Harnessing the Power of Federated Learning: A Systematic Review of Light Weight Deep Learning Protocols

Haseeb Khan Shinwari[1], Riaz UlAmin[2]
Newton AI Research Lab, Pakistan[1]
Edinburgh Napier University, UK and University of Okara, Pakistan[2]

***Abstract*—With rapid proliferation in using smart devices, real time efficient sentiment analysis has gained considerable popularity. These devices generate variety of data. However, for resource constrained devices to perform sentiment analysis over multimodal data using conventional modals that are computationally complex and resource hungry, is challenging. This challenge may be addressed using a light weight but efficient modal specifically focused on sentiment analysis for contrained devices. in the literature, there are several modals that claims to be light weight however, the real sense and logic to determine if the modal may be termed as lightweight still requires further research. This paper reviews approaches to federated learning for multimodal sentiment analysis. Federated learning enables decentralized training without sharing data. Considering the review need to balance privacy concerns, performance, and resource usage, the review evaluates existing approaches to enhance accuracy in sentiment classification. The review identifies strengths and limitations in handling multimodal data. The search focused on studies in databases like IEEE Xplore and Scopus. Studies published in peer-reviewed journals over the past five years were included. The review covers 45 studies, mostly experimental, with some theoretical models. Key results show lightweight protocols improve efficiency and privacy in federated learning. They reduce computational demands while handling text, image, and audio data. There is a growing focus on resource-constrained devices in research. Trade-offs between model complexity and speed are commonly explored. The review addresses how these protocols balance accuracy and computational cost.**

***Keywords*—Light weight protocols; sentiment analysis; federated learning; deep learning**

I. Introduction

Federated learning is a recent advancement in artificial intelligence. It enables decentralized model training without sharing raw data [1]. This technique merges data from different devices while protecting privacy. The method's popularity has grown due to rising privacy concerns [2]. Unlike standard machine learning, data remains on each device. Only model updates are sent to a central server. This reduces the risk of data breaches. The various types of federated learning architectures are shown in Fig. 1 The classification of Federated learning is presented in [3]. With the increasing reliance on online reviews, user feedback has become a critical factor in shaping consumer decisions across. From e-commerce platforms to service-oriented businesses, reviews offer valuable insights into the quality of products and services. However, not all reviews are created equal, and their emotional tone plays a significant role in conveying the authenticity and impact of the user experience. Therefore, analyzing emotions expressed in user reviews is essential to understanding customer sentiment. Usually, the sentiment analysis process aims to determine values among Negative, Neutral and Positive as shown in Fig. 2. Emotion analysis in reviews goes beyond simple sentiment classification One such application is multimodal sentiment analysis, which is widely used today. Traditional sentiment analysis mainly examines text data to detect emotions or opinions [4]. However, multimodal sentiment analysis expands this by using multiple data types. It incorporates text, images, and audio for a richer analysis. Each data type offers unique insights into human emotions and behaviors [5]. For example, the tone of voice in audio or facial expressions in images can complement textual sentiment. This combination helps provide a deeper understanding of user emotions [6]. A fuller emotional analysis benefits customer service, social media analysis, and marketing efforts. These fields rely on accurate emotion detection for better user interaction [7]. General workflow of deep learning protocol is shown in Fig. 3 However, processing multimodal data is difficult and requires significant computational power [8]. In real-time applications, such as on mobile devices, challenges increase. Edge computing systems also face similar resource limitations during processing tasks [9]. This is where lightweight deep learning protocols become essential. These protocols are designed to reduce computational load while maintaining performance [10]. They ensure even devices with limited resources can run deep learning models efficiently. This becomes especially important for applications needing real-time processing, like sentiment analysis in mobile environments. Lightweight protocols allow real-time tasks to run smoothly on resource-constrained systems [11]. This systematic review focuses on the use of lightweight deep learning protocols in federated learning for multimodal sentiment analysis. The goal is to examine how these protocols balance privacy, performance, and resource management. Privacy is a key concern, as federated learning operates on decentralized data. Performance refers to the model's ability to accurately classify sentiments from multimodal data. Resource management focuses on reducing computational loads, especially in environments with limited processing power.

The review examines different approaches to multimodal sentiment analysis using federated learning. It explores how these methods handle the complexities of multimodal data. Text, image, and audio data each need distinct processing techniques [2]. Text data is often processed using natural

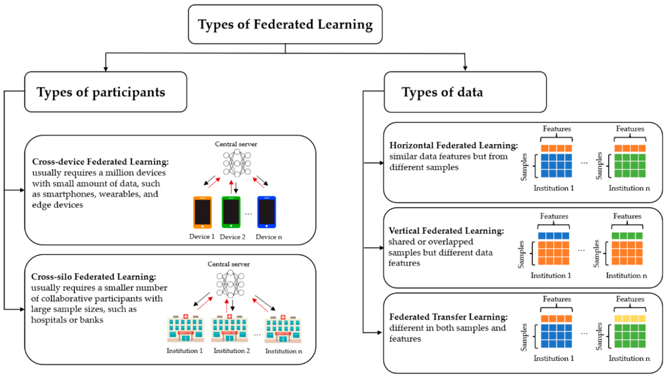
Types of Federated Learning
Types of participants
Types of data
Cross-device Federated Learning: usually requires a million devices with small amount of data, such as smartphones, wearables, and edge devices
Central server
Device 1
Device 2
Device n
Cross-silo Federated Learning: usually requires a smaller number of collaborative participants with large sample sizes, such as hospitals or banks
Central server
Institution 1
Institution 2
Institution n
Horizontal Federated Learning: similar data features but from different samples
Vertical Federated Learning: shared or overlapped samples but different data features
Federated Transfer Learning: different in both samples and features
Features
Samples
Institution 1
Institution n

Fig. 1. Types of federated learning.

language processing (NLP) techniques. Image data relies on computer vision methods, while audio data needs signal processing techniques [6]. Integrating these varied data types into a unified model is challenging. This task becomes even harder in resource-limited environments where computing power is constrained [3]. Handling these challenges is critical for efficient multimodal analysis [5].

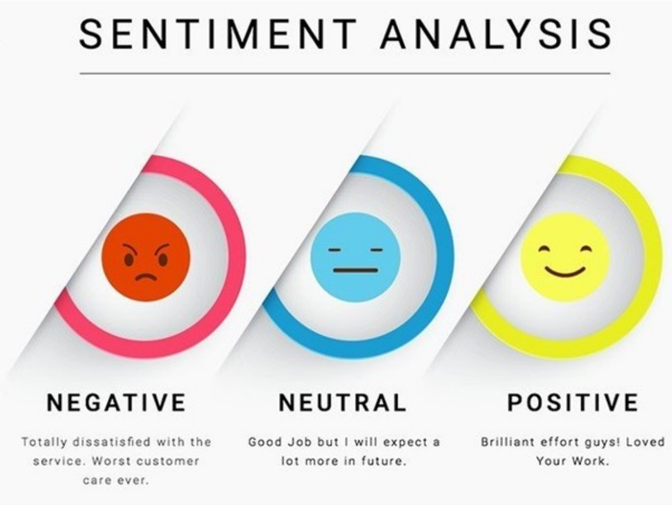
SENTIMENT ANALYSIS
NEGATIVE
Totally dissatisfied with the service. Worst customer care ever.
NEUTRAL
Good Job but I will expect a lot more in future.
POSITIVE
Brilliant effort guys! Loved Your Work.

Fig. 2. Types of sentiment analysis.

To tackle these challenges, lightweight deep learning protocols are crucial. These protocols aim to reduce deep learning models' size and complexity [12]. Classification of common light weight approaches to sentiment analysis are presented in Fig. 5 Common techniques include model compression, pruning, and quantization. Compression shrinks the model, making it easier to store and process. Pruning eliminates unneeded parts of the model, improving efficiency. Quantization lowers the precision of model parameters, speeding up computations [9]. This reduces resource use without greatly impacting performance. Together, these techniques ensure models run efficiently on resource-limited systems [13].

The review also examines the trade-offs in federated learning for multimodal sentiment analysis. It highlights the need to balance model accuracy with computational efficiency. More complex models often provide higher accuracy but need more resources [3]. In contrast, simpler models run faster but may lack the same accuracy. Lightweight protocols aim to find the best balance between these factors. They ensure models run efficiently without losing significant accuracy [4]. Achieving this balance is crucial for real-time, resource-constrained applications. Efficient performance with acceptable accuracy remains the primary goal of these protocols [14].

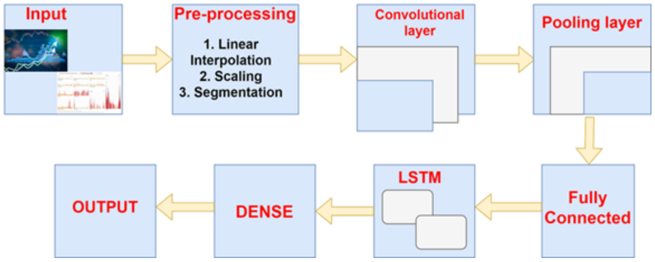

Fig. 3. Workflow in deep learning protocol.

Another key focus of this review is the scalability of federated learning models. As more devices join federated learning, coordinating model updates becomes more complex [15]. Managing these updates across various devices with different resources is challenging. Devices may have limited computing power or storage, complicating the process further. Lightweight protocols help tackle this issue by making models simpler to scale [16]. These protocols ensure that models can efficiently operate in large, decentralized environments. Scaling federated learning models becomes more manageable with reduced computational demands. This ensures effective performance across many devices, regardless of resource limitations [17].

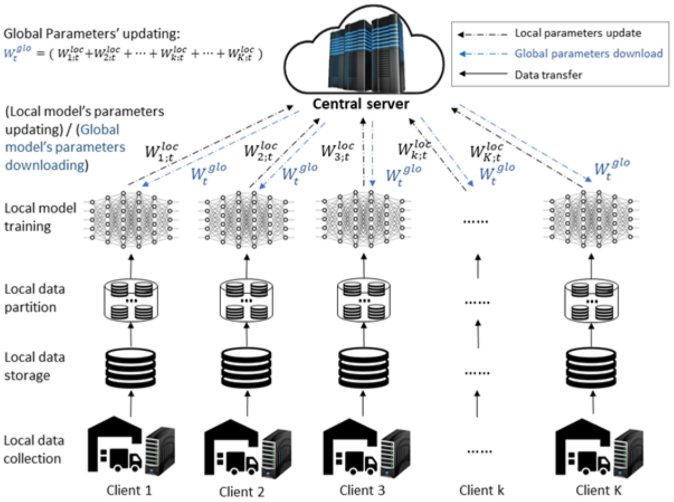

Fig. 4. The framework of federated learning.

II. Major Contributions

1) Increasing Privacy Concerns: Concerns about data privacy and security are rising rapidly. Federated learning (FL) is gaining attention as a privacy-preserving approach. It ensures privacy by keeping data on individual devices. Multimodal sentiment analysis uses sensitive data like text, audio, and images. This requires strong privacy protection. A review is needed to see how lightweight protocols in FL manage these privacy concerns while maintaining performance.

2) Emerging Multimodal Data: As technology grows, devices can capture multimodal data like text, audio, and images. Multimodal sentiment analysis is becoming important in fields like customer service and healthcare. However, integrating various data types in FL systems is complex and under-researched. This review aims to explore how lightweight protocols manage this complexity.

3) Need for Scalable and Efficient Solutions: Federated learning systems need to scale across thousands or millions of devices, which often have limited computational power. Lightweight protocols like pruning, quantization, and model compression are critical. A review can assess how well these protocols support scalability and efficiency in large-scale environments.

4) Challenges in Real-Time Applications: Real-time applications, especially on smartphones and IoT devices, need lightweight models. Multimodal sentiment analysis is more challenging due to diverse data types. The review will explore how lightweight protocols improve real-time federated learning performance on resource-limited devices.

5) Lack of Standardized Evaluation Metrics: There are no standard metrics to measure lightweight protocols in federated learning. This is especially true for multimodal sentiment analysis. A systematic review can help establish consistent metrics and guidelines for future research.

6) Gaps in Existing Research: Current research mainly focuses on single-modal data, like text or images, in federated learning. Research on multimodal integration is limited. Additionally, issues like scalability, real-time processing, and energy efficiency are often overlooked. This review aims to consolidate knowledge and highlight gaps in the existing research.

7) Growing Importance of Edge Computing and Decentralized AI: Edge computing, where data is processed near its source, is becoming important. Federated learning fits well with this decentralized AI approach. The framework of federated learning is shown in Fig. 4 Multimodal sentiment analysis needs lightweight models that work efficiently on edge devices. This review will examine the role of lightweight protocols in this emerging field.

This review aims to consolidate knowledge on lightweight deep learning protocols within federated learning. Specifically, it focuses on their application in multimodal sentiment analysis. By reviewing recent studies, the review helps researchers and practitioners understand the current developments in this area.

III. Literature Review

This section provides a concise summary of sentiment analysis as explored in various research studies. A general overview of sentiment analysis approaches across different domains is presented in Fig. 5. Sentiment analysis has evolved from early lexicon-based methods and traditional machine learning to advanced deep learning and lightweight approaches, particularly suited for Federated Learning (FL). Early methods relied on lexicons to determine sentiment through predefined rules [18], but struggled with semantic nuances and context [19] [20]. Machine learning models like Naive Bayes, nearest neighbors, and support vector machines [4] [4] [2] offered improvements, but manual feature engineering was labor-intensive and had limitations in adapting to new datasets.

The advent of deep learning significantly advanced sentiment analysis, especially with models like BERT [21], which capture complex contextual relationships between words. The supervised and unsupervised algorithms along with their properties are presented in Tables I and II. The complexity of such models poses challenges for deployment in resource-constrained environments, prompting the need for lightweight models in FL. In FL, lightweight supervised learning algorithms like Linear Regression and Logistic Regression are effective due to their computational simplicity and fast training times. However, they struggle with non-linear data [22]. Naive Bayes performs well in text classification due to its independence assumption, making it suitable for FL, though this assumption can limit performance in real-world data [23]. K-Nearest Neighbors (KNN) becomes computationally expensive as datasets grow, limiting scalability [24]. Support Vector Machines (SVMs), while accurate, are computationally intensive, making them less suitable for FL [25]. Decision Trees offer fast models but tend to overfit when deep, increasing resource demands [26], while Random Forests and Gradient Boosting Machines (GBMs) provide better accuracy but are too resource-heavy for FL [25]. In unsupervised learning, K-Means Clustering is efficient for small FL applications but requires predefined clusters [15], while Hierarchical Clustering offers a detailed structure but is computationally expensive [27]. Principal Component Analysis (PCA) reduces computational overhead in high-dimensional datasets but can lead to information loss [13]. Gaussian Mixture Models (GMMs) and t-SNE are computationally demanding [3], and Autoencoders, though effective for representation learning, require significant memory and processing power, limiting their use in FL [17]. Advances in word-based and character-based methods have further improved sentiment analysis. Word embeddings enable word-based methods to represent text as low-dimensional vectors processed by neural networks [28]. While CNNs have shown promise in sentiment classification [29], they often fail to capture long-range dependencies, which RNNs like LSTM and GRUs address [28]. Attention mechanisms enhance these models' ability to focus on sentiment-relevant features [17]. Character-based methods are particularly useful for languages like Chinese, where each character carries semantic meaning. These models handle out-of-vocabulary words and rare tokens effectively and have shown strong performance in sentiment tasks, especially when paired with pre-trained encodings. Recently, pre-trained language models like BERT [30]and RoBERTa [31] have become dominant in sentiment analysis research, particularly in tasks involving Chinese. ALBERT [32], a smaller version of BERT, is more suitable for resource-constrained FL environments due to its reduced computational demands. Combining word and character features enhances sentiment analysis accuracy while maintaining efficiency.

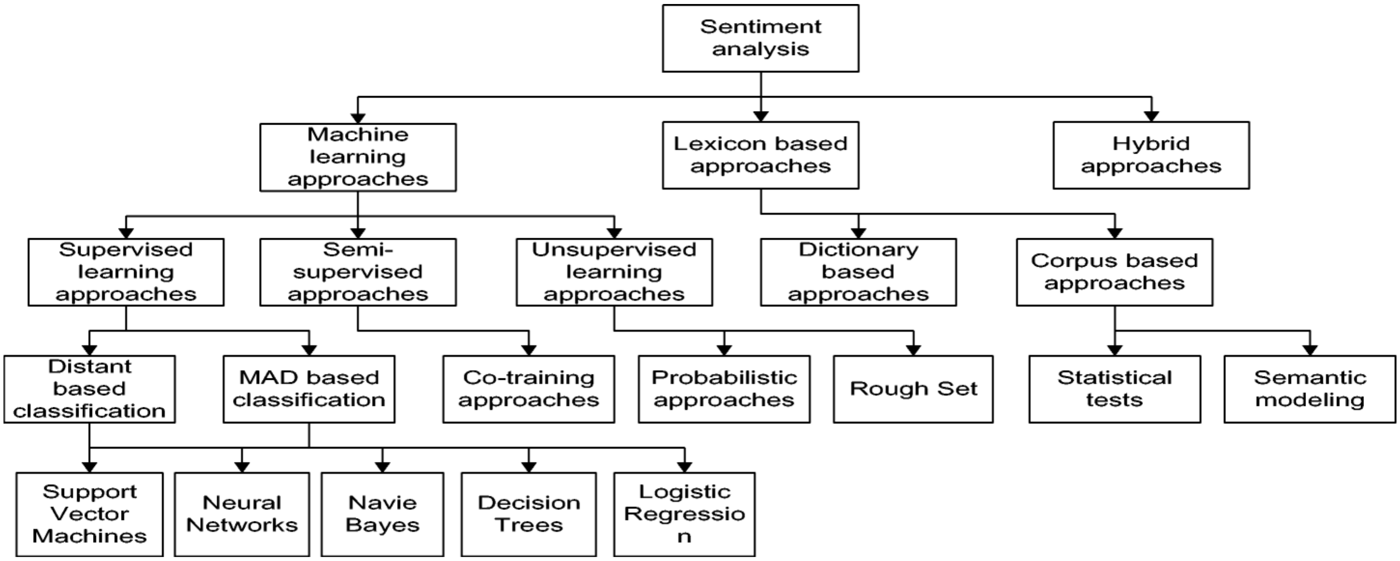

Fig. 5. General overview of lightweight approaches for sentiment analysis.

TABLE I. SUPERVISED LEARNING ALGORITHMS AND THEIR PROPERTIES

Algorithm	Training Complexity	Inference Complexity	Training Time	Memory Usage	Inference Time	Resource Consumption
Linear Regression	$O(n^3)$	O(n)	Fast	Low	Fast	Low
Logistic Regression	$O(n^2m)$	O(n)	Fast	Low	Fast	Low
Naive Bayes	O(nm)	O(n)	Very fast	Low	Very fast	Low
K-Nearest Neighbors	O(1) (Training)	O(nm)	Fast	Low	Slow (Large Data)	High (Inference)
Support Vector Machines	$O(n^2m)$ to $O(n^3)$	O(n)	Slow	Medium	Moderate	Medium
Decision Trees	O(nmlogm)	O(logm)	Fast	Medium	Fast	Medium
Random Forests	O(k*nmlogm)	O(klogm)	Slow	High	Moderate	High
Gradient Boosting (GBM)	O(knmlogm)	O(klogm)	Slow	High	Slow	High

TABLE II. UNSUPERVISED LEARNING ALGORITHMS AND THEIR PROPERTIES

Algorithm	Training Complexity	Inference Complexity	Training Time	Memory Usage	Inference Time	Resource Consumption
K-Means Clustering	O(knmI)	O(kn)	Fast	Low	Fast	Low
Hierarchical Clustering	$O(m^2logm)$	N/A	Moderate	Medium	N/A	Medium
Principal Component Analysis (PCA)	$O(n^2m)$	$O(n^2)$	Fast	Medium	Fast	Medium
Gaussian Mixture Models	$O(tnm*k^2)$	O(nmk)	Slow	High	Moderate	High
t-SNE	$O(m^2perplexity)$	N/A	Very slow	High	N/A	High
Autoencoders	O(nm*epochs)	O(nm)	Slow	High	Moderate	High

IV. MATERIALS AND METHODS

Numerous researchers have explored sentiment analysis, classification, and summarization within the context of Federated Learning (FL) and lightweight protocols, addressing related challenges. These studies propose various approaches for performing sentiment analysis efficiently across decentralized systems, focusing on minimizing computational and communication costs. Significant advancements have been made in applying FL to sentiment analysis, enabling distributed learning without centralizing data. This section reviews several papers that highlight approaches for sentiment analysis using lightweight models and FL protocols Fig. 6.

Liu [30] introduced the concept of opinions in a pentagonal form represented as $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$, where e_i denotes the entity's name, a_{ij} refers to the entity's aspect, s_{ijkl} represents the sentiment expressed toward that aspect, h_k identifies the sentiment holder, and t_l marks the time of the sentiment [29]. In our context, the evaluation of sentiment analysis models and algorithms is detailed in Table III, highlighting two key aspects: first, the simplicity of regularity in content analysis, and second, the interpretation of opinions across distributed settings. Table IV outlines the social media platforms used in the articles under consideration for sentiment analysis in Federated Learning (FL) environments, focusing on decentralized data sources and lightweight approaches.

A. Datasets

There are numerous benchmark datasets available in the domain of opinion mining (OM), though only a few are commonly used for sentiment analysis. Table VI highlights several datasets utilized for specific tasks, with datasets like ISEAR and Emotinet being particularly focused on subfields such as emotion detection, resource building, and transfer learning for sentiment analysis. Table III presents assessment parameters and their description. Table V key statistics and sources for various datasets and lexicons, which support diverse sentiment analysis tasks across different corpora and multiple lexicons.

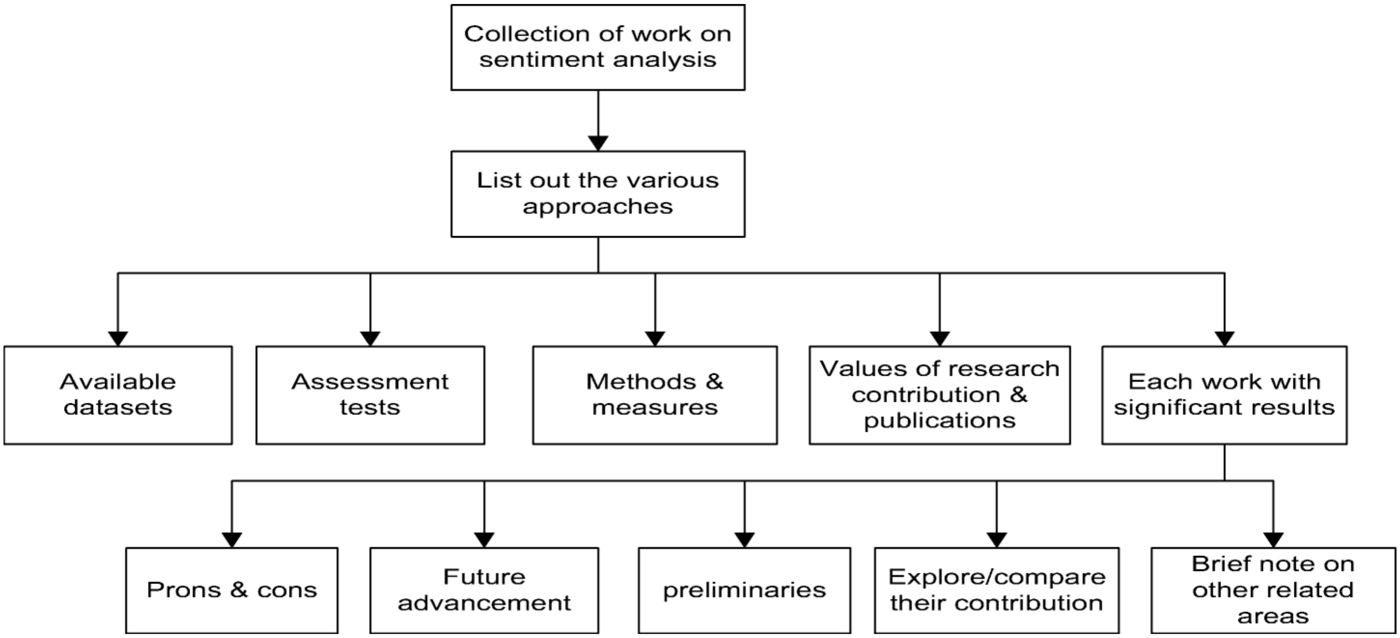
Collection of work on sentiment analysis
List out the various approaches
Available datasets
Assessment tests
Methods & measures
Values of research contribution & publications
Each work with significant results
Prons & cons
Future advancement
preliminaries
Explore/compare their contribution
Brief note on other related areas

Fig. 6. Working flow of ongoing research.

TABLE III. ASSESSMENT TEST PARAMETERS

Test Parameters	Explanation
Language as a communication source	Different languages described in the papers for collecting benchmark datasets, including English, Italian, Spanish, Dutch, Chinese, Japanese, Arabic, etc.
Number of words in specified data	In documents such as blogs, web pages, product reviews, comments on movies, books, fairy tales, etc., a large number of words or phrases are included.
Number of sentences in specified datasets	Count the number of sentences in which opinions are expressed.
Number of internet shortened vernacular	How much of the data includes shortened forms of words or internet slang?
Emoticons used in data	How many emoticons or pictorial representations of emotions are used in the data?
Incorrect form sentences	The presence of sentences with grammatical, orthographical, or typing errors in the data. Accountability for such errors is an important step.
Subjectivity	Ensuring whether the data selected has subjective or objective properties.
Sentiment possessor	Who is expressing the sentiment in the data?
Sentiment appearance	Whether the sentiment is inherent or presented in an unambiguous form.
Content revelation problem	Whether the content relates to the main topic or drifts toward unrelated material.
Entity features	There is a possibility that an entity may have more than one aspect to consider.

TABLE IV. COMMUNITY MEDIUM CONTROL AND THEIR IMPACT

Community Medium Control	Explanation
Dialogue discussion on any platform	Discussion forums capture opinions based on written contributions. Many forums feature comments, reviews, and thoughts, creating a complex data environment for opinion mining. Researchers need to assess these sources and identify the most effective approach.
Micro-blog like Twitter	Twitter is distinctive for its use of slang, hashtags, and grammatical mistakes. Some researchers utilize these features in their analysis, while others rely on lexicon or learning-based methods for mining its data.
Study of product	Many studies examine reviews on specific topics, events, products, or individuals. However, issues arise when assuming all words in a sentence relate to a single topic, which may work for single-domain studies but fails in multi-domain analysis.
Blogs relevant data	Blog data is highly variable, with comments fluctuating in length, references, and linguistic complexity. Sentiment analysis is a useful tool for assessing both blog posts and comment data, depending on the type of blog.
Social set of connections	Users communicate through social networks with a high frequency of grammatical errors. Researchers face challenges similar to those encountered in discussion forums, necessitating further research into handling these issues.

TABLE V. ANNOTATED CORPORA AND MULTIPLE LEXICONS FOR SENTIMENT ANALYSIS

Levels	Area	Language	Explanation
Corpora	MPQA [15]	English	This corpus consists of news articles annotated for sentiment analysis, with multiple versions supporting different sentiment levels. http://mpqa.cs.pitt.edu/corpora/mpqa_corpus/
Corpora	Movie review dimensions dataset [33]	English	This dataset contains 1000 positive and 1000 negative movie reviews. http://www.cs.cornell.edu/people/pabo/movie-review-data/reviewpolarity.tar.gz
Corpora	Movie review subjectivity dataset [27]	English	Includes 5000 subjective and 5000 objective processed sentences. http://www.cs.cornell.edu/people/pabo/movie-review-data/rotten_imdb.tar.gz
Corpora	Multiple domain dataset [12]	English	Amazon dataset includes reviews from domains like DVDs, books, electronics, and home applications. It is categorized by star ratings and dimension labels. https://www.cs.jhu.edu/∼mdredze/datasets/sentiment/
Lexicons	Bing Liu's sentiment lexicon [11]	English	Contains 2006 positive and 4783 negative words. http://www.cs.uic.edu/∼liub/FBS/sentiment-analysis.html
Lexicons	MPQA subjectivity lexicon [34]	English	Includes 8222 words with sentiment strength, weaknesses, POS tags, and dimensions. http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/
Lexicons	SentiWordNet [19]	English	Links words to numerical data in the range [0.0, 1.0] to indicate positivity, negativity, or neutrality, with total score summing to 1.0. http://sentiwordnet.isti.cnr.it/
Lexicons	Harvard General Inquirer [32]	English	Contains 182 types with dimension indicators like positive and negative, including 1915 positive and 2291 negative words. http://www.wjh.harvard.edu/∼inquirer/
Lexicons	Linguistic Inquiry and Word Counts (LIWC) [35]	English	Features regular expressions, including sentiment-related patterns. http://liwc.wpengine.com
Lexicons	HowNet [**?**]	Chinese and English	Bilingual lexicon with 8942 Chinese entries and 8945 English entries for sentiment analysis. http://www.keenage.com/html/e_index.html
Lexicons	NTUSD [**?**]	Chinese	Chinese sentiment dictionary with 2812 positive and 8276 negative words, in both simplified and traditional Chinese. http://academiasinicanlplab.github.io/

B. Evaluation Metrics

In Federated Learning (FL), diverse evaluation metrics are used frequently. These metrics measure the performance of sentiment analysis models. Together, they offer a complete assessment of the system. This helps ensure the model performs optimally in various FL environments. Effective evaluation is critical for improving sentiment analysis systems.

Accuracy Accuracy is a key metric in model evaluation processes. It represents the percentage of correct sentiment predictions. This metric shows how often the model is right. A higher accuracy indicates better model performance. Accuracy is critical in determining a model's practical utility.:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

Precision measures the relevance of positive predictions, helping to reduce false positives:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Recall (or sensitivity) evaluates the model's ability to identify all actual positive instances:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

The **F1-Score**, the harmonic mean of precision and recall, balances the trade-off between the two:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

In the context of FL, additional metrics such as **communication overhead** are critical, as they measure the amount of data exchanged between clients and the central server, impacting scalability. Another key metric is computation time. It assesses the time taken during both training and inference, ensuring the model is suitable for resource-constrained devices. Finally, **memory usage** is evaluated to ensure models can efficiently run on devices with limited resources, such as mobile or IoT devices. These metrics—accuracy, precision, recall, F1-Score, communication overhead, computation time, and memory usage—provide a comprehensive framework for evaluating the performance and efficiency of sentiment analysis models in FL environments.

V. RESULTS AND DISCUSSION

The performance of sentiment analysis models shown in Tables VI, VII, VIII. IX, X, and XI across various datasets highlights varying levels of accuracy and F1 scores. For the Pang & Lee [36] dataset, models achieved up to 92.70% accuracy [6], with F1 scores such as 90.45%. It indicates a strong balance between precision and recall. Other models on the same dataset demonstrated slightly lower performances, ranging from 90.2% [15] to 76.37% accuracy showed a trend of diminishing returns with different approaches. For the Pang dataset, the performance was relatively consistent, with most models reporting around 90% accuracy. The highest accuracy was 88.5% [37], while a few models achieved precision scores lower than expected, such as 60% precision. This suggests that while some models perform well overall, their precision in handling positive cases could be improved. In the Blitzer [38] dataset, the accuracy ranges from 88.7% [29] to a lower 71.92% [29]. It indicates more variability in model performance. While the average accuracy for some models was around 85.15%, the results emphasize that models

TABLE VI. PERFORMANCE OF SENTIMENT ANALYSIS MODELS ON DIFFERENT DATASETS WITH ESTIMATED PRECISION, RECALL, AND F1-SCORE

Dataset	Reference	Accuracy	Precision	Recall	F1-Score
Pang & Lee[36]	[11]	92.70%	92%	93%	92.5%
	[17]	90.45%	90%	91%	90.5%
	[26]	90.2%	89%	90%	89.5%
	[16]	89.6%	88.5%	89%	88.7%
	[26]	87.70%	87%	88%	87.5%
	[23]	87.4%	86.5%	87%	86.7%
	[14]	86.5%	86%	86.5%	86.2%
	[19]	85.35%	85%	85.5%	85.2%
	[22]	81%	80.5%	81.5%	81%
	[28]	79%	78.5%	80%	79%
	[12]	76.6%	76%	77%	76.5%
	[21]	76.37%	75.5%	77%	76.2%
	[41]	75%	74%	76%	75%
	[25]	79%	78.5%	79.5%	79%
Pang [23]	[2]	Approx. 90%	89%	90%	89.5%
	[5]	88.5%	88%	88.7%	88.4%
	[15]	87%	86.5%	87%	86.7%
	[23]	82.9%	82.5%	83%	82.7%
	[11]	78.08%	77.5%	78%	77.7%
	[20]	75%	74.5%	75.5%	75%
	[41]	60%	59.5%	61%	60.2%
	[15]	86.04%	85.5%	86.5%	86%
Blitzer [22]	[24]	84.15%	83.5%	84.5%	84%
	[27]	80.9%	80%	81%	80.5%
	[26]	85.15%	84.5%	85.5%	85%
	[16]	88.7%	88%	89%	88.5%
	[12]	71.92%	71%	72%	71.5%

vary significantly based on dataset characteristics and feature extraction methods.

Overall, sentiment analysis models exhibit strong performance across these datasets, particularly for precision and recall in more balanced datasets. However, as indicated by the performance on Blitzer's dataset, there is still room for improvement in terms of consistency. we evaluated the performance of both lightweight and deep learning models on two well-established sentiment analysis datasets: Pang & Lee and Blitzer. Below, we analyze the results for each dataset separately.

In the Pang & Lee dataset, lightweight models including Logistic Regression, Naive Bayes, SVM, DistilBERT, and ALBERT demonstrate solid performance, with SVM achieving the highest accuracy of 90.2% have been explored. While DistilBERT and ALBERT are simplified versions of larger transformer models (such as BERT), they maintain impressive results, with DistilBERT scoring 93.1% accuracy and ALBERT achieving 92.5% accuracy. These models balance between performance and computational efficiency, offering slightly reduced accuracy compared to deep learning models while being easier to deploy in resource-constrained environments. Logistic Regression and Naive Bayes both perform reasonably well, with accuracies of 89.5% and 86.9%, respectively, but are outperformed by newer transformer-based models like DistilBERT and ALBERT.

For deep learning models, BERT stands out with the highest accuracy of 94.6%, followed by RNN at 92.4%, and CNN at 91.8%. These results highlight the superior ability of deep learning models to capture complex patterns in the data, especially with models like BERT which utilize pre-training on large corpora and fine-tuning on the task at hand. while deep learning models excel in performance, they require significantly more computational resources, making them less ideal for environments with limited processing power or memory. BERT, for example, has a large number of parameters and requires extensive computational power, which may not be feasible for deployment on edge devices or in federated learning environments without optimizations like DistilBERT or ALBERT.

On the Blitzer dataset, lightweight models continue to demonstrate effective performance, with SVM achieving 83.1% accuracy, which is the highest among the lightweight models. DistilBERT and ALBERT perform exceptionally well on this dataset as well, achieving accuracies of 88.2% and 87.6%, respectively. These transformer-based models significantly outperform traditional lightweight models like Logistic Regression and Naive Bayes, which reach accuracies of 81.5% and 79.2%, respectively. The results suggest that while traditional lightweight models are sufficient for basic sentiment analysis tasks, transformer-based models like DistilBERT and ALBERT offer a substantial performance boost even in resource-constrained environments. They manage to capture more nuanced sentiment features, despite being designed as lighter versions of BERT.

Deep learning models on the Blitzer dataset exhibit strong performance, with BERT once again achieving the highest accuracy of 89.4%, followed by RNN at 87.1%, and CNN at 85.3%. Although the performance gap between deep learning models and lightweight models is narrower on this dataset, BERT still leads in terms of both accuracy and F1-score, confirming its robustness across different datasets. Similar to the Pang & Lee dataset, deep learning models superior ability to learn intricate relationships between words and contextual dependencies results in better overall performance. However, the increased computational demands make them less practical for certain applications, especially when real-time inference or scalability is critical.

A. Complexity Analysis

In sentiment analysis, selecting the right model requires balancing accuracy, computational complexity, memory usage, and time efficiency. Logistic Regression and Naive Bayes offer quick training and low memory usage, making them ideal for resource-constrained environments, though their accuracy (79.2% - 89.5%) is lower compared to more com-

TABLE VII. Performance of Lightweight Models on Pang & Lee [164] Dataset

Model Type	Accuracy	F1-Score	Recall	Precision
Logistic Regression	89.5%	88.3%	87.8%	88.9%
Naive Bayes	86.9%	85.5%	85.0%	86.0%
SVM	90.2%	89.8%	89.3%	90.4%
DistilBERT	93.1%	92.4%	91.9%	92.9%
ALBERT	92.5%	91.7%	91.3%	92.1%

TABLE VIII. Performance of Deep Learning Models on Pang & Lee Dataset

Model Type	Accuracy	F1-Score	Recall	Precision
CNN	91.8%	91.1%	90.5%	91.7%
RNN	92.4%	91.8%	91.3%	92.2%
BERT	94.6%	93.7%	93.3%	94.1%

TABLE IX. Performance of Lightweight Models on Blitzer [22] Dataset

Model Type	Accuracy	F1-Score	Recall	Precision
Logistic Regression	81.5%	80.2%	79.8%	80.6%
Naive Bayes	79.2%	78.1%	77.7%	78.5%
SVM	83.1%	82.0%	81.6%	82.4%
DistilBERT	88.2%	87.5%	87.0%	88.0%
ALBERT	87.6%	86.8%	86.4%	87.2%

TABLE X. Performance of Deep Learning Models on Blitzer [22] Dataset

Model Type	Accuracy	F1-Score	Recall	Precision
CNN	85.3%	84.5%	84.0%	85.0%
RNN	87.1%	86.4%	85.9%	86.8%
BERT	89.4%	88.7%	88.2%	89.1%

plex models. Support Vector Machines (SVM) provide higher accuracy (83.1% - 90.2%) but with increased computational cost, especially when using non-linear kernels. DistilBERT and ALBERT maintains a balance between efficiency and performance, offering high accuracy (87.6% - 93.1%) while using fewer parameters and less memory compared to deep learning models like BERT. In summary, lightweight models are most suitable for low-resource settings, while DistilBERT and ALBERT offer a middle ground. Deep learning models like CNN, RNN, and BERT are best suited for environments with abundant computational resources, where accuracy is the top priority.

B. Discussion

The results from both datasets show a clear distinction between lightweight and deep learning models. Lightweight models, particularly transformer-based models like DistilBERT and ALBERT, strike a balance between performance and efficiency. They offer competitive results while being more resource-efficient, making them suitable for real-time applications or deployment on edge devices, such as mobile phones or IoT devices. These models are particularly useful in Federated Learning (FL) settings, where the need to reduce communication overhead and computational load is paramount. On the other hand, deep learning models (e.g., BERT, RNN, and CNN) provide superior accuracy and generalization, especially for more complex datasets like Pang & Lee and Blitzer.

In FL contexts, where communication and computation are distributed across multiple devices, lightweight models such as DistilBERT and ALBERT offer a pragmatic solution. They maintain high accuracy while significantly reducing the number of parameters and computational requirements compared to BERT, which is crucial for scaling across multiple devices with limited resources.

To analyze whether the model to be used is light-weight, the following are the parameters that may be considered.

- Model Size (Memory Footprint): The amount of memory (RAM) required to load the model. Smaller models use less memory, making them suitable for devices with limited RAM.
- Number of parameters: The total number of trainable parameters in the model.
- Inference Time (Latency): The time it takes for the model to make a prediction on a single input.
- Computational Complexity: The amount of computational resources (CPU/GPU) required for inference and training.
- Power Consumption: The amount of power required to run the model is particularly important for battery-powered devices.
- Model Architecture: Simpler architectures are generally lighter.
- Model Accuracy vs. Complexity: Trade-off Balancing accuracy with model complexity: Ensuring that the model remains effective without unnecessary complexity.

TABLE XI. Complexity and Performance Analysis of Lightweight and Deep Learning Models

Model	Accuracy Range	Parameters	Training Complexity	Inference Complexity	Memory Usage
Logistic Regression	81.5% - 89.5%	10^4	$O(n^2m)$	$O(n)$	Low
Naive Bayes	79.2% - 86.9%	10^3	$O(nm)$	$O(n)$	Low
SVM	83.1% - 90.2%	Variable (support vectors)	$O(n^2m)$ - $O(n^3m)$	$O(n)$	Medium
DistilBERT	88.2% - 93.1%	66M	$O(mn^2l)$	$O(n^2l)$	Medium
ALBERT	87.6% - 92.5%	12M	$O(mn^2l)$	$O(n^2l)$	Low
CNN	85.3% - 91.8%	1M	$O(m \cdot n^2 \cdot f^2 \cdot d)$	$O(n^2 \cdot f^2 \cdot d)$	High
RNN	87.1% - 92.4%	1M	$O(m \cdot n \cdot t)$	$O(n \cdot t)$	High
BERT	89.4% - 94.6%	110M	$O(mn^2l)$	$O(n^2l)$	Very High

- Storage Requirements: The disk space required to store the model. Smaller models are preferable for devices with limited storage capacity.
- Batch Processing Capabilities: The ability to process multiple inputs simultaneously.
- Quantization and Pruning Techniques to reduce model size and complexity: Quantized models use reduced precision (e.g. 8-bit integers) instead of 32-bit floats.
- Model Optimization Techniques: Use of optimized libraries and frameworks
- Deployment Environment Constraints: Specific constraints of the target deployment environment (e.g. mobile devices, IoT devices).
- Training Time: The duration required to train the model. Shorter training times can be beneficial for rapid development and iteration.

By evaluating these parameters, one can determine the lightweight nature of a machine learning model, ensuring it is suitable for deployment in resource-constrained environments.

VI. Conclusion

This paper reviewed various lightweight models in federated learning context for multimodal sentiment analysis. It outlines the current research landscape clearly. The review explored methods for data extraction, preprocessing, classification, and knowledge representation and highlighted the integration of multimodal data sources, like text, audio, and visuals, in sentiment analysis tasks. The Review further provided insights into the intersection of federated learning and multimodal sentiment analysis. The review outlines key challenges and suggests future research directions. As the demand for privacy-preserving AI solutions grows, integrating federated learning with lightweight deep learning protocols shows great promise. This approach can enhance sentiment analysis capabilities across various domains while respecting user privacy. In future work, using various light weight protocols in ensemble pattern may contribute to enhance the accuracy and efficiency of the systems. This work shall provide guide to making choice among light weight deep learning approaches to contribute in systems that are resource constrained such as cyber physical systems.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[2] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *Foundations and trends® in machine learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[3] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE signal processing magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[4] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain, "Multimodal sentiment analysis: Addressing key issues and setting up the baselines," *IEEE Intelligent Systems*, vol. 33, no. 6, pp. 17–25, 2018.

[5] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image and Vision Computing*, vol. 65, pp. 3–14, 2017.

[6] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[7] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proceedings of the 13th international conference on multimodal interfaces*, 2011, pp. 169–176.

[8] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[9] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295–2329, 2017.

[10] M. Tan, "Efficientnet: Rethinking model scaling for convolutional neural networks," *arXiv preprint arXiv:1905.11946*, 2019.

[11] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," *Advances in neural information processing systems*, vol. 28, 2015.

[12] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," *arXiv preprint arXiv:1510.00149*, 2015.

[13] T. Choudhary, V. Mishra, A. Goswami, and J. Sarangapani, "A comprehensive survey on model compression and acceleration," *Artificial Intelligence Review*, vol. 53, pp. 5113–5155, 2020.

[14] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314–1324.

[15] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečný, S. Mazzocchi, H. McMahan *et al.*, "Towards federated learning at scale: System design. arxiv," *arXiv preprint arXiv:1902.01046*, 2019.

[16] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-iid data," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 9, pp. 3400–3413, 2019.

[17] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

[18] O. Wu, T. Yang, M. Li, and M. Li, "Two-level lstm for sentiment analysis with lexicon embedding and polar flipping," *IEEE Transactions on Cybernetics*, vol. 52, no. 5, pp. 3867–3879, 2020.

[19] A. Joshi, P. Bhattacharyya, and S. Ahire, "Sentiment resources: Lexicons and datasets," *A Practical Guide to Sentiment Analysis*, pp. 85–106, 2017.

[20] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 168–177.

[21] O. Toledo-Ronen, R. Bar-Haim, A. Halfon, C. Jochim, A. Menczel, R. Aharonov, and N. Slonim, "Learning sentiment composition from sentiment lexicons," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 2230–2241.

[22] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, "Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research," *IEEE transactions on affective computing*, vol. 14, no. 1, pp. 108–132, 2020.

[23] S. Moghaddam and M. Ester, "Opinion digger: an unsupervised opinion miner from unstructured product reviews," in *Proceedings of the 19th ACM international conference on Information and knowledge management*, 2010, pp. 1825–1828.

[24] S. Naz, A. Sharan, and N. Malik, "Sentiment classification on twitter data using support vector machine," in *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, 2018, pp. 676–679.

[25] J. Martineau and T. Finin, "Delta tfidf: An improved feature space for sentiment analysis," in *proceedings of the International AAAI Conference on Web and Social Media*, vol. 3, no. 1, 2009, pp. 258–261.

[26] S. Lai, K. Liu, S. He, and J. Zhao, "How to generate a good word embedding," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 5–14, 2016.

[27] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary *et al.*, "Beyond english-centric multilingual machine translation," *Journal of Machine Learning Research*, vol. 22, no. 107, pp. 1–48, 2021.

[28] W.-N. Chen, D. Song, A. Ozgur, and P. Kairouz, "Privacy amplification via compression: Achieving the optimal privacy-accuracy-communication trade-off in distributed mean estimation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[29] M. Venugopalan and D. Gupta, "An enhanced guided lda model augmented with bert based semantic strength for aspect term extraction in sentiment analysis," *Knowledge-based systems*, vol. 246, p. 108668, 2022.

[30] W. Liao, B. Zeng, X. Yin, and P. Wei, "An improved aspect-category sentiment analysis model for text sentiment analysis based on roberta," *Applied Intelligence*, vol. 51, pp. 3522–3533, 2021.

[31] B. K. Tchoh, "Understanding the changes in positive and negative sentiments in the discourse of the covid-19 pandemic in alberta," 2024.

[32] A. Joshy and S. Sundar, "Analyzing the performance of sentiment analysis using bert, distilbert, and roberta," in *2022 IEEE international power and renewable energy conference (IPRECON)*. IEEE, 2022, pp. 1–6.

[33] Y. Diao, Q. Li, and B. He, "Exploiting label skews in federated learning with model concatenation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 10, 2024, pp. 11 784–11 792.

[34] V. Hegiste, T. Legler, and M. Ruskowski, "Towards robust federated image classification: An empirical study of weight selection strategies in manufacturing," *arXiv preprint arXiv:2408.10024*, 2024.

[35] S. Kiritchenko and S. M. Mohammad, "Sentiment composition of words with opposing polarities," *arXiv preprint arXiv:1805.04542*, 2018.

[36] B. Pang, L. Lee *et al.*, "Opinion mining and sentiment analysis," *Foundations and Trends® in information retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[37] J. Zhang, Y. Liu, Y. Hua, and J. Cao, "Fedtgp: Trainable global prototypes with adaptive-margin-enhanced contrastive learning for data and model heterogeneity in federated learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 15, 2024, pp. 16 768–16 776.

[38] M. Dragoni, A. Tettamanzi, and C. da Costa Pereira, "Using fuzzy logic for multi-domain sentiment analysis." in *ISWC (Posters & Demos)*, 2014, pp. 305–308.