International Journal of Multidisciplinary Research and Publications
ISSN (Online): 2581-6187

Systematic Literature Review on Sentence Level Sentiment Analysis

Muhammad Garzali Qabasiyu[1], Musa Ahmed Zayyad[2], Shamsu Abdullahi[3]
[1]Department of Computer Science and Information Technology, Al-Qalam University Katsina, Katsina State, Nigeria
[2]Department of Information Technology, School of Mathematics and Computing, Kampala International University Uganda
[1,3]Department of Computer Studies, College of Science and Technology, Hassan Usman Katsina Polytechnic, Katsina State, Nigeria
Email address: garzaliqabasiyu@hukpoly.edu.ng, zayyad.musa@kiu.ac.ug, sasalihawa09@hukpoly.edu.ng

***Abstract*—*Sentiment analysis is the task of determining the sentiment expressed in a given text. It has gained significant attention in recent years due to its applications in various domains such as social media analysis, customer feedback analysis, and opinion mining. While many studies focus on document-level sentiment analysis, the analysis of sentiments at the sentence level provides a more fine-grained understanding of text sentiment. This paper presents a systematic literature review (SLR) on sentence-level sentiment analysis, employing the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. The objective of this study is to identify the state-of-the-art techniques, methodologies, datasets, and evaluation metrics used in sentence-level sentiment analysis studies published between 2010 and 2022. The SLR methodology involves a comprehensive search across multiple databases, such as Research Gate, Science Direct, Dimensions and Google Scholar, to retrieve relevant papers. A rigorous inclusion and exclusion criteria are applied to ensure the selection of papers that specifically address sentence-level sentiment analysis. Data extraction and synthesis are performed to capture essential information, including the contribution facets, types of datasets used, and evaluation metrics utilized. The findings of this SLR reveal several significant trends and advancements in sentence-level sentiment analysis. Various machine learning techniques have been widely adopted for sentiment classification at the sentence level. Additionally, the availability of large-scale annotated datasets, such as Tweets and Movie Review dataset, has played a crucial role in improving the performance of sentiment analysis models. Furthermore, evaluation metrics and benchmarks have emerged to assess the effectiveness of different models, including accuracy, precision, recall, F1-score, and AUC. This research paper contributes to the existing literature by providing a comprehensive overview of the current state-of-the-art in sentence-level sentiment analysis. The identified trends, contribution facets, datasets, and evaluation metrics can serve as a valuable resource for researchers and practitioners interested in further advancing the field.*

***Keywords*—** *Opinion classification, Sentence level, Sentiment analysis.*

I. INTRODUCTION

Before the internet, obtaining people's opinions required the use of tactics such as questionnaires and surveys, or just asking family members or friends [1]. These strategies are commonly used by organizations to obtain feedback on services or products. People may now voice their ideas about nearly anything on blogs, comment sections, and social media websites, thanks to the explosive development of social media websites [2]. Structured, semi-structured, and unstructured data can be found on the internet. Many scholars and practitioners have focused on methods and techniques for applying and extracting relevant information from this data [3].

Sentiment analysis is a computational technique designed to extract subjective information and sentiments are expressed in textual data [4]. It has become a crucial area of research due to its wide-ranging applications, including social media analysis, customer feedback analysis, market research, and brand reputation management. While sentiment analysis at the document level has been extensively studied, the analysis of sentiments at the sentence level offers a more granular understanding of text sentiment, enabling fine-grained analysis and interpretation of opinions [5].

In contemporary times, individuals, as well as private and government organizations, increasingly rely on content from these platforms to inform their decision-making processes. The application of opinion mining and sentiment analysis (SA) spans a wide range of fields, including business, where consumer satisfaction and expectations are assessed through online opinions [6]. The availability of public user reviews on the web means that one no longer needs to consult friends or family for product recommendations, and organizations can access open opinions without conducting surveys or polls [7]. However, keeping up with the rapid proliferation of diverse online platforms and sifting through the vast amount of sentiment-laden text they contain remains a formidable challenge. Identifying relevant sites, extracting sentiment, and summarizing the information within them are tasks that often pose difficulties for human readers. Hence, there is a need for a comprehensive review of sentiment-related issues.

SA involves the analysis of people's opinions, appraisals, attitudes, and emotions toward various entities, including products, services, individuals, organizations, issues, topics, events, and their attributes [8]. This field encompasses a broad range of studies, with "sentiment analysis" being more commonly used in industrial contexts, while "sentiment analysis" and "opinion mining" are frequently employed in academic settings, essentially referring to the same concept. SA research typically focuses on three levels: document, sentence, and aspect. Document-level SA aims to classify whether an entire document expresses a positive or negative sentiment, while sentence-level SA involves determining whether each sentence conveys a positive, negative, or neutral opinion.

Muhammad Garzali Qabasiyu, Musa Ahmed Zayyad, and Shamsu Abdullahi, "Systematic Literature Review on Sentence Level Sentiment Analysis," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 6, Issue 5, pp. 1-8, 2023.

The need for sentence-level sentiment analysis arises from the fact that sentiments are often expressed at a more nuanced level within a document. Analyzing sentiments at a sentence level offers several advantages over document-level analysis. It enables the identification of contrasting sentiments within a single text, the detection of sentiment shifts or transitions, and the ability to capture the sentiment associated with specific aspects or entities mentioned in the text.

While various reviews have been conducted in the field of sentiment analysis and sentence sentiment analysis, none have simultaneously focused on techniques for extracting sentence polarity. Therefore, this systematic literature review (SLR) fills this gap by providing a comprehensive overview of aspect extraction techniques. The primary objectives of this study are:

1. To identify dataset used in sentence-level sentiment analysis.
2. To analyze the evaluation metrics in sentence-level sentiment analysis.
3. To highlight contribution facet in sentence-level sentiment analysis.
4. To identify potential areas for future research.

This review follows the systematic review procedures outlined by [9] and is structured as follows: Section 2 describes the methodology, Section 3 presents the results of the review, Section 4 result and discusses the findings, Section 5 offers concluding remarks.

II. METHODOLOGY

We adopted the guidelines of writing systematic literature review by [9] and elements of PRISMA guideline in [10] as depicted in figure 1.

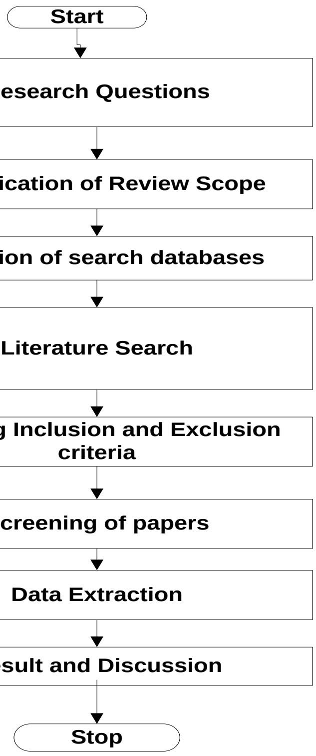

Fig. 1. Research methodology

A. *Research Questions (RQs):*

In this subsection, we define the RQs that our SLR intend to address. The RQs are in-line with the scope of our SLR on sentence level sentiment analysis. To achieve the desired objectives, the following RQs were outlined:

RQ1: What kinds of datasets are commonly used in sentence level sentiment analysis?
RQ2: What types of evaluation metrics are mostly employed by sentence level sentiment analysis?
RQ3: What are the contribution facets use in sentence level sentiment analysis?

B. *Search Databases*

This sub-section present the electronic databases used to sources the studies for our SLR. We selected 4 electronic databases with high precision as presented along with their URL in Table 1.

TABLE 1. Searched databases.

S. No.	Source	Link
1	Google scholar	http://scholars.google.com/
2	Research gate	https://www.researchgate.com
3	Science direct	https://www.sciencedirect.com
4	Dimensions	https://www.dimensions.ai/

C. *Literature Search Strategy*

The literature search begins with the formulation of search string based on the keywords "sentence level sentiment analysis. The keywords manipulated across the defined databases in section 2.2 to find the existing works on the sentence level sentiment analysis. The title, abstract and index were used to identify whether the journal articles, and conference proceedings could be used to conduct the research.

D. *Inclusion Criteria and Exclusion Criteria*

Inclusion and exclusion criteria were set in this study to further refine results obtained from the search process. The criteria were applied to all the studies retrieved from the adopted electronic databases as shown in Figure 1. To search for the relevant studies, the period of seven years was set (from January 2010 to June 2022), this is to ensure that only relevant and up-to-date articles published are included in the study. The inclusion and exclusion criteria in this study are describe as they were applied in the second and third stages of the defined protocol in Figure 2.

TABLE 2. Inclusion and exclusion criteria

S. No.	Inclusion criteria	Exclusion criteria
1	IC1: This study cover papers been published from 2010 to 2022	EC1: Papers which published before 2010 or after 2022
2	IC2: Papers that used Sentence base sentiment analysis	EC2: Papers which are not in English
3	IC3: Papers that enlisting method of sentiment analysis	EC3: article not addressing RQs
4	IC4: Papers that used machine learning method	EC2: Papers which did not used machine learning method

E. *Screening Papers and Selection*

The study selection process is intended to identify the relevant primary studies that provide direct evidence about the

Muhammad Garzali Qabasiyu, Musa Ahmed Zayyad, and Shamsu Abdullahi, "Systematic Literature Review on Sentence Level Sentiment Analysis," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 6, Issue 5, pp. 1-8, 2023.

formulated research questions. Figure 1 presents the procedure of study selection in our SLR. It comprises of four stages which were achieved by all the authors to have consensus and avoid bias. The initial results of the search string resulted to 145 articles, next, the duplicate articles were removed prior to the screening based on the titles. Furthermore, 59 articles were fully reviewed and another screening was carried out on the full articles based on inclusion/exclusion and quality assessment criteria. Finally, 24 studies were selected as the primary studies in this study.

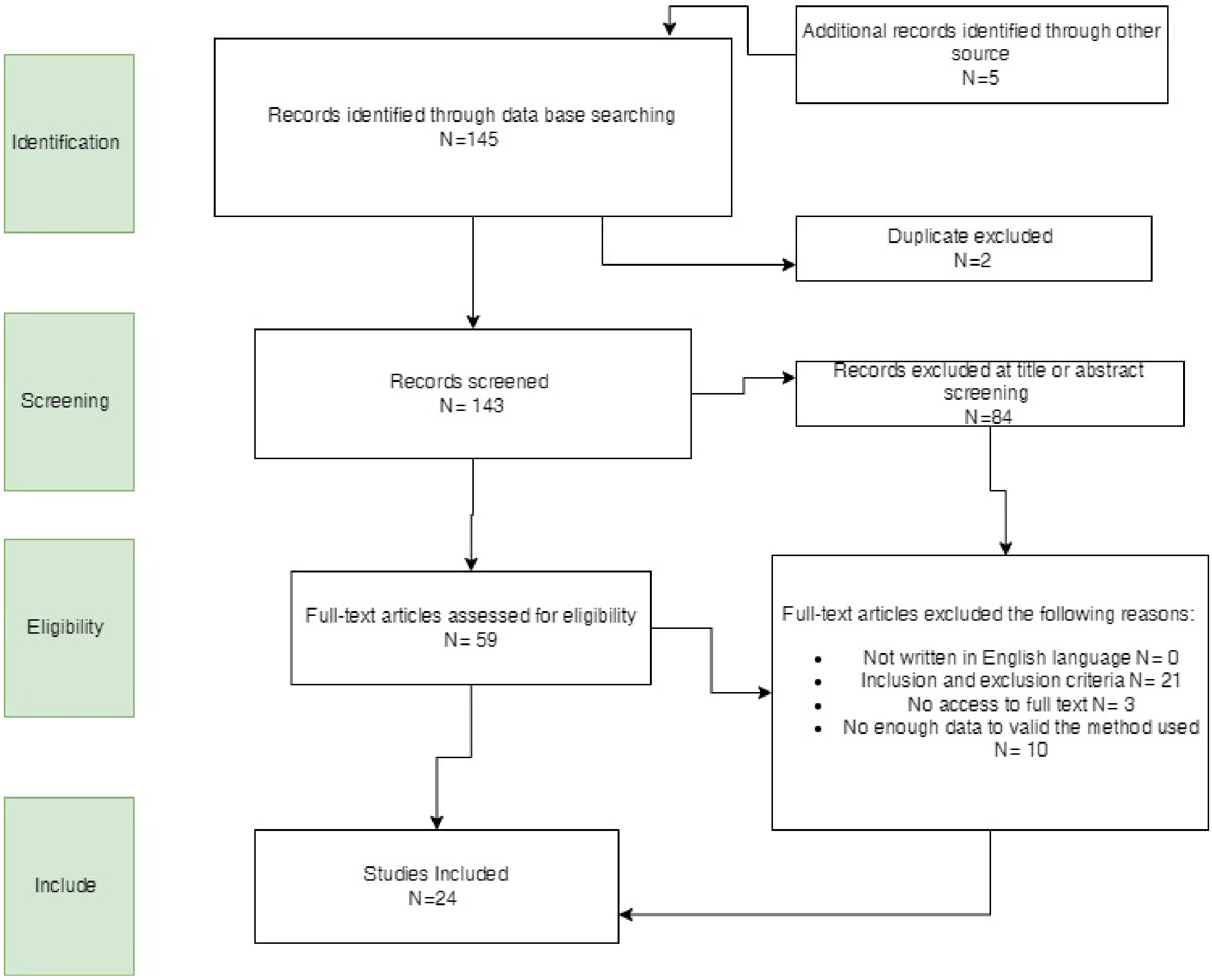

Fig. 2. Screening protocol

F. Data Extraction

These steps help the researchers to accurately extract the relevant information from the selected studies. The authors reviewed the full text of the primary studies selected in this study and the information extracted was stored in Microsoft excel file designed for this purpose. The extracted information included are: title, authors name, technique, contribution facet, performance metrics, and dataset used.

III. RESULT AND DISCUSSION

The objective of this research paper is to provide a systematic literature review (SLR) on sentence-level sentiment analysis, focusing on papers published between 2010 and 2022. By following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, this study aims to identify the datasets, contribution facet, and evaluation metrics employed in sentence-level sentiment analysis research.

RQ1: What kinds of datasets are commonly used in sentence level sentiment analysis?

This RQ investigates the datasets used in sentence level sentiment analysis as presented in Figure 3. Our finding indicates that there are 12 datasets; they are categorized as (Product review, Customers comments, Tweets, Movie review, Review dataset, Restaurant review, News article, Blog comment, Hotel review, Digital product Review and Food review). However, we have found that the study of [26] used more than one dataset. Among the identified Datasets in the study, Tweets and Movies Review constitutes the majority with 20% of the selected studies each contrary to what has been presented in [10] that found experiments as the most utilized evaluation technique, then followed by product review which have 16%, Blog comment 12%, food review, digital marketing

International Journal of Multidisciplinary Research and Publications
ISSN (Online): 2581-6187

review, hotel review, news article, restaurant review, financial review and customer comments review with 4% each.

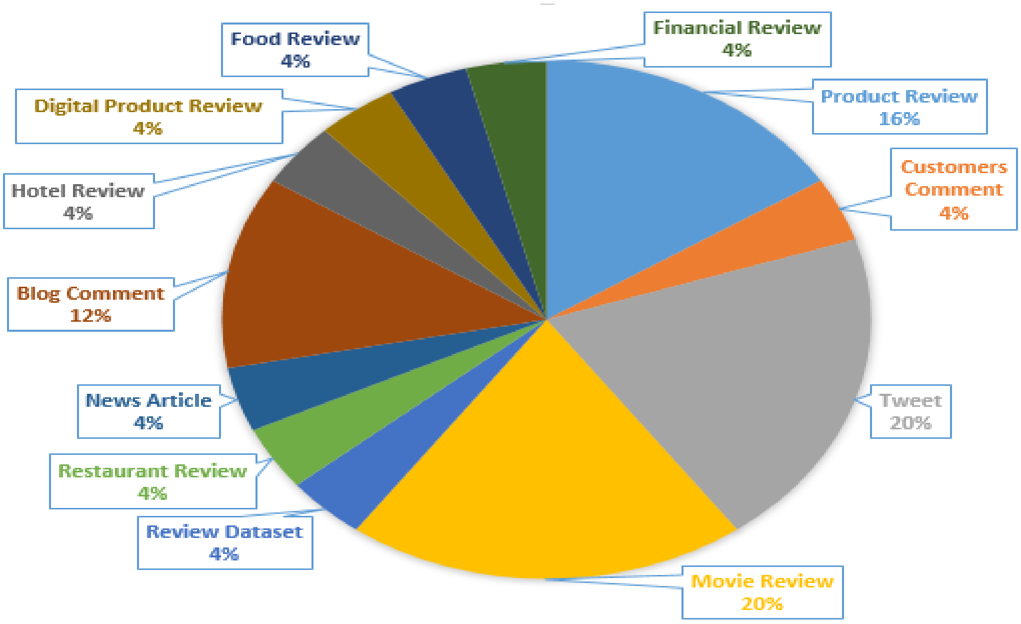

Fig. 3. Datasets

The dataset used in sentence level sentiment analysis are discussed below**.**

a. Product Reviews: Product reviews encompass a broad category of consumer opinions on various goods and services. They are widely utilized in sentence-level sentiment analysis due to their availability and relevance in understanding customer sentiment towards specific products. These datasets enable researchers to analyze sentiment expressions related to different product categories and assess customer satisfaction, sentiment polarity, and feature-specific sentiments.
b. Customer Comments: Customer comments, often extracted from online platforms such as e-commerce websites or social media platforms, provide a valuable source of sentiment-rich data. These comments offer insights into customer experiences, opinions, and sentiments towards a wide range of products, services, or brands. Customer comments datasets are particularly useful for capturing real-time sentiment expressions and monitoring customer feedback.
c. Tweets: Twitter, being a microblogging platform, offers a concise and dynamic source of sentiment-rich data for sentence-level sentiment analysis. Tweets are short, user-generated messages that often express personal opinions, emotions, or reactions. Due to their brevity, tweets datasets are commonly used to analyze sentiments related to trending topics, events, or public opinions on various subjects.
d. Movie Reviews: Movie reviews datasets consist of sentiment-labeled sentences or texts that express opinions about films or specific cinematic elements. These datasets are extensively used in sentiment analysis research due to their availability and the inherent subjectivity associated with movie opinions. Movie reviews provide insights into sentiment expressions related to plot, acting, direction, and overall cinematic experiences.
e. Review Datasets: Review datasets encompass generic datasets that compile various types of reviews, including product reviews, movie reviews, restaurant reviews, hotel reviews, and more. These datasets are valuable resources for training and evaluating sentence-level sentiment analysis models across multiple domains. They offer diverse perspectives on sentiment expressions and enable researchers to analyze sentiment patterns across different review categories.
f. Restaurant Reviews: Restaurant reviews datasets focus on sentiments expressed in customer reviews specifically related to dining experiences, food quality, service, ambiance, and other aspects of the restaurant industry. These datasets provide valuable insights into customer preferences, satisfaction levels, and sentiment variations across different cuisines or dining establishments.
g. News Articles: News articles datasets provide a source of sentiment-rich data for sentence-level sentiment analysis within the realm of journalism and current affairs. These datasets often cover diverse topics and domains, enabling researchers to analyze sentiment expressions related to news events, political opinions, or public sentiment towards specific issues.
h. Blog Comments: Blog comments datasets capture sentiment-rich opinions expressed by users in response to blog articles or posts. These datasets allow researchers to analyze sentiments related to specific blog comments,

including sentiment lexicons, deep learning models, and sentiment coherence analysis.

i. Hotel Reviews: This dataset comprises reviews of hotels and accommodations provided by customers. Record typically includes the review text, the reviewer's rating, the hotel name, location, and additional details such as check-in date and room type. Hotel reviews help potential travelers make informed decisions about their accommodations.

j. Digital Product Reviews: Digital product reviews encompass a wide range of products, including software, mobile applications, and electronic gadgets, taking into account specific aspects such as usability, features, and customer support.

k. Food Reviews: Food reviews offer insights into customers' opinions on restaurants, dishes, and overall dining experiences. This dataset contains reviews related to food establishments, including restaurants, cafes, and food delivery services.

l. Financial News: Sentiment analysis in financial news plays a crucial role in understanding market sentiments and predicting stock market trends. This dataset focuses on news articles and reports related to financial markets, stocks, investments, and economic trends. Record typically includes the news article text, publication date, author, source, and possibly other relevant information.

TABLE 3. Dataset

Name of Dataset	References
Product review	[11], [12], [13], [14]
Customers comments	[15]
Tweets	[13], [16], [17], [18], [14]
Movie review	[17], [19], [20], [14], [21]
Review dataset	[22], [18]
Restaurant review	[23]
News article	[24]
Blog comment	[25], [26], [27]
Hotel review	[26]
Digital product Review	[26]
Food review	[14]
Financial new	[28]

Tweets and movie reviews are commonly used datasets for sentence-level sentiment analysis due to their availability, concise nature, opinion-rich content, sentiment labeling, and real-time relevance. Tweets offer a vast amount of user-generated content in a short format, while movie reviews provide concise evaluations of films. Both datasets are valuable for sentiment analysis as they contain opinionated content and are often annotated with sentiment labels. Additionally, tweets capture real-time conversations and events, making them useful for analyzing current trends and news. Overall, these datasets are widely used and provide valuable resources for training and evaluating sentiment analysis models.

RQ2: What types of evaluation metrics are mostly employed by sentence level sentiment analysis?

This RQ investigates the evaluation metrics used by the selected studies. Our analysis reveal the existence of 6 prominent evaluation metrics as depicted in Figure 4. Additionally, almost all of the papers in the study used more than one evaluation metric as such studies are counted on all the evaluation metric used. However, from the primary studies, 12 used accuracy, then 11 used F1 Score, also precision and recall are used by 9 and 8 studies respectively. Finally, Macro-F1, Loss and AUC score are used by 3, 2 and 1 studies respectively.

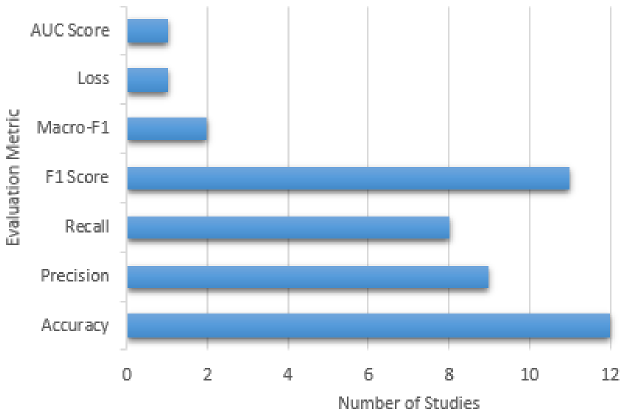

Fig. 4. Evaluation metrics

Metrics for evaluation are used to gauge how well a statistical or machine learning model is performing. Every project has to evaluate machine learning models or algorithms. To test a model, a wide variety of evaluation measures are available. They are different evaluation metrics employed by sentence level sentiment analysis.

a. Accuracy: Accuracy is a widely used evaluation metric that measures the overall correctness of sentiment predictions. It calculates the ratio of correctly classified sentences to the total number of sentences in the dataset. While accuracy provides a general indication of model performance, it may not be suitable for imbalanced datasets where the majority class dominates. In such cases, other metrics like precision, recall, and F1 score are more informative.

$$\text{Accuracy} = \frac{Total\ Number\ of\ Predictions}{Number\ of\ correct\ prediction} \text{ x } 100\%$$

b. Precision: Precision measures the proportion of correctly predicted positive or negative sentences out of all sentences predicted as positive or negative. It focuses on the correctness of positive or negative predictions, irrespective of missed predictions or misclassifications of neutral sentences. Precision is useful when the goal is to minimize false positive errors and ensure high precision in sentiment predictions.

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

c. Recall: Recall, also known as sensitivity or true positive rate, calculates the proportion of correctly predicted positive or negative sentences out of all actual positive or negative sentences in the dataset. It emphasizes the ability of a model to capture positive or negative sentiment instances correctly. Recall is particularly important when the goal is to minimize false negative errors and ensure high recall in sentiment predictions.

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

d. F1 Score: The F1 score is the harmonic mean of precision and recall. It provides a balanced measure of a model's performance by considering both precision and recall. The F1 score is useful in situations where there is an imbalance between positive, negative, and neutral classes. It is a popular metric for evaluating sentence-level sentiment analysis models due to its ability to capture the trade-off between precision and recall.

$$\text{F1 Score}=\frac{2\ x\ Precison\ x\ Recall}{Precision + Recall}$$

e. Macro-F1 Score: The macro-F1 score calculates the average F1 score across all classes (positive, negative, and neutral). It treats each class equally and does not consider class imbalance. The macro-F1 score is suitable when all classes are of equal importance, and the goal is to achieve balanced performance across all sentiment categories.

$$\text{Macro-F1 Score}=\frac{\sum_{i=1}^{N} F1_i}{N}$$

Where N is the number of classes and $F1_i$ *is the F1 score for class i.*

f. Loss: Loss functions, such as cross-entropy loss, are commonly used during the training phase of sentiment analysis models. These metrics quantify the discrepancy between predicted sentiment labels and the true labels. While loss is not typically used as a standalone evaluation metric, it provides valuable insights into the optimization process and model convergence during training.

g. AUC Score: The Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) score is often used in binary sentiment analysis tasks. It measures the model's ability to distinguish between positive and negative sentiments at various threshold settings. The AUC score provides a summary of the model's performance across all possible classification thresholds and is particularly useful when the focus is on ranking or prioritizing sentiment predictions.

AUC Score = Area Under ROC Curve

TABLE 4. Evaluation metrics

Evaluation metrics	References
Accuracy	[15], [16], [17], [29], [30], [18], [23], [24], [25], [31], [21], [28]
Precision	[15], [13], [16], [17], [29], [24], [25], [26], [28]
Recall	[15], [13], [16], [17], [29], [25], [26], [28]
F1 score	[15], [12], [13], [16], [17], [29], [24], [25], [26], [14], [28]
Macro-f1	[32], [19]
Loss	[31]
AUC Score	[31]

Accuracy and F1 score are commonly employed metrics in sentence-level sentiment analysis because they provide a comprehensive evaluation of the model's performance in capturing sentiment information.

The accuracy is widely used as a performance metric because it is easy to interpret and understand. However, accuracy alone may not be sufficient in sentiment analysis tasks, especially when the dataset is imbalanced. For instance, if a sentiment analysis dataset contains a large number of positive sentences but only a few negative sentences, a model that classifies all sentences as positive will achieve a high accuracy even though it fails to capture the negative sentiment. Therefore, accuracy should be used in conjunction with other metrics.

The F1 score combines precision and recall, providing a balanced evaluation of the model's performance. Precision measures the proportion of correctly classified positive sentences among all sentences predicted as positive, while recall measures the proportion of correctly classified positive sentences among all actual positive sentences in the dataset. The F1 score is the harmonic mean of precision and recall, giving equal weight to both metrics. It is particularly useful when the dataset is imbalanced or when both precision and recall are equally important. In sentiment analysis, it is crucial to capture both positive and negative sentiments accurately, so the F1 score is commonly used to assess the model's performance.

By considering both accuracy and F1 score, researchers and practitioners can gain a more comprehensive understanding of the model's effectiveness in sentence-level sentiment analysis. While accuracy provides an overall assessment of correctness, the F1 score takes into account the balance between precision and recall, considering the trade-offs between false positives and false negatives. Together, these metrics offer valuable insights into the model's ability to capture sentiment information accurately, making them the commonly employed evaluation measures in this field.

RQ3: What are the contribution facets use in sentence level sentiment analysis?

This RQ investigates the contribution facet by the selected studies. As presented in Figure 5, Method and Approach are the most addressed facet discussed in the selected studies with 8 and 7 studies each, followed by model proposed 5 of the selected studies. Also, 3 of the selected studies presented framework as their contribution facets, 1 of the study used new resources method and 1 other is not clear on its contribution facet.

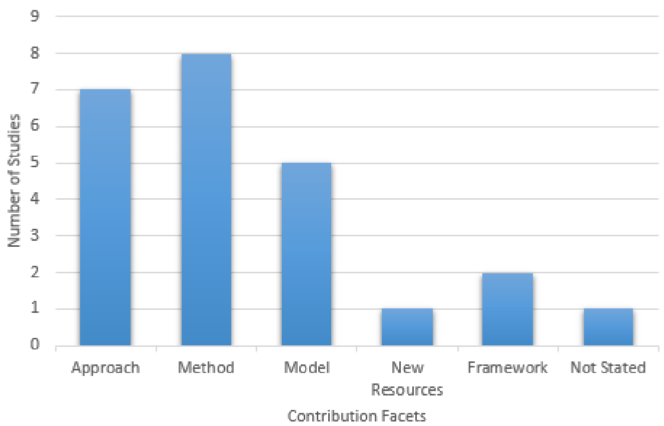

Fig. 5. Contribution Facets

The primary research is also divided into categories depending on research methodologies and contribution elements. In addition, contribution facets are grouped into model, Framework, approach, technique, and method.

a. Approach: This refers to the novel or unique strategies and techniques used to tackle sentence-level sentiment analysis. It could involve the integration of multiple methods or the application of a specific set of rules to handle sentiment classification at the sentence level.
b. Method: The method facet focuses on the specific techniques and algorithms employed to extract and analyze sentiment from individual sentences. It may involve feature engineering, statistical approaches, machine learning algorithms, or deep learning architectures tailored for sentence-level sentiment analysis.
c. Model: The model facet highlights the development and utilization of specific models or frameworks designed to capture the sentiment expressed within sentences. This could include the use of pre-trained language models, recurrent neural networks, convolutional neural networks, or other advanced architectures for sentiment classification.
d. New Resource: The new resource facet pertains to the creation or enhancement of sentiment-related linguistic resources specifically tailored for sentence-level sentiment analysis. This could involve the construction of sentiment lexicons, sentiment-specific datasets, or annotated corpora that facilitate accurate sentiment classification at the sentence level.
e. Framework: The framework facet encompasses the overall architecture or structure proposed to conduct sentence-level sentiment analysis. It includes the integration of various components such as data preprocessing, feature extraction, sentiment classification models, and evaluation metrics, providing a systematic and cohesive approach to analyze sentiment at the sentence level.

These contributions, spanning across the facets of approach, method, model, new resource, and framework, collectively contribute to advancing the field of sentence-level sentiment analysis by improving the accuracy, efficiency, and effectiveness of sentiment classification at the sentence level.

TABLE 5. Contribution facets

Contribution facets	Reference
Approach	[11], [33], [16], [29], [25], [14], [27]
Method	[12], [17], [22], [24], [34], [26], [21], [28]
Model	[13], [30], [19], [31], [20]
New resource	[15]
Frame work	[18], [23],

Approach, method, and model are the key facets used in sentence-level sentiment analysis to make significant contributions to the field.

The approach encompasses the high-level considerations and decisions made during the analysis process. For example, an approach may involve preprocessing steps such as tokenization, removing stop words, or performing part-of-speech tagging. It may also involve feature extraction techniques, such as bag-of-words or word embedding. The approach sets the foundation for the subsequent steps in sentiment analysis and significantly influences the overall performance and outcomes of the analysis.

Various methods have been proposed and utilized in sentence-level sentiment analysis, including rule-based methods, machine learning, and deep learning techniques. Rule-based methods rely on predefined linguistic rules or patterns to determine sentiment polarity. Machine learning methods involve training models on labeled sentiment data and using them to predict sentiment labels for unseen sentences. Deep learning methods, such as recurrent neural networks (RNNs) or transformers, leverage neural network architectures to capture complex patterns and dependencies in text data. The choice of method depends on the specific requirements, available resources, and the desired level of accuracy in sentiment analysis.

The model involves training a specific algorithm or neural network architecture on a given dataset to learn the patterns and relationships between textual features and sentiment labels. The model parameters are optimized during the training process to minimize the prediction errors. The performance of the model is assessed using evaluation metrics such as accuracy, F1 score, or precision and recall. Models can be further improved through techniques such as hyperparameter tuning, assembling, or transfer learning.

By focusing on approach, method, and model, researchers and practitioners in sentence-level sentiment analysis can contribute to the field through the development and application of novel strategies, techniques, and implementations. These facets allow for advancements in understanding sentiment in text, improving the accuracy of sentiment classification, and exploring new possibilities for sentiment analysis in various domains and languages.

IV. CONCLUSION

This systematic literature review provides a comprehensive analysis of datasets, evaluation metrics, and contribution facets commonly employed in sentence-level sentiment analysis. The insights gained from this review contribute to the understanding of current trends and advancements in the field, facilitating further research and the development of more accurate and robust sentence-level sentiment analysis models. By utilizing diverse datasets, employing appropriate evaluation metrics, and exploring various contribution facets, researchers can continue to advance the field and address the challenges associated with sentiment analysis at the sentence level. Using the PRISMA framework, this review paper only included the publications written in English, which may lead to an inadequate understanding of the sentiment analysis of non-English texts, for example, the methods for non-English data extraction, cleaning, tokenization and analysis. However, future research in sentence-level sentiment analysis could focus on developing standardized datasets that cover a wider range of domains and languages. Additionally, the exploration of novel evaluation metrics that consider the nuances and complexities of sentiment analysis can further enhance the evaluation of models. Furthermore, investigating advanced techniques, such as deep learning and transfer learning, can improve the performance of sentence-level sentiment analysis models.

In conclusion, this systematic literature review highlights the importance of datasets, evaluation metrics, and contribution facets in sentence-level sentiment analysis. It provides valuable insights for researchers, practitioners, and stakeholders interested in understanding the current state of the field and identifying avenues for future research and development.

REFERENCES

[1] K. R. Haase, T. Cosco, L. Kervin, I. Riadi, and M. E. J. J. a. O'Connell, "Older adults' experiences with using technology for socialization during the COVID-19 pandemic: Cross-sectional survey study," vol. 4, no. 2, p. e28010, 2021.

[2] A. B. Pawar, M. Jawale, D. J. S. a. Kyatanavar, and o. e. A. e. o. c. intelligence, "Fundamentals of sentiment analysis: concepts and methodology," pp. 25-48, 2016.

[3] P. Del Vecchio, A. Di Minin, A. M. Petruzzelli, U. Panniello, S. J. C. Pirri, and I. Management, "Big data for open innovation in SMEs and large corporations: Trends, opportunities, and challenges," vol. 27, no. 1, pp. 6-22, 2018.

[4] B. Bhavitha, A. P. Rodrigues, and N. N. Chiplunkar, "Comparative study of machine learning techniques in sentimental analysis," in *2017 International conference on inventive communication and computational technologies (ICICCT)*, 2017, pp. 216-221: IEEE.

[5] J. Luo, S. Huang, R. J. J. o. H. M. Wang, and Management, "A fine-grained sentiment analysis of online guest reviews of economy hotels in China," vol. 30, no. 1, pp. 71-95, 2021.

[6] V.-D. Păvăloaia, E.-M. Teodor, D. Fotache, and M. J. S. Danileţ, "Opinion mining on social media data: sentiment analysis of user preferences," vol. 11, no. 16, p. 4459, 2019.

[7] A. K. Jain, S. R. Sahoo, J. J. C. Kaubiyal, and I. Systems, "Online social networks security and privacy: comprehensive review and analysis," vol. 7, no. 5, pp. 2157-2177, 2021.

[8] R. S. Kumar *et al.*, "Exploration of sentiment analysis and legitimate artistry for opinion mining," pp. 1-16, 2021.

[9] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," ed: UK, 2007.

[10] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and P. G. J. A. o. i. medicine, "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement," vol. 151, no. 4, pp. 264-269, 2009.

[11] V. Jagtap, K. J. I. J. o. S. E. Pawar, and Technology, "Analysis of different approaches to sentence-level sentiment classification," vol. 2, no. 3, pp. 164-170, 2013.

[12] B. Yang and C. Cardie, "Context-aware learning for sentence-level sentiment analysis with posterior regularization," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014, pp. 325-335.

[13] O. Täckström and R. McDonald, "Semi-supervised latent variable models for sentence-level sentiment analysis," in *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.

[14] M. Araújo, A. Pereira, and F. J. I. S. Benevenuto, "A comparative study of machine translation for multilingual sentence-level sentiment analysis," vol. 512, pp. 1078-1102, 2020.

[15] M. E. Basiri and A. Kabiri, "Sentence-level sentiment analysis in Persian," in *2017 3rd International Conference on Pattern Recognition and Image Analysis (IPRIA)*, 2017, pp. 84-89: IEEE.

[16] A. Shoukry and A. Rafea, "Sentence-level Arabic sentiment analysis," in *2012 international conference on collaboration technologies and systems (CTS)*, 2012, pp. 546-550: IEEE.

[17] T. Hayashi and H. J. A. P. H. Fujita, "Word embeddings-based sentence-level sentiment analysis considering word importance," vol. 16, no. 7, pp. 7-24, 2019.

[18] D. Tang *et al.*, "A joint segmentation and classification framework for sentence level sentiment classification," vol. 23, no. 11, pp. 1750-1761, 2015.

[19] Z. Zhang and M. Lan, "Learning sentiment-inherent word embedding for word-level and sentence-level sentiment analysis," in *2015 International Conference on Asian Language Processing (IALP)*, 2015, pp. 94-97: IEEE.

[20] Y. Long, Y. Li, J. Luo, C. Miao, and J. Fu, "MCP-LSTM Network for Sentence-Level Sentiment Classification," in *2019 International Conference on Virtual Reality and Visualization (ICVRV)*, 2019, pp. 124-128: IEEE.

[21] K. Dashtipour, M. Gogate, A. Gelbukh, and A. Hussain, "Persian sentence-level sentiment polarity classification," in *ICOTEN*, 2021.

[22] U. Farooq, H. Mansour, A. Nongaillard, and M. A. Qadir, "Negation handling in sentiment analysis at sentence level," 2016.

[23] D. Meškelė, F. J. I. P. Frasincar, and Management, "ALDONAr: A hybrid solution for sentence-level aspect-based sentiment analysis using a lexicalized domain ontology and a regularized neural attention model," vol. 57, no. 3, p. 102211, 2020.

[24] V. Shirsat, R. Jagdale, K. Shende, S. N. Deshmukh, S. J. I. J. o. C. S. Kawale, and Engineering, "Sentence level sentiment analysis from news articles and blogs using machine learning techniques," vol. 7, no. 5, pp. 1-6, 2019.

[25] F. Frederick and L. F. Proceso, "Blog Comments Sentence Level Sentiment Analysis for Estimating Filipino ISP Customer Satisfaction," in *International Conference Data Mining, Civil and Mechanical Engineering (lCDMCME, 2015)*, 2015.

[26] H. Peng, E. Cambria, and X. Zou, "Radical-based hierarchical embeddings for Chinese sentiment analysis at sentence level," in *The thirtieth international flairs conference*, 2017.

[27] Z. Hao, R. Cai, Y. Yang, W. Wen, and L. Liang, "A dynamic conditional random field based framework for sentence-level sentiment analysis of Chinese microblog," in *2017 IEEE international conference on computational science and engineering (CSE) and IEEE international conference on embedded and ubiquitous computing (EUC)*, 2017, vol. 1, pp. 135-142: IEEE.

[28] B. Lutz, N. Pröllochs, and D. J. a. p. a. Neumann, "Sentence-level sentiment analysis of financial news using distributed text representations and multi-instance learning," 2018.

[29] A. M. Shoukry, "Arabic sentence-level sentiment analysis," 2013.

[30] X. Fu, W. Liu, Y. Xu, C. Yu, and T. Wang, "Long short-term memory network over rhetorical structure theory for sentence-level sentiment analysis," in *Asian conference on machine learning*, 2016, pp. 17-32: PMLR.

[31] E. J. C. C. R. de Vries, "The Sentiment is in the Details: A Language-agnostic Approach to Dictionary Expansion and Sentence-level Sentiment Analysis in News Media," vol. 4, no. 2, pp. 424-462, 2022.

[32] M. Araujo, J. Reis, A. Pereira, and F. Benevenuto, "An evaluation of machine translation for multilingual sentence-level sentiment analysis," in *Proceedings of the 31st annual ACM symposium on applied computing*, 2016, pp. 1140-1145.

[33] T. I. Jain and D. J. I. J. o. C. A. Nemade, "Recognizing contextual polarity in phrase-level sentiment analysis," vol. 7, no. 5, pp. 12-21, 2010.

[34] M. Araujo *et al.*, "ifeel 2.0: A multilingual benchmarking system for sentence-level sentiment analysis," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2016, vol. 10, no. 1, pp. 758-759.