nguyen, T.-M.-M. Nguyen, and H.-P. Nguyen, "VNHSGE: Vietnamese high school graduation examination dataset for large language models," 2023, *arXiv:2305.12199*.

[152] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. Wei Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," 2021, *arXiv:2109.01652*.

[153] Y. Tan, D. Min, Y. Li, W. Li, N. Hu, Y. Chen, and G. Qi, "Can ChatGPT replace traditional KBQA models? An in-depth analysis of the question answering performance of the GPT LLM family," 2023, *arXiv:2303.07992*.

[154] G. Moraes Rosa, L. Bonifacio, V. Jeronymo, H. Abonizio, M. Fadaee, R. Lotufo, and R. Nogueira, "No parameter left behind: How distillation and model size affect zero-shot retrieval," 2022, *arXiv:2206.02873*.

[155] E. Taktasheva, T. Shavrina, A. Fenogenova, D. Shevelev, N. Katricheva, M. Tikhonova, A. Akhmetgareeva, O. Zinkevich, A. Bashmakova, S. Iordanskaia, A. Spiridonova, V. Kurenshchikova, E. Artemova, and V. Mikhailov, "TAPE: Assessing few-shot Russian language understanding," 2022, *arXiv:2210.12813*.

[156] J. Huang, S. Shane Gu, L. Hou, Y. Wu, X. Wang, H. Yu, and J. Han, "Large language models can self-improve," 2022, *arXiv:2210.11610*.

[157] R. Anil et al., "PaLM 2 technical report," 2023, *arXiv:2305.10403*.

[158] M. Yasunaga, J. Leskovec, and P. Liang, "LinkBERT: Pretraining language models with document links," in *Proc. ICML*, 2022, pp. 8003–8016.

[159] H. Cheng, Y. Shen, X. Liu, P. He, W. Chen, and J. Gao, "UnitedQA: A hybrid approach for open domain question answering," 2021, *arXiv:2101.00178*.

[160] X. Ma, Z. Zhang, and H. Zhao, "Enhanced speaker-aware multi-party multi-turn dialogue comprehension," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 2410–2423, 2023.

[161] I. Schlag, T. Munkhdalai, and J. Schmidhuber, "Learning associative inference using fast weight memory," 2020, *arXiv:2011.07831*.

[162] G. T. Hudson and N. A. Moubayed, "MuLD: The multitask long document benchmark," 2022, *arXiv:2202.07362*.

[163] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," 2019, *arXiv:1903.10676*.

[164] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based model for Arabic language understanding," 2020, *arXiv:2003.00104*.

[165] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, "Pre-training with whole word masking for Chinese BERT," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 3504–3514, 2021.

[166] I. Abu Farha and W. Magdy, "Mazajak: An online Arabic sentiment analyser," in *Proc. 4th Arabic Natural Lang. Process. Workshop*, Stroudsburg, PA, USA, 2019, pp. 192–198.

[167] I. M. Moosa, M. E. Akhter, and A. B. Habib, "Does transliteration help multilingual language modeling?" 2022, *arXiv:2201.12501*.

[168] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le, "Unsupervised data augmentation for consistency training," in *Proc. NIPS*, vol. 33, 2020, pp. 6256–6268.

[169] M. Cliche, "BB_twtr at SemEval-2017 task 4: Twitter sentiment analysis with CNNs and LSTMs," 2017, *arXiv:1704.06125*.

[170] D. Araci, "FinBERT: Financial sentiment analysis with pre-trained language models," 2019, *arXiv:1908.10063*.

[171] Y. Meng, Y. Zhang, J. Huang, Y. Zhang, C. Zhang, and J. Han, "Hierarchical topic mining via joint spherical tree and text embedding," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, Aug. 2020, pp. 1908–1917.

[172] G. Lai, B. Oguz, Y. Yang, and V. Stoyanov, "Bridging the domain gap in cross-lingual document classification," 2019, *arXiv:1909.07009*.

[173] H. Soyer, P. Stenetorp, and A. Aizawa, "Leveraging monolingual data for crosslingual compositional word representations," 2014, *arXiv:1412.6334*.

[174] J. Martin Eisenschlos, S. Ruder, P. Czapla, M. Kardas, S. Gugger, and J. Howard, "MultiFiT: Efficient multi-lingual language model finetuning," 2019, *arXiv:1909.04761*.

[175] A. Rashed, M. Kutlu, K. Darwish, T. Elsayed, and C. Bayrak, "Embeddings-based clustering for target specific stances: The case of a polarized Turkey," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 15, May 2021, pp. 537–548.

JOHN FIELDS (Graduate Student Member, IEEE) received the B.S. degree in engineering (industrial distribution) from Texas A&M University and the M.S. degree in applied data science from Syracuse University. He is currently pursuing the Ph.D. degree in computer science with Marquette University, Milwaukee, WI, USA.

He is an Assistant Professor of business analytics with Concordia University Wisconsin-Ann Arbor. He has collaborated as the coauthor on an upcoming paper scheduled for presentation at the IEEE Big Data 2023 Conference, delving into the application of natural language processing (NLP) in the realm of education. He is the co-inventor on a pending patent (62/935,928) that explores the utilization of machine learning and AI for higher education applications. His research interests include the integration of artificial intelligence (AI) in education and its impact on student success, with a specific interest in text classification (including multimodal approaches), graph databases, and addressing AI bias and fairness.

Mr. Fields was a recipient of the 2015 SAP IGgie Award for information governance.

KEVIN CHOVANEC is a data scientist in the Office of Institutional Research at Marquette University in Milwaukee, WI, USA. He received his B.S. in Math and English from Marquette, an M.A. from the University of Chicago, and a Ph.D. in English and Comparative Literature from the University of North Carolina. Currently, he is pursuing a Ph.D. in Computer Science at Marquette. His research embraces an interdisciplinary approach, focusing on the digital humanities, editing, natural language processing, fairness and accountability in AI, and educational analytics, and his work has appeared in journals and conference proceedings across academic disciplines, including Digital Humanities Quarterly, Renaissance Drama, and a forthcoming paper in IEEE Big Data 2023.

PRAVEEN MADIRAJU received the Ph.D. degree in computer science from Georgia State University.

He directs the Data Science and Text Analytics Laboratory. The Data Laboratory focuses on solving real-world problems by applying techniques from the broad area of data science and data analytics on both structured and unstructured data. The laboratory also conducts research on applying machine learning techniques to analyze textual and social media data. He is currently a Professor with the Department of Computer Science, Marquette University, Milwaukee, WI, USA. He is also the Graduate Program Chair of the Computer Science Program, Marquette University. He has published over 50 peer-reviewed articles and has organized workshops on middleware systems in conjunction with ACM SAC and IEEE COMPSAC. His research interests include data science, healthcare informatics, text analytics, and databases. He regularly serves on NSF panels.