Social Network Analysis and Mining (2025) 15:45
https://doi.org/10.1007/s13278-025-01461-8

REVIEW PAPER

Generalizing sentiment analysis: a review of progress, challenges, and emerging directions

Khaled Alahmadi[1,2] · Sultan Alharbi[1] · Juan Chen[3] · Xianzhi Wang[1]

Received: 12 November 2024 / Revised: 31 March 2025 / Accepted: 2 April 2025

Abstract
Sentiment analysis is an increasingly vital technique within natural language processing for interpreting human emotions expressed in text. This survey explores the trajectory of sentiment analysis research, examining advancements from traditional machine learning approaches to state-of-the-art deep learning models, including Transformers and hybrid architectures. We highlight key challenges such as domain adaptation, linguistic diversity, and the evolving nuances of digital communication. This review distinguishes itself by adopting a multidisciplinary approach, integrating advancements from machine learning, cognitive science, and linguistics to address generalization, multimodal data integration, and the potential of self-supervised learning. Unlike prior surveys, our work provides a comprehensive synthesis of recent and emerging methodologies, although introduced in previous literature, remain scattered across domain specific studies such as hybrid models combining RoBERTa-GRU and Capsule Networks with semantic rules, while emphasizing ethical considerations and novel directions like adaptive feature selection and fairness-aware training. By providing comprehensive insights into applications across domains like healthcare, finance, and disaster management, this survey serves as a foundational resource for the next generation of sentiment analysis tools.

Keywords Sentiment analysis · Natural language processing · Transfer learning · Generalization techniques · Future directions

1 Introduction

Sentiment analysis is a subfield of natural language processing (NLP) that has emerged as a crucial tool for understanding and interpreting human emotions, opinions, and attitudes expressed in text, speech, or video (Liu 2020). As digital communication continues to thrive, the ability to automatically detect and analyze sentiment has become increasingly valuable across various domains, including business intelligence, social media monitoring, and political analysis. The proliferation of user-generated content on platforms like social media and e-commerce sites has amplified the demand for robust sentiment analysis technologies (Afriliana and Iswari 2022). Applications range from tailoring personalized recommendations and enhancing the efficacy of virtual assistants to informing public policy decisions and refining financial forecasting models. Moreover, sentiment analysis is pivotal in monitoring mental health trends by analyzing social media posts, improving customer service through automated sentiment detection in communications, and optimizing product strategies with feedback analysis (Gupta and Kumar 2023).

The field's journey is fraught with complexities, primarily stemming from the variability and nuance of human emotions, compounded by the dynamic and diverse fabric of language. The challenges of linguistic diversity,

✉ Khaled Alahmadi
khaledmohammadatiqm.alahmadi@student.uts.edu.au

Sultan Alharbi
sultan.a.alharbi@student.uts.edu.au

Juan Chen
jane.chen@sophos.com.au

Xianzhi Wang
xianzhi.wang@uts.edu.au

[1] University of Technology Sydney, Sydney, NSW 2007, Australia

[2] Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Mecca, Saudi Arabia

[3] Sophos Pty. Ltd., North Sydney, NSW 2060, Australia

Published online: 28 April 2025

contextual subtleties, and the continuous evolution of online communication modalities necessitate advanced machine learning techniques capable of navigating these intricacies (Hasan et al. 2019). Understanding and interpreting human emotions through text not only enhances human-computer interactions but also provides valuable insights into public opinions and trends, facilitating better decision-making across various domains (Chatterjee et al. 2019).

Recent advancements in machine learning, particularly deep learning models, have marked a significant leap forward in sentiment analysis. By leveraging sophisticated neural network architectures, these models have demonstrated an ability to grasp nuanced linguistic patterns embedded in text (Dong et al. 2020; Bharti et al. 2022). However, challenges persist, including a reliance on basic features and a predominant focus on English-language datasets, which raises concerns about the generalizability and inclusivity of these models (Oueslati et al. 2020). Addressing these gaps is critical for developing sentiment analysis systems that are linguistically inclusive and culturally sensitive. Efforts to transcend these limitations have catalyzed a multidisciplinary approach, integrating insights from computational linguistics, cognitive science, and social psychology. This review aims to synthesize the current trends, techniques, and challenges in sentiment analysis while identifying opportunities for innovation. By advocating for cross-disciplinary collaboration and focusing on cutting-edge methodologies, such as transfer learning, multimodal data fusion, and self-supervised learning, this survey provides a forward-looking perspective on the evolution of sentiment analysis.

Distinctions from Existing Surveys Several recent surveys have significantly advanced the field of sentiment analysis but have done that from narrowly scoped perspectives. For example, Schouten and Frasincar (Schouten and Frasincar 2015) focused on aspect-level sentiment analysis, particularly for structured review data. while (Yue et al. 2019) emphasized sentiment analysis in social media, addressing informal language and real-time dynamics, highlighting challenges like informal language and evolving online slang. Liu et al. (2020) centred on deep learning approaches for sentiment analysis, while (Oueslati et al. 2020) explored challenges in Arabic-language sentiment analysis. While each offers meaningful contributions, they tend to be domain-specific, modality-limited, or technologically constrained, and do not fully address the broader challenge of building generalizable sentiment analysis systems. In contrast, this survey adopts a generalization-focused and multidisciplinary approach, integrating perspectives from machine learning, linguistics, and cognitive science. We systematically examine adversarial robustness, cross-domain adaptability, multilingual inclusion, and ethical fairness dimensions that are largely absent or underexplored in prior surveys. These areas are critical for building sentiment analysis models that can perform reliably across different domains, languages, and modalities. By emphasizing generalization and inclusivity, our work fills a crucial gap in the literature and serves as a forward-looking resource that synthesizes both foundational techniques and emerging directions in sentiment analysis research.

Our Contributions This survey examines sentiment analysis through a comprehensive lens, covering the progression from early machine learning techniques to cutting-edge methodologies like hybrid models and transformer architectures. By addressing linguistic diversity, domain adaptation, and ethical challenges, it presents a detailed synthesis of existing research. Furthermore, this review emphasizes the importance of cross-disciplinary approaches and identifies future directions, such as adaptive feature selection and fairness-aware training, to foster innovation in sentiment analysis.

2 State-of-the-art approaches

This review is structured to reflect both the historical development and the methodological evolution of sentiment analysis techniques. We begin by discussing traditional machine learning and lexicon-based approaches, which laid the groundwork for sentiment classification. We then progress to more advanced methods, including deep learning, transfer learning, and hybrid models. This sequential structure mirrors the field's trajectory toward increasingly sophisticated and generalizable models. Later sections (Section 3 onward) focus specifically on the challenges to generalization such as domain shifts, multilingual complexity, and model robustness and the emerging techniques designed to address them.

2.1 Traditional machine learning approaches

Traditional machine-learning approaches have played a pivotal role in the early development of sentiment analysis. These methods typically involve extracting features from text data, such as word frequencies or patterns, and then applying classification algorithms to predict sentiment. This general procedure lays the groundwork for more advanced techniques to follow.

Specifically, the machine learning process (illustrated in Fig. 1) involves several key steps: (1) data acquisition, where raw data is collected from various sources; (2) data preprocessing, which cleans, transforms, and reduces data to prepare it for analysis; (3) model development, where an algorithm is selected and trained on the data; and (4) model evaluation and testing, which assesses the model's

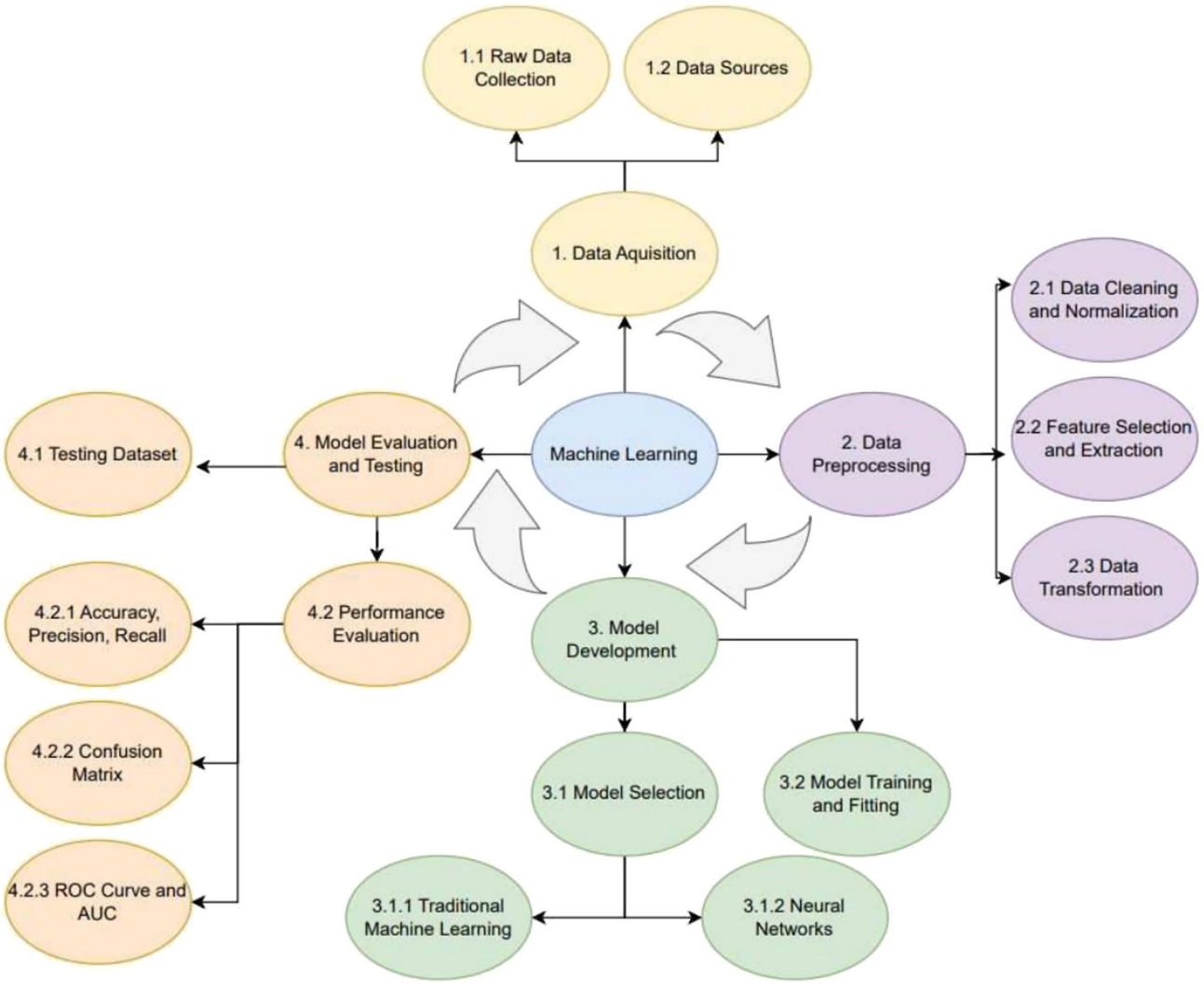

Fig. 1 Machine learning processes

performance using metrics and ensures it is generalized to new data. These steps are widely recognized as standard practices in the machine learning pipeline, as described in foundational works on the subject (Bishop and Nasrabadi 2006).

Early studies relied on simple bag-of-words features processed by Support Vector Machine (SVM) classifiers, achieving a robust **82.9% accuracy** on the IMDb dataset, setting a strong baseline for binary positive/negative predictions (Pang et al. 2002). These methods demonstrated efficiency but struggled with intricate linguistic phenomena such as sarcasm and negation. Optimizations in SVMs through genetic algorithm-based support vector selection further improved their performance, achieving **80.6% accuracy** and an F1-score of **71.4%** on mobile phone reviews (Fei 2016). Random Forest classifiers emerged as another strong approach, excelling on imbalanced datasets like product reviews (digital camera, mobile phone, laptop) with an AUC-ROC of **83.0%** (Gopalakrishnan and Ramaswamy 2014).

Recent advancements included hierarchical models such as Recursive Neural Tensor Networks (RNTN) applied to the Stanford Sentiment Treebank (SST). These models excelled in binary sentiment classification with an accuracy of **85.4%** at the root level but struggled with fine-grained classification, achieving **80.7%** accuracy (Socher et al. 2013). Hybrid approaches combining Naïve Bayes and SVM further pushed performance boundaries, achieving **88.94% accuracy** on IMDb by integrating efficient n-gram feature selection with robust classification (Tripathy et al. 2016).

The results summarized in Table 1 summarizes the machine learning techniques and datasets used in these studies, along with their results.

Table 1 Performance of traditional machine learning techniques on benchmark datasets

References	Technique	Dataset	Accuracy (%)	F1-Score (%)
Pang et al. (2002)	Naïve Bayes	IMDb	78.7	–
Pang et al. (2002)	SVM	IMDb	82.9	–
Fei (2016)	Genetic Algorithm-Optimized SVM	Mobile Phone Reviews	80.6	71.4
Gopalakrishnan and Ramaswamy (2014)	Modified Bagging	Product Reviews (Imbalanced)	AUC-ROC: 83.0	–
Socher et al. (2013)	Recursive Neural Tensor Network (Binary)	Stanford Sentiment Treebank (SST)	85.4	–
Socher et al. (2013)	Recursive Neural Tensor Network (Fine-Grained)	Stanford Sentiment Treebank (SST)	80.7	–
Tripathy et al. (2016)	Hybrid (Naïve Bayes + SVM)	IMDb	88.94	–

F1-scores are reported where available

The machine learning techniques outlined in Table 1 demonstrate significant advancements in sentiment analysis. Naïve Bayes classifiers excel in handling large datasets with simplicity and computational efficiency, but their reliance on feature independence assumptions can lead to biases toward dominant classes. SVMs, with their robust performance on small and medium-sized datasets like IMDb, effectively capture non-linear decision boundaries but are computationally intensive and require careful parameter tuning. Ensemble methods, such as Modified Bagging, address imbalanced datasets effectively by aggregating multiple decision trees but often lack interpretability.

Recursive neural tensor networks, applied to datasets like SST, highlight the importance of compositional semantics in sentiment analysis, offering strong binary classification performance. However, their fine-grained classification accuracy remains limited due to high computational demands and sensitivity to input structure. Hybrid approaches, such as those combining Naïve Bayes and SVM, leverage the strengths of both techniques, demonstrating the value of integrated feature selection and classification. The performance of traditional models varies significantly depending on feature selection and dataset structure. SVMs outperform Naïve Bayes on datasets like IMDb due to their ability to model non-linear boundaries, especially when high-dimensional sparse features are present. Hybrid approaches that combine Naïve Bayes and SVM benefit from the former's efficiency and the latter's precision, resulting in superior accuracy. However, these models still struggle with linguistic nuances like sarcasm or negation, which limits their fine-grained classification capabilities. Despite these advancements, these methods generally rely on predefined features and struggle to adapt to new domains or incorporate semantic knowledge effectively.

2.2 Rule-based and Lexicon-based approaches

Rule-based and lexicon-based approaches represent some of the earliest methods in sentiment analysis. These methods are celebrated for their simplicity, interpretability, and computational efficiency. Rule-based approaches rely on predefined linguistic rules to classify sentiment, while lexicon-based techniques utilize dictionaries of sentiment-labeled words to assign sentiment scores. Despite their foundational role, these methods are often limited in their ability to handle complex linguistic phenomena, such as sarcasm, negation, and ambiguous expressions. For instance, sarcasm often contradicts literal sentiment, rendering rule-based approaches ineffective (Filatova 2012). Similarly, negation can reverse the sentiment of words, requiring sophisticated handling beyond simple rules (Councill et al.

Table 2 Performance of Rule-based and Lexicon-based approaches on benchmark datasets

References	Technique	Dataset	Accuracy (%)	F1-Score (%)
Hutto and Gilbert (2014)	VADER (Lexicon-based)	Twitter Data	–	96.0
Srivastava et al. (2022)	SentiStrength Lexicon	Sentiment140 Dataset	81.70	–
Iqbal et al. (2019)	Lexicon + ML (Genetic Algorithm Feature Reduction)	Twitter Sentiment Dataset	79.00	–
Siddiqua and Ahsan (2016)	Rule-based + Weakly Supervised Learning	Stanford Sentiment140	84.96	86.08
Fadel and Öz (2020)	Hybrid Lexicon + ML	Twitter (Terrorism Sentiment)	94.80	96.20 (pos), 91.60 (neg)

F1-scores are reported only where available in the referenced studies

2010). These limitations make rule-based methods less reliable for nuanced sentiment analysis tasks.

As illustarated in Table 2, VADER (Hutto and Gilbert 2014), a lexicon-based model, has proven effective in handling sentiment in social media contexts. By integrating linguistic and grammatical heuristics, it achieved an impressive **96.0% F1-score** on Twitter datasets, outperforming several human raters. This demonstrates the potential of rule-based models when tailored to specific domains. Similarly, SentiStrength (Srivastava et al. 2022) demonstrated its efficacy in analyzing customer feedback, achieving an accuracy of **81.70%** on sentiment classification tasks for the Sentiment140 dataset. SentiStrength also incorporates rules to account for negation and amplify sentiment intensity, which adds a degree of flexibility over traditional lexicon-based methods.

Hybrid frameworks have emerged as a natural extension of these approaches, combining lexicon-based methods with machine learning to overcome the inherent limitations of static rules. Iqbal et al. (2019) proposed a genetic algorithm-based feature reduction framework that integrates lexicon-based methods with supervised classifiers, achieving **79.0% accuracy** on the Twitter Sentiment dataset. This approach demonstrates the potential for combining the interpretability of lexicons with the predictive power of machine learning models.

In the context of social media sentiment analysis, Siddiqua and Ahsan (2016) combined rule-based classifiers with weakly supervised learning to enhance performance on the Stanford Sentiment140 dataset. Their approach achieved an impressive **84.96% accuracy** and an **86.08% F1-score**, showcasing the value of integrating rule-based systems with adaptive learning techniques for large-scale datasets.

An application of hybrid methods in a critical domain was presented by Fadel and Öz (2020), who analyzed Twitter sentiment related to terrorist attacks. By integrating lexicon-based analysis with machine learning, their model achieved **94.8% accuracy** and an **F1-score of 96.2% (positive)** and **91.6% (negative)**, highlighting the applicability of these techniques to domains requiring nuanced sentiment understanding.

Although rule-based and lexicon-based approaches are computationally efficient and excel in domains with limited training data, they are inherently constrained by their reliance on static lexicons and predefined rules. These methods often fail to generalize to novel datasets and struggle with the semantic subtleties of natural language, such as idiomatic expressions and context-specific sentiment. Another challenge with lexicon-based approaches is maintaining and updating sentiment lexicons to accommodate domain-specific vocabulary and evolving language usage. For instance, terms that carry specific sentiment in one context may have a neutral meaning in another (Taboada et al. 2011). The static nature of most lexicons makes them unsuitable for rapidly changing domains, such as social media or emerging technologies (Mohammad and Turney 2013).

These methods in Table 2 provide valuable insights, especially in contexts where labeled training data is scarce or domain-specific customization is essential. Models like VADER and SentiStrength perform well on social media due to their heuristic rules tailored for informal language. However, their static lexicons limit adaptability. Hybrid models, such as those by Siddiqua and Ahsan (2016). and Fadel and Öz (2020), integrate rule-based features with machine learning to improve performance. The approach by Fadel et al. [26] achieved the highest accuracy and F1-score in this group by combining lexicon-based analysis with domain-specific training on terrorism-related tweets, demonstrating the value of contextual adaptation in hybrid frameworks.

However, as datasets grow larger and more diverse, the limitations of rule-based and lexicon-based methods become evident. These challenges have catalyzed the adoption of deep learning techniques, which leverage automated feature extraction and semantic understanding to achieve superior performance.

2.3 Deep neural networks

Deep neural networks (DNNs) have brought significant advancements in sentiment analysis by enabling the automatic learning of rich representations from text. Various architectures excel in modeling different linguistic features, making them powerful tools for sentiment classification. Figure 2 illustrates a typical DNN architecture for sentiment analysis, providing a visual representation of how different layers interact to process input text.

Convolutional Neural Networks (CNNs) have proven effective in extracting local semantic patterns from text. Kim (2014) demonstrated the effectiveness of CNN architectures for sentence classification. Multi-channel CNNs (MVCNN), proposed by Yin and Schütze (2016), further advanced this approach by capturing both word- and sentence-level features, achieving **88.2%** accuracy on Senti140 and **93.9%** on Subj. Recurrent Neural Networks (RNNs), especially Long Short-Term Memory networks (LSTMs), excel at capturing sequential dependencies in text. Bharti et al. (2022) achieved **88.08% accuracy** on Twitter sentiment classification using an optimized LSTM architecture. Tree-structured LSTMs (Tai et al. 2015) improved upon traditional LSTMs by incorporating hierarchical sentence structures, scoring **88.0% accuracy** on binary classification tasks for the Stanford Sentiment Treebank. Capsule networks, a novel deep learning architecture, capture hierarchical relationships in data for nuanced sentiment analysis. Dong et al. (2020) proposed a capsule network model with BiLSTM (caps-BiLSTM), achieving **91.96% accuracy** on IMDB. Similarly, Wang

et al. (2018) introduced RNN-Capsule, which achieved **91.6% accuracy** on a hospital feedback dataset. Hybrid deep learning models that integrate CNNs, LSTMs, and attention mechanisms have further enhanced performance. Chen et al. (2018) proposed a CNN-LSTM model that achieved **78.42% accuracy** and an F1-score of **78.42%** on a multi-class sentiment dataset. Attention-based LSTMs, like ATAE-LSTM (Wang et al. 2016), improved sentiment analysis by focusing on sentiment-laden words, achieving **89.9% accuracy** on binary classification for the SemEval 2014 dataset. Recent developments include integrating Transformer-based models, such as BERT, which set new benchmarks in transfer learning for downstream sentiment analysis tasks (Abdullah and Ahmet 2022). These advancements indicate a trend toward hybrid architectures and attention mechanisms for improving sentiment classification performance.

The results summarized in Table 3 highlight the diverse performance of deep learning architectures across a range of datasets. CNN-based models, such as those proposed by Kim (2014) and Yin and Schütze (2016), demonstrated strong accuracy on datasets like MR and Subj by effectively capturing localized semantic patterns. RNN-based architectures, particularly LSTMs, excelled in processing sequential data, achieving high accuracy on Twitter and binary sentiment classification tasks (Bharti et al. 2022; Tai et al. 2015). Capsule networks, a more recent innovation, showed robust performance on IMDB and hospital feedback datasets, reflecting their strength in modeling hierarchical relationships (Dong et al. 2020; Wang et al. 2018).

Hybrid models, such as CNN-LSTM (Chen et al. 2018) and attention-based LSTMs like ATAE-LSTM (Wang et al. 2016), further enhanced classification accuracy by combining local feature extraction with sequential and context-aware processing. Attention mechanisms, in particular, proved effective in aspect-based sentiment tasks by focusing on sentiment-relevant words.

Overall, the performance of these models reflects their adaptability to specific task requirements and dataset structures. LSTMs and Tree-LSTMs are well-suited for structured datasets like SST and Twitter due to their capacity to capture sequential and hierarchical patterns. Capsule networks are particularly effective in datasets with rich compositional sentiment, such as IMDB. CNNs and CNN-LSTM hybrids perform well when both local and contextual understanding is required. These findings illustrate the trade-offs between architectural complexity and performance, and underscore the scalability of deep learning approaches in handling the nuanced challenges of sentiment analysis.

2.4 Transfer learning

Transfer learning has emerged as a pivotal technique in sentiment analysis, enabling models to leverage knowledge from one domain and apply it to another, often with limited labeled data. This approach is particularly beneficial where sentiment expressed in text varies significantly across domains or languages. Transfer learning, often based on deep neural networks, enables the use of pre-trained models to address sentiment classification challenges across diverse

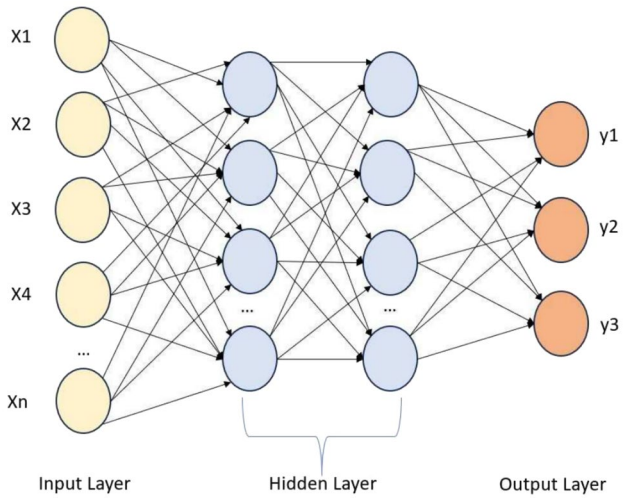

Fig. 2 Deep neural network architecture

Table 3 Performance of deep learning models for sentiment analysis

References	Techniques	Dataset	Accuracy (%)	F1-Score (%)
Bharti et al. (2022)	LSTM	Twitter dataset	88.08	–
Dong et al. (2020)	Capsule network (caps-BiLSTM)	MR, IMDB, SST	81.47 (MR)91.96 (IMDB)48.34 (SST)	–
Yin and Schütze (2016)	Multi-channel CNNs (MVCNN)	SST-1, Senti140, Subj	88.2 (Senti140)93.9 (Subj)	–
Tai et al. (2015)	Tree-LSTM	SemEval 2014, Stanford Sentiment Treebank	88.0 (Binary)51.0 (Fine-grained)	–
Wang et al. (2018)	RNN-Capsule	MR, SST, Hospital Feedback	83.8 (MR)91.6 (Hospital Feedback)	–
Chen et al. (2018)	CNN-LSTM (proposed model)	Multi-class dataset	78.42	78.42

F1-scores are reported only where available in the referenced studies

domains effectively. This technique involves pre-training a deep neural network model on a large dataset (source domain) and fine-tuning it on a smaller, domain-specific dataset (target domain). This process enables the model to adapt to new contexts with minimal additional training, leveraging deep neural networks' capability to extract complex features.

Several studies highlight the effectiveness of transfer learning for sentiment analysis. Zhu et al. (2022) demonstrated its application in biomedical question answering by fine-tuning BioBERT, T5, and RoBERTa models. Their approach achieved an accuracy of **83.91%** and an F1-score of **76.92%**, outperforming state-of-the-art methods and human annotators by significant margins. Prottasha et al. (2022) explored transfer learning for Bangla sentiment analysis using a BERT-based CNN-BiLSTM model, achieving an accuracy of **94.15%** and an F1-score of **93.04%**. These results underscore the potential of transfer learning in enhancing sentiment analysis for underrepresented languages. Sahar et al. (2022) leveraged BERT for sentiment analysis of cosmetics product reviews, achieving an impressive accuracy of **93.21%**, highlighting the model's adaptability to domain-specific tasks with minimal labeled data.

Khan and Shahid (2022) applied BiLSTM with Self-Attention and Joint Dual Input Learning (JDIL) for sentiment analysis in Hindi and Bengali languages. Their model achieved an accuracy of **78%** and an F1-score of **78%** for Bengali and **76%** accuracy for Hindi, emphasizing the utility of joint learning in multilingual sentiment analysis. Ahmed et al. (2024) extended the application of transfer learning to Urdu sentiment analysis, achieving an impressive accuracy of **96%** and an F1-score of **91%** using an LSTM model. Their study highlighted the effectiveness of manual annotation and inter-annotator reliability for building reliable sentiment datasets in low-resource languages.

The summarized results in Table 4 highlight the diverse applications of transfer learning in sentiment analysis across languages and domains. Zhu et al. (2022) achieved notable improvements in biomedical QA tasks by integrating sentiment information through fine-tuning BioBERT, T5, and RoBERTa, setting new benchmarks in domain-specific sentiment modeling. Prottasha et al. (2022) demonstrated the adaptability of BERT for Bangla sentiment analysis, achieving state-of-the-art performance and showcasing transfer learning's potential for underrepresented languages. Sahar et al. (2022) applied transfer learning to the cosmetics domain, achieving high accuracy with minimal labeled data. Similarly, Ahmed et al. (2024) achieved the highest reported accuracy and F1 scores in Urdu sentiment analysis, underscoring the model's effectiveness in low-resource settings. Khan and Shahid (2022) demonstrated the usefulness of BiLSTM architectures in multilingual sentiment analysis, particularly for Hindi and Bengali, though the lower accuracy suggests challenges in managing linguistic diversity.

These findings affirm the effectiveness of transfer learning in leveraging pre-trained models such as BERT, BioBERT, and RoBERTa-not only for general-purpose NLP tasks, but also for domain-specific and multilingual sentiment analysis. Their ability to generalize across tasks is driven by pre-training on large-scale corpora, while fine-tuning enables adaptation to specific domains and languages. As shown across the surveyed studies, this combination yields strong results, especially in low-resource contexts. Continued advancements in transfer learning, combined with techniques like attention mechanisms and hybrid architectures, are expected to further improve sentiment classification performance across domains.

Multilingual Transfer Learning Multilingual capabilities play a pivotal role in generalizing machine learning models for sentiment analysis by broadening their applicability across a diverse linguistic landscape. The shift towards incorporating multiple languages in sentiment analysis research addresses the critical need to transcend the limitations imposed by focusing predominantly on English. This expansion is crucial for creating models that are both inclusive and reflective of the global diversity of language and sentiment expression.

Recent studies underscore the significance of devising sentiment analysis models capable of effective operation across multiple languages. Badawi (2023) investigated the application of multilingual BERT for sentiment classification on the Kurdish Medical Corpus, as represented in Table 5, achieving an accuracy of **92%** and an F1-score of **92%**. Their

Table 4 Performance of transfer learning models for sentiment analysis

References	Techniques	Dataset	Accuracy (%)	F1-Score (%)
Zhu et al. (2022)	BioBERT, T5, RoBERTa	PubMedQA	83.91	76.92
Prottasha et al. (2022)	BERT, CNN-BiLSTM	Bangla Sentiment Dataset	94.15	93.04
Sahar et al. (2022)	BERT	Cosmetics Reviews	93.21	93.00
Khan and Shahid (2022)	BiLSTM + JDIL	HASOC (Hindi, Bengali)	78.00 (Bengali)	78.00
Ahmed et al. (2024)	LSTM	Urdu Sentiment Dataset	96.00	91.00

Table 5 Performance of multilingual transfer learning models in sentiment analysis

References	Techniques	Dataset	Accuracy (%)	F1-Score (%)
Badawi (2023)	BERT-Multilingual	Kurdish Medical Corpus	92.00	92.00
Rathod et al. (2022)	XLM-RoBERTa	L3CubeMahaSent (Marathi Twitter)	83.82	–
Draskovic et al. (2022)	NB-LR Hybrid Model	Serbian Music Reviews	79.00	–
Rastogi (2023)	BERT with Weak Supervision	English, Hindi, Urdu Twitter Data	–	91.00

F1-scores are reported only where available in the referenced studies

findings highlighted the capability of transformer-based architectures to bridge the resource gap for underrepresented languages. Similarly, Rathod et al. (2022) tackled sentiment analysis challenges in Marathi using XLM-RoBERTa. Their study achieved **83.82%** accuracy on the L3CubeMahaSent dataset, underscoring the potential of large multilingual transformer models to perform effectively in low-resource settings.

Draskovic et al. (2022) introduced a hybrid model combining Naive Bayes with Logistic Regression (NB-LR) for sentiment analysis in Serbian. The model achieved an accuracy of **79%** for binary classification on the Music Reviews dataset, emphasizing the importance of hybrid approaches in addressing low-resource language challenges. Lastly, Rastogi (2023) explored multilingual sentiment analysis for English, Hindi, and Urdu using weak supervision and BERT. Their model achieved an F1-score of **91%** for positive sentiments on translated datasets, highlighting the effectiveness of weak supervision in managing multilingual data and the robustness of transformer-based models in multilingual sentiment scenarios.

The summarized results in Table 5 highlight the growing effectiveness of multilingual transfer learning approaches. For instance, the hybrid sentiment analysis model for Serbian datasets achieved notable accuracy for binary classification, while BERT-based models demonstrated robust performance across diverse languages like Kurdish and Hindi. These findings emphasize the adaptability of transformer architectures to multilingual sentiment analysis challenges.

2.5 Comparative summary of sentiment analysis approaches

To provide a holistic perspective on the state-of-the-art techniques in sentiment analysis, Table 6 presents a comparative overview of the aforementioned four primary categories traditional machine learning, rule and lexicon-based approaches, deep learning models, and transfer learning methods. This summary highlights key trade-offs in terms of accuracy, computational efficiency, explainability, data requirements, and generalization capability.

Traditional machine learning models, such as SVM and Naïve Bayes, are computationally efficient and interpretable but rely heavily on manual feature engineering and struggle with linguistic nuance. Rule-based and lexicon-based approaches offer transparency and low resource requirements, however their static nature limits scalability and generalization to evolving language.

Deep learning models have shown significant gains in sentiment classification accuracy by learning hierarchical

Table 6 Comparative summary of sentiment analysis approaches

Aspect	Traditional ML	Rule / Lexicon-based	Deep learning	Transfer learning
Accuracy	Moderate; depends on dataset and feature engineering	Moderate in domain-specific tasks	High with sufficient labeled data	Very high across domains and tasks
Computational cost	Low	Very low	High, especially during training	Very high due to pretraining and fine-tuning
Explainability	Moderate (e.g., SVM weights)	High (transparent, rule-based logic)	Low; typically black-box	Low; often difficult to interpret
Data requirements	Requires labeled training data	No training data; depends on lexicon quality	Requires large labeled datasets	Effective with few labeled examples
Domain adaptability	Limited without retraining	Weak due to static lexicons	Moderate with domain-specific tuning	Strong; adaptable through fine-tuning
Robustness	Sensitive to noise and domain shift	Poor for dynamic or informal language	Improved with hybrid or regularized models	High; handles multilingual and domain-shifted data well

and semantic representations directly from data. However, they require considerable labeled datasets and computational resources and are generally nontransparent in decision-making. Transfer learning techniques, particularly those using pre-trained transformer-based models (e.g., BERT, RoBERTa), have pushed state-of-the-art results by enabling robust, cross-domain performance with minimal supervision. Their primary drawbacks lie in the high cost of pretraining and reduced interpretability.

This comparative overview helps identify the operational strengths and limitations of each approach, offering practical guidance when selecting models for specific sentiment analysis tasks and contexts.

3 Generalizing sentiment analysis

Generalization in the context of machine learning and sentiment analysis refers to the ability of a model to perform well on new, previously unseen data beyond the specific examples used during training. This concept is critical because it determines the robustness and applicability of a model in real-world scenarios. A model that generalizes well can accurately analyze sentiments across different texts, domains, and contexts without retraining.

The need for generalization in sentiment analysis is underscored by the dynamic nature of language used on digital platforms. New slang, expressions, and even languages emerge regularly, and sentiment-laden text often includes irony, sarcasm, and context-dependent meanings. Models trained on too specific or outdated datasets may perform poorly when exposed to such evolving language patterns. Therefore, enhancing the generalization capability of sentiment analysis models ensures they remain effective as language use changes over time.

In this section, we explore the challenges of generalizing sentiment analysis, including overfitting, domain differences, and sparsity, followed by a discussion on strategies such as parameter regularization, transfer learning, multi-task learning, and ensemble methods to enhance model robustness and applicability across diverse datasets.

3.1 Challenges

Successfully generalizing sentiment analysis models requires overcoming several key challenges with varying impacts depending on the context. Among them, overfitting is a generic challenge affecting all areas of machine learning, where models perform well on training data but poorly on unseen data. This issue is critical in sentiment analysis, where models must generalize across varied linguistic expressions not covered during training. Domain differences present a challenge that is particularly significant to sentiment analysis. Sentiment expression can vary dramatically across domains, such as social media, customer reviews, or news articles, each with unique linguistic styles and contextual nuances. Sparsity in data, characterized by high-dimensional feature spaces with relatively few data points, is another challenge that, while common in many machine-learning contexts, poses specific difficulties in sentiment analysis due to the subtle and diverse ways emotions are conveyed in the text.

3.1.1 Overfitting

Overfitting occurs when machine learning models become too specialized in the quirks and noise of their training data, making them fail to work with new data. Complex deep neural networks are flexible models with many parameters that can easily overfit by memorizing fine patterns in training data. This leads to two problems: 1) the model may perform poorly on new data because it relies on superficial cues rather than robust representations. 2) the model may show artificially high performance on training data, which does not reflect its usefulness in real-world applications.

Overfitting makes models focus on spurious correlations instead of grasping the essence of the data. It's a persistent challenge because language is complex, and training data is limited. Simpler models are prone to underfitting and cannot represent nuanced language phenomena. This trade-off between simplicity and complexity is an ongoing challenge across diverse tasks and data distributions (Reimers and Gurevych 2017). Inductive biases from transfer learning and architectural constraints offer some protection against overfitting to spurious correlations. However, modern networks rely heavily on surface-level cues rather than in-depth understanding (Niven and Kao 2019). For instance, an end-to-end neural natural language interface model sometimes uses dataset biases instead of genuinely understanding meaning and semantics for mapping to database queries (Utama et al. 2018).

Various strategies have been proposed to address overfitting. These include early stopping to prevent overfitting by halting training before performance optimization stops (Ying 2019), network reduction to exclude noise in training data, data expansion for fine-tuning hyperparameters in complex models, and regularization to select more useful features (Trivedi et al. 2021). Sparse Bayesian Learning with weakly informative hyperpriors and extended predictive information criteria have also been explored to prevent overfitting in regression tasks (Murayama and Kawano 2021). Careful regularization, data augmentation, and thorough evaluation are needed to ensure models learn broadly applicable representations not tied to narrow training data. For instance,

Hayatin et al. (2022) examine the efficacy of different Naïve Bayes models in sentiment analysis, particularly focusing on their performance with small datasets to combat overfitting. The study finds the Complement Naïve Bayes model is particularly effective, achieving an F1-score of approximately 0.82, underscoring the potential of probabilistic approaches in mitigating overfitting on sentiment analysis tasks. In a related vein, Ahmed et al. (2022) conduct a comprehensive review and assessment of feature selection techniques, including TF-IDF, document frequency, word frequency, sparsity reduction, and chi-square statistics, within the context of sentiment analysis. This study emphasizes on the critical role of effective feature selection in overcoming data sparsity and enhancing model performance. Furthermore, Jin and Zhang (2022) introduce a novel filter feature selection method, the Discriminant Variance Criterion (DVC), designed for sentiment analysis. By leveraging feature and class variances, DVC effectively selects discriminative features, thereby addressing the challenges associated with high dimensionality and sparsity of text data.

3.1.2 Domain differences

Sentiment expression varies across different text domains due to nuanced socio-cultural norms. For example, reviews, social media, and news have distinct language patterns. When models are trained on one domain, they often rely on surface-level cues that correlate with sentiment in that domain. However, these cues may not transfer well to new domains. For instance, reviews tend to contain more explicit sentiments than social media posts. A model trained solely on reviews struggles when applied to social media since it cannot rely on the same lexical cues (Minanovic et al. 2014). Multi-task learning helps by jointly training on diverse domains with varying data distributions (Liu et al. 2019). This exposes models to different training signals. However, challenges remain in bridging the domain gap. Additional methods to address domain differences include adapting sentiment lexicons to domain-specific social media texts (Deng et al. 2017), learning disentangled representations for multimodal cross-domain sentiment analysis (Zhang et al. 2022), and user-guided cross-domain sentiment classification (Nelakurthi et al. 2017). Techniques like cross-domain sentiment classification based on key sentiment sentence extraction (Zhang et al. 2015) and using word embeddings with canonical correlation analysis (Bach et al. 2016) have also been explored. Despite the advancements in addressing domain differences in sentiment analysis, this area remains a challenging and active field of research. The complexity of language and the dynamic nature of sentiment expression across various domains continue to pose significant hurdles Nevertheless, fundamental improvements in transfer learning are necessary for sentiment models to generalize effectively across diverse text domains.

3.1.3 Sparsity

Sparsity refers to the fact that text embeddings used to represent words and sentences have a large number of dimensions, often in the millions. However, each text example only contains meaningful information in a small subset of these dimensions. As a result, the text representation becomes sparse, with most dimensions having zero values and only a few dimensions having relevant information for a specific example. This is similar to searching for a needle in a haystack, where the crucial information is scattered across a vast space. This sparsity poses a challenge for models as it makes it hard to learn useful patterns. The important semantic connections often get lost in the noise of the high-dimensional space. Models can easily overfit by latching onto chance correlations in the training data that do not generalize well. Text embedding spaces in modern NLP models have millions of semantic features, resulting in an extremely sparse representation space (Bengio et al. 2013). The essential relationships between words and concepts are spread thinly across dimensions, making it challenging to find meaningful patterns. Standard learning algorithms struggle to uncover relevant patterns in such sparse, high-dimensional spaces. Effective capacity control through techniques like dropout can help prevent overfitting to spurious correlations (Nitish 2014). However, striking the right balance between underfitting the complexity of language and overfitting superficial patterns remains a subtle challenge. Various methods have been proposed to address the issue of sparsity in text embeddings. For example, context-dependent models have been developed for spam detection on social networks, leveraging the BERT model for a context-dependent representation of text (Ghanem and Erbay 2020). Joint text embedding models have been proposed for personalized content-based recommendation, combining text embedding with personalized recommendation (Chen et al. 2017). Additionally, methods like interpretable adversarial training for text have been introduced, using sparse projected gradient descent (SPGD) to craft interpretable adversarial examples for text (Barham and Feizi 2019). Text-graph enhanced knowledge graph representation learning has been explored to address structure sparsity in knowledge graphs by incorporating auxiliary texts of entities (Hu et al. 2021).

3.2 Generalization strategies

Transitioning from the challenges of overfitting, domain differences, and sparsity, we explore strategies to enhance the generalization of sentiment analysis models. These approaches, including Parameter Regularization, Data

Augmentation, Transfer Learning, Multi-Task Learning, and Ensemble Methods, are designed to build robust models that adapt effectively to varied linguistic contexts and evolving expressions of sentiment.

3.2.1 Parameter regularization

Regularization remains crucial to prevent overfitting and enhance generalization in overparameterized neural networks. Combining multiple complementary regularization techniques further improves model robustness. Regularization techniques are essential for constraining model complexity and preventing overfitting. These methods include L2 regularization, which imposes penalties on weight magnitudes, encouraging smaller weights (Krogh and Hertz 1991). Another approach is dropout, which randomly deactivates hidden units during training to prevent co-adaptation (Nitish 2014). Data augmentation involves expanding training datasets by introducing perturbations like noise injection. This helps enhance the model's robustness to various input patterns (Wei and Zou 2019). Recent research explores innovative regularization techniques rooted in information theory. One such approach is entropy regularization, which penalizes low-entropy output distributions, indicating overconfidence (Grandvalet and Bengio 2006). Another technique is mutual information regularization, which reduces shared information between outputs and distorted inputs (Belghazi et al. 2018). Bayesian methods model weight distributions and regularize networks through posterior inference. An example is applying dropout during testing for Monte Carlo model averaging, which enhances uncertainty estimation (Gal and Ghahramani 2016). Pruning is another technique that eliminates redundant parameters by sparsifying weights or removing channels. The "lottery ticket hypothesis" identifies subnetworks that achieve full model performance through iterative magnitude pruning (Frankle and Carbin 2018).

3.2.2 Data augmentation

Data augmentation is a valuable strategy for expanding training datasets through input perturbations, such as random insertion, deletion, swapping, and substitution. This exposes models to more linguistic variability and mitigates overfitting (Wei and Zou 2019). Simple augmentation methods encompass synonym replacement, random insertion, and mix-up interpolation between examples (Guo 2020). More advanced techniques, like contextual augmentation, mask and reconstruct portions of the input based on context, thereby enhancing robustness to perturbations (Kobayashi 2018). Generative approaches, such as Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs), augment data by generating examples that follow the training data distribution (Bowman et al. 2016). Reinforcement learning explores optimizations, such as maximizing rewards based on validation accuracy, to generate augmented data and improve performance selectively. Back-translation employs machine translation to create paraphrased versions of the original training data (Corbeil and Ghadivel 2020). This exposes models to syntactic and lexical diversity. Another approach is adversarial learning, which is a strategy to enhance the model's resilience by exposing blind spots through deliberate input perturbation. Small adversarial changes like inserted or swapped words can dramatically flip predictions, revealing limitations (Jia and Liang 2017). This motivates training models to be robust to such perturbations. Whitebox adversarial examples assume full knowledge of the model to craft the worst-case perturbations. Gradient-based attacks like Fast Gradient Sign Method FGSM use the model's own gradients to find small damaging modifications within an allowed norm ball (Goodfellow et al. 2014). The FGSM efficiently utilizes the gradient sign directly. Iterative attacks like Projected Gradient Descent PGD extend this by iteratively employing it with incremental adjustments. Adversarial training fine-tunes the model on such examples, optimizing for worst-case robustness. For text, gradient signals are sparse, so whitebox attacks are less effective. Blackbox approaches like synonym replacement generate adversarial examples by substituting words with semantic equivalents without gradients. Models fine-tuned on these unambiguous perturbations learn to rely less on brittle surface word statistics (Alzantot et al. 2018). However, discrete texts lack the continuous input space for gradient-based attacks. Generative approaches like SeqGAN train adversarial example generators by modeling text perturbation as a reinforcement learning task (Yu et al. 2017). The target classifier provides the reward signal for the generator, which is co-trained to fool the classifier. This approach does not require gradient access but is less efficient.

Recent work explores constraint-based textual adversarial attacks that more subtly flip sentiment by inserting or modifying phrases within syntactic constraints, revealing model biases (Iyyer et al. 2018). Human studies find such perturbations increase psychological realism compared to prior lexical substitutions. While adversarial training provides robustness, models remain vulnerable to adversarial attacks designed to bypass their defences (Jin et al. 2020). Developing truly robust NLP models remains an open challenge.

3.2.3 Transfer learning

Transfer learning techniques are pivotal in initializing deep neural networks for NLP tasks. The process involves pre-training on large generic corpora before fine-tuning for specific tasks, providing strong inductive biases and mitigating overfitting, thereby enabling the training of deep

models. Early transfer learning methods, such as word embeddings (Mikolov et al. 2013), used pre-trained word vector representations to incorporate semantic and syntactic knowledge. Self-supervised objectives have become the primary pre-training approach, which involves masking parts of the input and training models to predict the masked content. RoBERTa (Liu et al. 2019) trained Transformer encoders using masked language modeling, demonstrating performance improvements through increased model size, increased data, and removing BERT's next sentence prediction. Following BERT's success, several extensions have been proposed to enhance pre-trained representations further. ALBERT (Lan et al. 2019) employs techniques to reduce the number of parameters, resulting in reduced memory usage and faster model training times. T5 (Xue et al. 2020) adopts a unified text-to-text format for all NLP tasks. Recent work explores transfer learning techniques requiring much less data, addressing the need for massive pre-trained corpora (Gururangan et al. 2020). pre-trained models have also been adapted to various downstream tasks beyond text classification, including document summarization (Liu and Lapata 2019) and language generation (Radford et al. 2019).

3.2.4 Multi-task learning

Multi-task learning aims to leverage inductive biases by training a model across a range of related tasks. It encourages the acquisition of more generalizable features that perform well across tasks, preventing over-specialization in a single objective. Early multi-task learning focused on neural networks with shared hidden layers trained on different tasks to improve generalization (Caruana 1997). Modern approaches extend this idea to deep neural networks using techniques like hard parameter sharing, where hidden layers are shared while task-specific output layers are maintained. In natural language processing (NLP), multi-task learning has proven effective for various language comprehension tasks, including entity recognition, relation extraction, part-of-speech tagging, and semantic role labeling. Sharing representations for related tasks improves overall performance and trains more robust models (Collobert and Weston 2008). Recent advances explore adaptive architectures that dynamically route parameters to optimize sharing. Standalone tasks can access dedicated modules, while related tasks share appropriate subnetworks. This provides the flexibility benefits of multi-task learning with less destructive interference when combining disparate objectives (Ruder et al. 2019). pre-trained models like BERT are inherently multi-task learners. The masked language modeling objective trains a generic textual representation useful for many downstream tasks. Finetuning on a target task leverages this inductive transfer. Extensions like MT-DNN (Liu et al. 2019) simultaneously fine-tune multiple objectives to improve generalization further. Ongoing research also studies multi-task learning across modalities like vision and language. Joint training on aligned image captioning and textual entailment tasks improves representations for grasping semantics across both modalities (Kiela et al. 2019).

3.2.5 Ensemble methods

These approaches involve merging several models to enhance the overall predictive accuracy. Training diverse base models on distinct data, parameters, and objectives reduces variance and overfitting (Hansen and Salamon 1990). Simple averaging ensemble techniques like bootstrap aggregating (bagging) train base models on different subsets of the data. Models are then aggregated by averaging predictions. Boosting adaptively reweights training data to focus on hard examples. More advanced selective ensembling dynamically chooses a subset of base models per example based on uncertainty estimates. Models confident in a sample abstain from prediction. Ensembles can also be formed in output space by stacking base model predictions as input features to a top-level combiner model (Wolpert 1992). For sentiment analysis, ensemble diversity stems from varied lexical features, neural architectures, and adversarial training objectives (Wei and Zou 2019). Transfer learning ensembles using the same pre-trained model fine-tuned on different target datasets improves robustness. Multitask ensembles targeting related objectives also generalize better on unseen data (Liu et al. 2019). In a notable contribution to this paradigm, Alsayat (2022) advance sentiment analysis through a customized ensemble deep learning model. This model leverages advanced word embedding techniques within a long short-term memory (LSTM) network framework, integrating the baseline classifier with other leading classifiers for sentiment analysis. Demonstrating enhanced classification accuracy across diverse datasets, including Twitter coronavirus hashtags, Amazon and Yelp reviews, this approach underscores the efficacy of ensemble methods in harnessing deep learning for nuanced sentiment classification tasks. Overall, selectively combining diverse, complementary models trained on different views of the data builds more robust sentiment classifiers. Future work on optimized ensemble construction and more efficient inference will further improve generalization. Integrating complementary regularization techniques holds promise for creating robust, generalizable sentiment systems. However, rigorous evaluation practices are critical for measuring real progress.

3.2.6 Novel methodologies and hybrid models

Building on ensemble strategies, hybrid methodologies have emerged as a promising direction for enhancing generalization, robustness, and interpretability in sentiment analysis. These approaches combine complementary learning paradigms-such as neural networks, rule-based systems, and generative models-to overcome the limitations of individual techniques. In this section, we categorize hybrid models into five distinct groups based on their structural design and methodological intent: (1) neural architecture fusion, (2) ensemble-based hybrids, (3) unsupervised hybrid models, (4) semantic and rule-based hybrids, and (5) domain-specific hybrids. This classification clarifies the organizational logic behind the selected works and underscores the range of strategies employed to achieve more adaptable and effective sentiment analysis systems.

Neural Architecture Fusion A prominent class of hybrid models combines multiple neural architectures to leverage their complementary strengths. For instance, integrating the RoBERTa Transformer model with Gated Recurrent Units (GRU) allows the model to benefit from RoBERTa's attention-based contextual embeddings and GRU's ability to capture long-term dependencies in textual data. This fusion has demonstrated improved sentiment classification across IMDb, Sentiment140, and Twitter datasets, particularly in handling imbalanced data (Tan et al. 2023).

Ensemble-Based Hybrids Other approaches utilize ensemble learning to combine multiple neural models. For example, RoBERTa, LSTM, and BiLSTM can be combined within an ensemble framework using averaging or majority voting. In this setup, RoBERTa provides rich contextual embeddings, while LSTM and BiLSTM capture sequential and bidirectional dependencies. Such ensemble hybrids have shown strong generalization and predictive accuracy in complex domains such as finance and healthcare, where sentiment often exhibits subtle, context-dependent patterns (Tan et al. 2022).

Unsupervised Hybrid Models In low-resource scenarios, hybrid unsupervised models have been employed to improve aspect-based sentiment analysis (ABSA). The Hierarchical Dirichlet Process-Latent Dirichlet Allocation (HDP-LDA) model, for example, identifies aspect-specific sentiment terms without labeled data. Unlike standard LDA, HDP-LDA can dynamically determine the number of topics and separate factual from opinionated content, leading to improved performance in unsupervised sentiment classification (Ding et al. 2013).

Semantic and Rule-based Hybrids Another group of hybrid methods integrates traditional lexicon-based or rule-based strategies with learning-based models to enhance precision and interpretability. For instance, fuzzy logic combined with sentiment lexicons such as SentiWordNet enables models to better capture intensity and polarity, especially in sentence-level sentiment analysis. These hybrid models have been shown to outperform classical machine learning baselines such as Naive Bayes and Maximum Entropy, particularly on short, opinion-rich datasets like movie reviews (Appel et al. 2016) Their rule-driven components contribute to explainability, while the learned components help overcome brittleness to domain shifts making them well suited for sentiment tasks in highly variable social text.

Domain-Specific Hybrid Models Finally, certain hybrid architectures are designed for specific domains or multi-class sentiment settings. An example is the NH-ResNeXt-RNF model, which integrates ResNeXt with Recurrent Neural Networks (RNNs) and uses preprocessing to reduce noise in social media and e-commerce datasets. This architecture combines ResNeXt's capacity for deep feature extraction with RNNs' strength in sequence modeling. Experimental results on Amazon and Twitter datasets show that NH-ResNeXt-RNF achieves over 95% accuracy, highlighting the effectiveness of tailoring hybrid architectures to domain-specific requirements (Krosuri and Aravapalli 2023).

These hybrid and adaptive methodologies not only enhance generalization and robustness across varied domains, but also contribute to maintaining interpretability-an important factor for user's trust and real-world deployment. By strategically combining complementary modeling techniques, hybrid approaches continue to advance the frontiers of sentiment analysis by enabling more flexible, accurate, and context-aware systems.

3.2.7 Summary and perspectives on generalization strategies

The six generalization strategies discussed in Sections 3.2.1 through 3.2.6 represent complementary yet distinct approaches to improving model robustness, especially in the context of sentiment analysis. In this section, we offer a comparative reflection on these methods, highlighting their relative strengths, limitations, and common application contexts.

Regularization techniques, such as L2 penalties, dropout, entropy-based loss functions, and pruning, are foundational tools that are widely used across nearly all deep learning models. They are especially valuable in controlling overfitting when training on limited or noisy sentiment data. However, they offer limited support for addressing domain shifts or linguistic variability across different datasets.

Data augmentation and *adversarial training* are increasingly popular in sentiment classification, particularly in social media and noisy-text settings. These methods improve robustness by exposing models to syntactic and lexical variations. While simple augmentations are easy to implement, more sophisticated generative or adversarial techniques

often require additional resources and tuning. Their impact is most noticeable in tasks with sparse or imbalanced data.

Transfer learning remains the most impactful and widely adopted strategy in recent years. Pretrained language models like BERT, RoBERTa, and T5 have demonstrated strong generalization across domains and languages. These models are particularly useful in low-resource settings and for domain adaptation. However, their high computational cost and limited interpretability remain concerns in real-world deployment.

Multi-task learning is well suited for scenarios involving multiple related NLP tasks (e.g., sentiment + aspect extraction). It promotes the learning of shared representations and reduces task-specific overfitting. Nevertheless, it requires careful design to avoid negative transfer, especially when the related tasks have competing objectives or data distributions.

Ensemble methods are effective at reducing variance and improving predictive performance, particularly when the base models capture complementary aspects of the data. They are commonly applied in multi-domain applications and are especially useful when high accuracy is prioritized over interpretability and efficiency.

Hybrid models, as discussed in Section 3.2.6, offer a flexible design space by combining different modeling paradigms-rule-based, neural, generative, or domain-specific architectures. These models often excel in specialized contexts such as aspect-based sentiment analysis or low-resource domains, offering both improved generalization and, in some cases, better explainability. However, their design and tuning complexity can be a barrier to wider adoption.

In practice, the choice of generalization strategy should be informed by the dataset characteristics, domain constraints, and computational budget. While transfer learning and hybrid approaches show the most promise in broad, cross-domain applications, regularization and augmentation remain essential in all stages of model development. Continued research into adaptive, interpretable, and efficient generalization techniques will be key to the next generation of sentiment analysis systems.

4 Future directions

The journey of sentiment analysis in natural language processing (NLP) from its inception to today showcases a remarkable evolution of methodologies and applications. Initially, reliant on traditional machine learning approaches, sentiment analysis benefited from these early methods' foundational insights. However, while effective in their time, these techniques often fell short of capturing the complex nuances of human sentiment due to their reliance on surface-level text patterns and limited handling of language subtleties.

Initially, sentiment classification heavily depended on traditional machine learning techniques, including Support Vector Machines (SVMs), Naive Bayes, and Decision Trees. These methods, leveraging bag-of-words models, n-grams, and basic lexical features, encountered difficulties in capturing nuanced sentiments, contextual meanings, and the inherent sequential nature of language. Particularly, they struggled with interpreting sarcasm, negation, and context-dependent expressions of sentiment, prompting a shift towards more sophisticated models capable of a deeper understanding of language. The advent of neural network approaches, especially those involving deep learning, represented a significant leap forward in sentiment classification. This transition was led by DNNs, CNNs, RNNs, LSTM, and transformers, which offer enhanced capabilities in processing sequential data and understanding long-term dependencies.

4.1 Feature selection

Feature selection is pivotal in optimizing machine learning models for sentiment analysis. Its importance extends beyond mere data dimensionality reduction to selecting a subset of relevant features crucial for model training. The primary aim is to improve model performance by mitigating overfitting and enhancing generalizability to new, unseen data. Moreover, it significantly reduces computational costs and enhances model interpretability, which is crucial for making models accessible and understandable to users and stakeholders.

Recent advancements have highlighted the significance of intrinsic feature selection algorithms, such as Lasso regression (Li et al. 2005), which streamline model complexity while improving interpretability and computational efficiency. Additionally, Mutual information has been extensively utilized as a criterion to enhance feature relevance and reduce redundancy in high-dimensional datasets. For instance, correlation measures based on mutual information help calculate the association between features, enabling the removal of irrelevant features and streamlining models (Nagpal et al. 2014; Manikandan et al. 2018). Additionally, sophisticated approaches integrate feature-feature and feature-class mutual information to refine feature subsets, thereby improving model interpretability and classification accuracy (Zheng and Kwoh 2011).

Hybrid feature selection models integrate filter, wrapper, and embedded methods, offering a versatile and context-adaptive strategy. Ansari et al. (2019) proposed a hybrid filter-wrapper approach using Recursive Feature Elimination (RFE) combined with Particle Swarm Optimization (PSO). Their study demonstrated enhanced

accuracy across sentiment benchmarks, achieving an accuracy of **95.85%** on the Movie dataset with a feature subset reduction from 39,389 to 1,276 features. Similarly, Jeon and Oh (2020) developed a hybrid recursive feature elimination framework leveraging the strengths of Support Vector Machines (SVMs) and Random Forests (RF). Their method achieved significant dimensionality reduction while maintaining high classification accuracy, achieving **95.8%** accuracy on the Ionosphere dataset and **91.9%** on the Satellite dataset.

Metaheuristic approaches have gained traction in sentiment analysis for feature selection. Kristiyanti et al. (2023) employed the Salp Swarm Algorithm (SSA) with a novel binary version to enhance feature selection efficiency. Their approach achieved an accuracy improvement of **80.95%** on benchmark sentiment datasets using the KNN classifier, showcasing its capability to address high-dimensional data challenges effectively. Parlar and Sarac (2019) introduced an Intelligent Water Drops (IWD) based algorithm, demonstrating the utility of physics-inspired models for feature prioritization. Their approach achieved an F1-score of **72.1%** on Turkish Twitter data with a reduced feature set (250 features), highlighting both computational efficiency and performance improvements.

The summarized results in Table 7 highlight the effectiveness of these techniques. Feature selection techniques show varied effectiveness based on the nature of the dataset and the method used. Hybrid approaches, such as RFE combined with PSO, achieve high accuracy on structured datasets like movie reviews by effectively reducing dimensionality while preserving discriminative features. Metaheuristic algorithms like the Salp Swarm Algorithm (SSA) demonstrate strong performance in high-dimensional settings due to their adaptive search capabilities. Meanwhile, physics-inspired methods such as IWD perform well on domain-specific, informal data (e.g., Turkish Twitter), prioritizing features dynamically and offering computational efficiency alongside moderate classification gains.

Future research should explore adaptive mechanisms for feature selection, incorporating metaheuristic strategies and real-time analytics. Developing hybrid frameworks that seamlessly integrate feature selection within deep learning pipelines could unlock new paradigms in sentiment analysis, enabling models to dynamically adjust to evolving data landscapes and ensure robust performance across diverse applications.

4.2 Data augmentation

Data augmentation holds significant promise for enhancing the generalization capabilities of machine learning models in sentiment analysis. In scenarios with limited labeled data, augmentation techniques artificially expand the dataset, introducing a wider range of linguistic variations and sentiment expressions. This approach has proven particularly effective in addressing data sparsity and imbalances, which are common challenges in sentiment analysis tasks.

Omran et al. (2023) demonstrated the impact of simple augmentation techniques, such as random swap, applied in conjunction with deep learning models like LSTM. Their study revealed substantial performance improvements across multilingual datasets, including English, Modern Standard Arabic (MSA), and Bahraini Dialects (BDs). For instance, applying random swap augmentation improved accuracy on the BDs dataset from **82.66%** to **96.72%**, highlighting its efficacy in low-resource scenarios. Similarly, Xiang et al. (2021) proposed Part-of-Speech Lexical Substitution for Data Augmentation (PLSDA), which generated semantically rich training data. Their method significantly elevated model performance, achieving **94.7%** accuracy on SST-2 and **92.8%** accuracy on IMDB. This approach emphasizes the importance of generating diverse yet semantically relevant samples.

Karimi et al. (2021) explored adversarial training as a data augmentation strategy for aspect-based sentiment analysis. Their BERT Adversarial Training (BAT) framework achieved an F1-score of **85.57%** for aspect extraction on the Laptop domain and an accuracy of **86.03%** for aspect sentiment classification on the Restaurant domain, demonstrating robustness to domain shifts. Lastly, Ombabi et al. (2024) applied a context-aware deep learning architecture, Deep Conv-ABiLSTM, to Arabic sentiment analysis. Their model,

Table 7 Performance of feature selection techniques in sentiment analysis

References	Techniques	Dataset	Accuracy (%)	F1-Score (%)
Ansari et al. (2019)	RFE + PSO Hybrid	Movie Dataset	95.85	95.00
Jeon and Oh (2020)	Hybrid Recursive Feature Elimination	Satellite, Ionosphere	91.90 (Satellite)95.80 (Ionosphere)	–
Kristiyanti et al. (2023)	Salp Swarm Algorithm (Binary Version)	Benchmark Sentiment Datasets	80.95	–
Parlar and Sarac (2019)	Intelligent Water Drops (IWD)	Turkish Twitter	–	72.10

F1-scores are reported only where available in the referenced studies

Table 8 Performance of data augmentation techniques in sentiment analysis

References	Techniques	Dataset	Accuracy (%)	F1-Score (%)
Omran et al. (2023)	Random Swap + LSTM	English, MSA, BDs	96.72 (BDs), 97.01 (MSA)	–
Xiang et al. (2021)	Lexical Substitution (PLSDA)	SST-2, IMDB, Twitter	94.7 (SST-2), 92.8 (IMDB)	–
Karimi et al. (2021)	BERT Adversarial Training (BAT)	Laptop, Restaurant	86.03 (Rest14), 85.57 (Laptop)	–
Ombabi et al. (2024)	Deep Conv-ABiLSTM	LABR, Ar-Twitter	96.08 (LABR), 93.73 (Ar-Twitter)	95.51 (LABR)

Note: F1-scores are reported only where available in the referenced studies

leveraging both convolutional and attention-based Bi-LSTM layers, achieved **96.08%** accuracy on the LABR dataset and **93.73%** accuracy on the Ar-Twitter dataset, showcasing its effectiveness in handling Arabic text with diverse contexts.

The summarized results in Table 8 highlight the efficacy of these techniques. For instance, random swap augmentation (Omran et al. (2023)) improved accuracy by over 10% across multiple datasets, while PLSDA (Xiang et al. (2021)) demonstrated its ability to elevate performance across sentiment benchmarks. These findings underscore the potential of advanced augmentation techniques to improve model generalization, particularly in resource-scarce settings.

Future research should explore real-time adaptive augmentation techniques. By leveraging state-of-the-art generative models, it is feasible to produce contextually relevant, sentiment-laden text, significantly enhancing the diversity and representativeness of training datasets. Additionally, integrating data augmentation with technologies such as transfer learning, few-shot learning, and domain adaptation presents an exciting opportunity to unlock new paradigms in sentiment analysis. This could lead to the development of highly generalizable and efficient models, particularly in data-sparse environments. Moreover, research should prioritize the ethical implications of augmentation strategies, ensuring that synthetic data accurately reflects the sentiment and context of real-world communication without introducing bias. Exploring techniques to dynamically adjust augmentation strategies based on evolving trends in online discourse will also ensure that sentiment analysis models remain effective and relevant over time.

4.3 Adversarial training

Adversarial training reduces model biases and improves generalizability by injecting perturbations into the input data or model parameters to mimic potential adversarial attacks. This technique is particularly relevant in sentiment analysis, where models can be susceptible to adversarial noise and domain shifts, potentially skewing their predictions. Karimi et al. (2021) showcased the success of adversarial training in aspect-based sentiment analysis (ABSA). Their BERT Adversarial Training (BAT) framework achieved significant improvements on the SemEval benchmark datasets, with an F1-score of **85.57%** in aspect extraction and an accuracy of **86.03%** in aspect sentiment classification for the Laptop and Restaurant domains, respectively. These results demonstrated the efficacy of adversarial examples in enhancing robustness to domain shifts.

Wang and Gan (2023) explored multi-level adversarial training at character, word, and sentence levels to enhance stock sentiment prediction models. Their best-performing model, RoBERTa-Twitter with word-level adversarial training, achieved an accuracy of **69.54%** and a macro F1-score of **65.72%** on the TweetFinSent dataset. This study highlighted the effectiveness of word-level perturbations for improving predictions in the volatile domain of financial sentiment analysis.

Lina et al. (2024) proposed an adversarial training framework for cross-lingual sentiment classification, incorporating heteroscedastic uncertainty estimation. Their model, combining mBERT with dual-channel feature extraction, achieved an average accuracy of **88.48%** on the MLDoc dataset and **86.63%** on the PAWS-X dataset. These results underscore the potential of adversarial perturbations for improving robustness across multiple languages.

Kitada and Iyatomi (2021) introduced an attention-based adversarial training model (Attention AT) that demonstrated improved robustness and interpretability. On the SST dataset, their Attention AT model achieved an F1-score of **82.20%**, outperforming baseline approaches such as Vanilla (79.27%) and Word AT (79.61%). Similarly, their approach improved performance on the IMDB and AGNews datasets, with F1-scores of **90.21%** and **96.69%**, respectively. In a follow-up study, Kitada and Iyatomi (2023) extended this work with Attention VAT, combining attention mechanisms with virtual adversarial perturbations. This model achieved state-of-the-art performance on single-sequence tasks, including **83.18%** F1-score on SST and **92.48%** F1-score on IMDB, and pair-sequence tasks such as SNLIE, with a micro F1-score of **77.87%**.

The results in Table 9 showcase the effectiveness of adversarial training techniques in improving sentiment analysis performance across various datasets and domains.

Table 9 Performance of adversarial training models for sentiment analysis

References	Techniques	Dataset	Accuracy (%)	F1-Score (%)
Karimi et al. (2021)	BERT Adversarial Training (BAT)	Laptop, Restaurant	86.03 (Rest14)85.57 (Laptop)	–
Wang and Gan (2023)	Multi-level Adversarial Training	TweetFinSent	69.54	65.72
Lina et al. (2024)	Cross-Lingual Adversarial Training	MLDoc, PAWS-X	88.48 (MLDoc)86.63 (PAWS-X)	–
Kitada and Iyatomi (2021)	Attention Adversarial Training (Attention AT)	SST, IMDB, AGNews	82.20 (SST)90.21 (IMDB)	–
Kitada and Iyatomi (2023)	Attention VAT	SST, IMDB, AGNews, SNLIE	83.18 (SST)92.48 (IMDB)96.35 (AGNews)	77.87 (SNLIE)

F1-scores are reported only where available in the referenced studies

Notable results include BAT's performance in ABSA tasks, with an accuracy of **86.03%** on the Restaurant domain, and Attention VAT's performance on single-sequence tasks, achieving **92.48%** on IMDB. These findings highlight the adaptability and effectiveness of adversarial training in addressing challenges such as domain shifts, data noise, and imbalances.

Future research could delve into domain-adaptive adversarial training techniques, focusing on creating models that can seamlessly adapt to and perform well across domains and languages in sentiment analysis. This involves developing adversarial training methods to strengthen model resilience and ensure the model's adaptability to domain-specific nuances. Moreover, as adversarial training can potentially increase model complexity, future work should explore strategies to balance adversarial robustness with computational efficiency. This could involve investigating lightweight adversarial training frameworks or optimizing existing adversarial training algorithms to maintain high performance without significantly increasing computational demands.

4.4 Multimodal sentiment analysis

Multimodal modeling in sentiment analysis represents a dynamic and evolving frontier, harnessing the power of diverse data types to capture the complexity of human emotions more accurately. By integrating textual, visual, and auditory data, this approach offers unprecedented opportunities for deeper, more nuanced sentiment analysis. Multimodal fusion allows for a holistic understanding of sentiment, enabling better generalization across communication forms. This capability is crucial for decision-making processes that involve significant investments of time and resources. The increasing availability of multimodal data on social media and online platforms underscores the need for robust models capable of handling such complexities (Cambria et al. 2013).

Recent advancements highlight the efficacy of various approaches in multimodal sentiment analysis. Chen et al. (2022) proposed the Video-based Cross-modal Auxiliary Network (VCAN), achieving binary accuracy scores of **75.9%** and **73.9%** on the CMU-MOSI and CMU-MOSEI datasets, respectively. Their model leverages the Audio Features Map Module (AFMM) and Cross-Modal Selection Module (CMSM), simplifying sentiment analysis into an image classification task. Huddar et al. (2020) developed a multi-level context extraction model combined with attention-based contextual inter-modal fusion. Their approach demonstrated superior performance, achieving an accuracy of **80.33%** on the CMU-MOSI dataset and **80.87%** on the IEMOCAP dataset.

Huang et al. (2019) introduced a Deep Multimodal Attentive Fusion (DMAF) model, which integrates intermediate and late fusion schemes to capture discriminative features and internal correlations between modalities. The model achieved an accuracy of **88.0%** and an F1-score of **87.6%** on the Flickr-m dataset, emphasizing the importance of deeper modality integration. Similarly, Lai and Yan (2022) proposed the Asymmetric Window Multi-attentions (AWMA) model, which achieved state-of-the-art results on the CMU-MOSI dataset with an accuracy of **80.0%** and an F1-score of **79.9%**. Their model differentiates historical and future contexts, capturing dynamic interactions within and across modalities.

Gkoumas et al. (2021) introduced an Entanglement-driven Fusion Neural Network (EFNN) for video sentiment analysis, achieving an accuracy of **82.8%** and an F1-score of **82.6%** on the CMU-MOSEI dataset. Their model leverages quantum-inspired entanglement mechanisms to capture non-separable interactions between modalities, setting a new benchmark for video-based sentiment analysis. Lopes et al. (2021) explored the potential of automation in multimodal sentiment analysis through an AutoML-based approach. By integrating individual textual and visual predictions into a fused classification, they highlighted the role of automation

Table 10 Performance of multimodal sentiment analysis models

References	Techniques	Dataset	Accuracy (%)	F1-Score (%)
Chen et al. (2022)	VCAN with AFMM and CMSM	CMU-MOSI, CMU-MOSEI	75.9 (MOSI)73.9 (MOSEI)	76.0 (MOSI)
Huddar et al. (2020)	Multi-level Context Extraction and Fusion	CMU-MOSI, IEMOCAP	80.33 (MOSI)80.87 (IEMOCAP)	–
Huang et al. (2019)	DMAF (Deep Multimodal Attentive Fusion)	Flickr-m, Getty Images	88.0 (Flickr-m)86.9 (Getty)	87.6 (Flickr-m)86.6 (Getty)
Lai and Yan (2022)	Asymmetric Window Multi-attentions	CMU-MOSI	80.0	79.9
Gkoumas et al. (2021)	Entanglement-driven Fusion Neural Network	CMU-MOSEI	82.8	82.6

F1-scores are reported only where available in the referenced studies

in model selection for improving sentiment detection accuracy. Baecchi et al. (2016) investigated multimodal feature learning for social media content, achieving improved sentiment classification through a combination of neural network-based models. Lastly, Das and Singh (2023) presented a comprehensive survey on multimodal sentiment analysis, emphasizing the evolution from unimodal to multimodal approaches and highlighting challenges in multimodal fusion and domain adaptation.

The summarized results in Table 10 highlight the efficacy of these models. Notable results include the Entanglement-driven Fusion Neural Network proposed by Gkoumas et al. (2021), which achieved an accuracy of **82.8%** and an F1-score of **82.6%** on the CMU-MOSEI dataset. Similarly, Huddar et al. (2020) demonstrated the benefits of multi-level context extraction and inter-modal fusion with an accuracy of **80.33%** on the CMU-MOSI dataset.

Future research in multimodal sentiment analysis should prioritize adaptive fusion mechanisms capable of dynamically adjusting to the data's specificities. Advanced techniques that enhance the interaction between modalities will be crucial for capturing the nuanced interplay of textual, auditory, and visual cues. Additionally, addressing challenges related to interpretability and explainability will be essential as multimodal models become increasingly complex. Innovative methods to enhance model transparency will foster trust and accessibility, enabling their broader application across industries.

4.5 Self-supervised learning

Self-supervised learning represents a transformative shift in sentiment analysis, offering innovative ways to harness unlabeled data for deeper, more accurate text interpretation. Unlike supervised learning, which relies on human-annotated labels, self-supervised approaches exploit the inherent structure of data to learn meaningful representations without explicit labels. This method significantly reduces the dependency on large labeled datasets, mitigating the associated costs and efforts of annotation. By leveraging vast amounts of unlabeled data, self-supervised learning enriches the generalization capabilities of sentiment analysis models, exposing them to a diverse array of linguistic styles, expressions, and sentiment-laden content, including slang and cultural nuances.

Key advancements in this domain include techniques such as contrastive pre-training, which has shown immense promise in acquiring universal text representations. Le and Mikolov (2014) introduced the Paragraph Vector algorithm, an unsupervised method that produces fixed-length feature vectors from variable-length text segments. Their approach achieved an accuracy of **92.58%** on the IMDB dataset and **87.8%** on the SST dataset, demonstrating its ability to capture semantic and contextual relationships in text.

Liu et al. (2023) employed multi-view contrastive learning to enhance modality representation in multimodal sentiment analysis. Their framework achieved accuracies of **83.7%** on the CMU-MOSI dataset and **84.95%** on the CMU-MOSEI dataset, with corresponding F1-scores of **84.2%** and **85.01%**, respectively. Zhang and He (2013) proposed a partitioned self-training approach for Chinese sentiment classification, achieving **95.2%** accuracy on THU Hotel Reviews and **90.5%** on ChnSentiCorp-4000. Zhu et al. (2022) demonstrated the potential of self-supervised learning for multimodal sentiment analysis by leveraging image-text matching. Their Senti-ITEM model achieved an accuracy of **76.6%** on the B-T4SA dataset and **84.2%** on IMDB, illustrating the model's ability to generalize from multimodal data.

In the realm of video-based sentiment analysis, Qian et al. (2023) proposed Sentiment Knowledge Enhanced Self-supervised Learning (SKESL), which incorporates sentiment knowledge and non-verbal behavior. Their model achieved an accuracy of **86.77%** on CMU-MOSI and **86.25%** on CMU-MOSEI, setting new benchmarks in multimodal sentiment analysis.

The summarized results in Table 11 illustrate the effectiveness of various self-supervised learning techniques

Table 11 Performance of self-supervised learning models for sentiment analysis

References	Techniques	Dataset	Accuracy (%)	F1-Score (%)
Le and Mikolov (2014)	Paragraph Vector	IMDB, SST	92.58 (IMDB)87.8 (SST)	–
Liu et al. (2023)	Multi-view Contrastive Learning	CMU-MOSI, CMU-MOSEI	83.7 (MOSI)84.95 (MOSEI)	84.2 (MOSI)85.01 (MOSEI)
Zhang and He (2013)	Partitioned Self-training	Chinese Sentiment Datasets	95.2 (THU Hotel)90.5 (ChnSentiCorp)	–
Zhu et al. (2022)	Image-Text Matching	B-T4SA, IMDB	76.6 (B-T4SA)84.2 (IMDB)	–
Qian et al. (2023)	SKESL (Sentiment Knowledge Enhanced SSL)	CMU-MOSI, CMU-MOSEI	86.77 (MOSI)86.25 (MOSEI)	86.82 (MOSI)86.25 (MOSEI)

F1-scores are reported only where available in the referenced studies

in advancing sentiment analysis. Notable results include the Paragraph Vector method by Le and Mikolov (2014), which achieved **92.58%** accuracy on the IMDB dataset, and SKESL by Qian et al. (2023), which demonstrated **state-of-the-art performance** on CMU-MOSI and CMU-MOSEI with accuracies exceeding **86%**. These findings underscore the potential of self-supervised learning to tackle diverse challenges in sentiment analysis, including multimodal and domain-specific applications.

Despite these advancements, challenges persist in adapting self-supervised learning to underrepresented languages and domains. Future research should prioritize the integration of multimodal data, such as text, audio, and images, to create richer and more nuanced sentiment analysis models. Combining self-supervised learning with pre-trained models and transfer learning can further enhance generalization and accuracy. Ethical considerations, particularly in mitigating biases across demographics and addressing cultural differences, will also be vital to the continued success of self-supervised sentiment analysis. These efforts will ensure self-supervised models cater to diverse linguistic and demographic contexts, fostering inclusivity and fairness.

4.6 Multilingual sentiment analysis

Multilingual sentiment analysis bridges linguistic and cultural diversity, enabling a comprehensive understanding of sentiment across languages. Challenges such as limited labeled datasets and cultural nuances make sentiment modeling in underrepresented languages particularly difficult. To address these challenges, researchers have leveraged cross-lingual embeddings, multilingual models, and zero-shot transfer learning, achieving significant advances in sentiment analysis for diverse linguistic contexts.

Ghasemi et al. (2022) investigated cross-lingual and bilingual frameworks for Persian sentiment analysis. Using the Digikala dataset, the monolingual setup with dynamic embeddings achieved an F-measure of **73.36%**. The cross-lingual setup improved this result to **79.05%**, while the bilingual setup achieved the highest F-measure of **84.29%**, demonstrating the benefits of leveraging bilingual resources for low-resource languages.

Robnik-Sikonja et al. (2020) studied cross-lingual sentiment classification using LASER embeddings and multilingual BERT (mBERT). On Twitter datasets spanning 13 languages, their model achieved an F1-score of **63%** for cross-lingual transfer between Polish and Slovak using LASER embeddings, and **52%** for Russian to English transfer using mBERT. These results highlight the challenges and opportunities of sentiment transfer across both similar and different language families.

Phan et al. (2021) explored zero-shot cross-lingual aspect-based sentiment analysis using XLM-R and mBERT. For Aspect Category Detection (ACD) in the monolingual setting, XLM-R achieved F1-scores of **82.66%** (English) and **80.67%** (Spanish), while in the zero-shot setting, it achieved **78.94%** (English from Dutch) and **75.57%** (Spanish from French). For Opinion Target Expression (OTE), XLM-R achieved F1-scores of **75.31%** (English) and **73.63%** (Spanish) in the monolingual setup, and **68.15%** (English from Russian) and **67.55%** (Spanish from English) in the zero-shot setup. These results demonstrate the utility of multilingual pre-trained models for handling low-resource languages.

Dong and De Melo (2018) introduced a dual-channel convolutional neural network (DC-CNN) with cross-lingual embeddings. Their model achieved **96.48% accuracy** on the Allocine dataset (French TV reviews) and **89.49% accuracy** on the SemEval 2016 Task 5 Spanish dataset, demonstrating the effectiveness of domain-specific sentiment embeddings in multilingual settings.

The performance of these models is summarized in Table 12, which provides a comprehensive comparison of techniques, datasets, and key metrics across the studies. For example, the bilingual approach by Ghasemi et al. achieved a notable improvement (F1-score: **84.29%**) over the monolingual baseline (F1-score: **73.36%**). Similarly, the zero-shot capabilities of XLM-R are evident, where F1-scores for tasks like Aspect Category Detection and Opinion Target Expression consistently surpassed 75% in low-resource languages such as Spanish and French. The

Table 12 Performance of multilingual sentiment analysis models

References	Techniques	Dataset	Accuracy (%)	F1-Score (%)
Ghasemi et al. (2022)	Cross-lingual and bilingual embeddings	Persian Digikala Dataset	–	73.36 (Monolingual)79.05 (Cross-lingual)84.29 (Bilingual)
Robnik-Sikonja et al. (2020)	LASER embeddings, mBERT	Twitter (Polish, Slovak, Russian)	–	63 (Polish to Slovak)52 (Russian to English)
Phan et al. (2021)	XLM-R, mBERT (Zero-shot cross-lingual)	SemEval 2016 Task 5	–	78.94 (ACD, English)75.57 (ACD, Spanish) 68.15 (OTE, English) 67.55 (OTE, Spanish)
Dong and De Melo (2018)	DC-CNN, cross-lingual embeddings	Allocine (French TV Reviews)	96.48	–

F1-scores are reported only where available in the referenced studies

highest reported accuracy (**96.48%**) was achieved by Dong and de Melo on the Allocine dataset (Row 4), showcasing the utility of domain-specific embeddings.

Future research in multilingual sentiment analysis should focus on enhancing the generalization capabilities of sentiment models across diverse languages, particularly low-resource languages. Developing *language-agnostic embeddings* and leveraging multilingual datasets can address linguistic and cultural nuances, reducing bias and improving transferability across languages (Liu et al. 2024). Additionally, researchers should prioritize *cultural-contextual modeling* to better account for cultural nuances in sentiment expression, as cultural differences can impact model interpretation (Mabokela et al. 2023). Expanding *multimodal sentiment analysis* by integrating textual, visual, and audio cues could also open new opportunities for understanding sentiment in diverse, real-world contexts (Agüero-Torales et al. 2021). Lastly, creating *robust benchmarks* and advancing transformer-based architectures for multilingual aspect-based sentiment analysis can further elevate performance in low-resource settings (Zhao et al. 2024).

4.7 Capsule networks

Capsule Networks (CapsNets) offer unique advantages in sentiment analysis by capturing hierarchical part-whole relationships, making them particularly effective at recognizing patterns and nuances in textual data. Their ability to encode spatial relationships and dynamically route information between capsules enhances their robustness, even with limited data or across domains. This makes CapsNets an appealing choice for sentiment analysis, especially for tasks requiring high precision and adaptability across different textual contexts.

Recent research highlights the effectiveness of Capsule Networks across various sentiment analysis tasks. Dong et al. (2020) proposed the caps-BiLSTM model, which integrates Capsule Networks with BiLSTM to improve performance on datasets such as MR, IMDB, and SST. As presented in Table 13, Their model achieved an accuracy of **91.96%** on the IMDB dataset, **81.47%** on the MR dataset, and **48.34%** on the SST dataset, demonstrating its versatility in handling sentiment-laden texts. Zhang et al. (2018) introduced CapsuleDAR, a model designed for cross-domain sentiment classification, which achieved an

Table 13 Performance of capsule network models for sentiment analysis

References	Techniques	Dataset	Accuracy (%)	F1-Score (%)
Dong et al. (2020)	Capsule network (caps-BiLSTM)	MR, IMDB, SST	81.47 (MR)91.96 (IMDB)48.34 (SST)	–
Zhang et al. (2018)	CapsuleDAR	Amazon Product Reviews	89.4 (Overall)93.2 (Best Domain)	–
Wang et al. (2018)	RNN-Capsule	MR, SST, Hospital Feedback	83.8 (MR)91.6 (Hospital Feedback)49.3 (SST)	–
Demotte et al. (2023)	Enhanced Capsule Networks	STSGd, CrowdFlower, US Airline	86.87 (STSGd)82.04 (US Airline)	86.70 (STSGd)81.11 (US Airline)

F1-scores are reported only where available in the referenced studies

overall accuracy of **89.4%** on the Amazon Reviews dataset, with the highest domain-specific accuracy reaching **93.2%** (Books → Kitchen). This result highlights the model's ability to generalize effectively across domains.

Wang et al. (2018) developed RNN-Capsule, a hybrid model combining recurrent neural networks with Capsule Networks. As presented in This model excelled in capturing contextual sentiment, achieving an accuracy of **91.6%** on the Hospital Feedback dataset and **83.8%** on the MR dataset, illustrating its potential for real-world applications in healthcare sentiment analysis. Demotte et al. (2023) proposed an enhanced capsule network architecture for social media content analysis (Stanford Twitter Sentiment Gold and CrowdFlower US Airline datasets), achieving an accuracy of **86.87%** and an F1-score of **86.70%** on the STSGd dataset, and an accuracy of **82.04%** with an F1-score of **81.11%** on the US Airline dataset.

The results from these studies, summarized in Table 13, demonstrate the versatility of Capsule Networks in addressing various sentiment analysis challenges, ranging from cross-domain generalization to multimodal sentiment extraction. However, their computational complexity remains a significant challenge, motivating future research to explore efficient implementations and hybrid approaches.

4.8 Graph-based models

Graph-based models have shown exceptional promise in sentiment analysis by leveraging structured representations of textual data. Unlike traditional sequence-based methods, these approaches represent text as structured graphs, enabling the extraction of syntactic and semantic dependencies between words, phrases, and sentences. These models are particularly effective in aspect-based sentiment analysis, where understanding the relationships between aspects and sentiment expressions is critical.

Recent advancements in graph neural networks (GNNs) have significantly extended their applicability to sentiment analysis tasks. For instance, Zhao et al. (2024) proposed the Structured Dependency Tree-based Graph Convolutional Network (SDTGCN), which enhances syntactic dependency trees with sentiment commonsense knowledge and part-of-speech tags. As illustrated in Table 14, this model achieved an accuracy of **91.53%** and an F1-score of **77.08%** on the Rest16 dataset. Similarly, Niu et al. (2021) introduced the Syntactic Edge-Enhanced Gated Graph Neural Network (SEE-GGNN), which combines syntactic dependency and word co-occurrence graphs. The model achieved an F1-score of **86.95%** on the Chinese Hotel Review dataset and **87.14%** on the Chinese Takeout Review dataset. Graph-based approaches that incorporate heterogeneous structures have also shown substantial progress. An An et al. (2022) developed the Heterogeneous Aspect Graph Neural Network (HAGNN), which integrates word, sentence, and aspect nodes to capture complex interactions. Using BERT-based embeddings, HAGNN achieved an accuracy of **86.20%** and an F1-score of **76.92%** on the Rest2014 dataset. In addition, Jiang et al. (2023) proposed the Multiscale Graph Attention Network (MSGAT), which integrates syntactic dependency trees with multiscale feature extraction to enhance sentence-level sentiment analysis. MSGAT achieved a remarkable accuracy of **94.79%** and an F1-score of **96.23%** on the Chinese Hotel Review dataset.

Lastly, Guan et al. (2023) introduced the Enhanced Syntactic and Semantic Graph Convolutional Network (ESSGCN). This model constructs aspect-oriented syntactic and semantic graphs, dynamically weighting syntax and semantics through contrastive learning. ESSGCN achieved an accuracy of **87.86%** and a macro-F1 score of **83.14%** on the Restaurant dataset.

Table 14 provides a summary of these graph-based models and their performance metrics. These studies collectively highlight the versatility and effectiveness of graph-based models in capturing syntactic and semantic relationships for sentiment analysis. Future research could explore optimizing these models for low-resource domains, reducing computational overhead, and integrating multimodal data for a more comprehensive understanding of sentiment across diverse contexts.

Table 14 Performance of graph-based models for sentiment analysis

References	Technique	Dataset	Accuracy (%)	F1-Score (%)
Zhao et al. (2024)	SDTGCN (Structured Dependency Tree-based GCN)	Rest16	91.53	77.08
Niu et al. (2021)	SEE-GGNN (Syntactic Edge-Enhanced GGNN)	Chinese HotelTakeout Reviews	–	86.95 (Hotel)87.14 (Takeout)
An et al. (2022)	HAGNN (Heterogeneous Aspect Graph Neural Network)	Rest2014	86.20	76.92
Jiang et al. (2023)	MSGAT (Multiscale Graph Attention Network)	Chinese Hotel Reviews	94.79	96.23
Guan et al. (2023)	ESSGCN (Enhanced Syntactic and Semantic GCN)	Restaurant	87.86	83.14

4.9 Ethical concerns: bias in sentiment analysis models

Bias in sentiment analysis models poses significant ethical challenges due to their reliance on biased training data and model development practices. These biases can lead to unfair predictions, as demonstrated in age and racial disparities.

4.9.1 Sources of bias

Sociodemographic Bias Sentiment analysis models encode biases based on sociodemographic attributes such as age and race. For instance, research analyzing 15 sentiment models and 10 GloVe embeddings demonstrated systematic age-related biases, which could only be mitigated by carefully processing the training data (Diaz et al. 2018). Similarly, models have shown racial biases linked to names popular among specific demographics, highlighting broader algorithmic concerns Rozado (2020).

Bias in sociodemographic attributes often arises from the uneven representation of demographic groups in training data. Overrepresentation of certain groups or stereotypes within the training datasets amplifies biases in model predictions. For example, age-related biases are rooted in underrepresentation of older adults in sentiment training data, while racial disparities in sentiment lexicons perpetuate stereotypes. These biases can compound downstream, influencing applications in areas like hiring and content moderation, leading to unfair outcomes for certain demographic groups (Rozado 2020).

Linguistic Bias Language biases in multilingual sentiment analysis models have been observed, such as favoritism towards certain languages (e.g., French over English), leading to inconsistent sentiment scoring (Wong and M'hiri 2024). Multilingual models often assign higher positive sentiment scores to certain languages due to differences in cultural or linguistic expressions, resulting in inconsistent behavior. Such biases highlight the need for balanced datasets and careful calibration across languages. Additionally, linguistic biases may manifest as discrepancies in the sentiment scoring of dialects and informal speech patterns, as sentiment models are often trained on formal text. The inclusion of diverse linguistic variations could mitigate discrepancies and reduce favoritism in multilingual sentiment classification (Wong and M'hiri 2024).

Annotation and Political Bias Annotator biases, often unnoticed, significantly influence model outputs. Research auditing Polish sentiment models showed that human annotators' political biases were reflected in sentiment predictions, necessitating dataset pruning to reduce bias Plisiecki et al. (2024). The socio-political context of annotators affects the consistency of labeled datasets. For example, sentiment annotation involving politically sensitive topics showed that annotators from opposing political ideologies labeled the same data differently, injecting bias into the training process. This finding underscores the need for diverse annotator pools and neutral labeling protocols.

4.9.2 Mitigation strategies

Causal and Counterfactual Analysis Employing causal mediation analysis and targeted counterfactual training can significantly reduce bias while maintaining prediction accuracy. For example, gender fairness improved in models trained using counterfactual sentences without degrading sentiment accuracy Da et al. (2024). Causal mediation analysis helps identify the pathways through which bias propagates in sentiment predictions. By isolating the causal relationships between demographic attributes and sentiment scores, developers can better understand the sources of bias and design counterfactual interventions. This approach can be extended to other demographic attributes like age and race, improving fairness across diverse user populations.

Bias-Aware Thresholding Adjusting polarity prediction thresholds in lexicon-based methods has demonstrated success in reducing prediction bias and improving overall accuracy (Iqbal et al. 2015). Threshold adjustment helps address the disproportionate influence of extreme sentiment polarities, which are often misclassified due to skewed training data. Dynamic thresholds that adapt to the context of sentiment analysis tasks ensure that models remain sensitive to nuanced sentiments, reducing overgeneralization of negative or positive sentiment biases.

Dataset Curation and Audits Pruning biased content from training datasets has proven effective in reducing biases related to political affiliations and may reduce biases related to sociodemographic attributes (Plisiecki et al. 2024).

Dataset audits should go beyond simple keyword-based filtering and involve a more comprehensive review of the data collection process. This includes identifying and removing data points with inherent annotation biases or demographic imbalances. For example, political datasets with skewed representation of specific ideologies can be balanced to mitigate downstream political bias in sentiment predictions.

Bias in sentiment analysis models originates from multiple sources, including data, annotations, and language inconsistencies. Mitigating these biases requires rigorous data processing and algorithmic refinements, such as causal mediation analysis and bias-aware thresholding, to ensure equitable and trustworthy AI systems. Future research should integrate these ethical considerations into the sentiment analysis model development pipeline. Addressing these challenges will ensure that sentiment analysis systems are not only accurate but also fair, transparent, and accountable,

paving the way for their responsible application in diverse and sensitive domains.

5 Real-world applications

The advancements in sentiment analysis have profound implications across various sectors. By applying sophisticated techniques discussed in this survey, practitioners can address real-world challenges effectively. Below are some potential applications:

5.1 Healthcare (Mental health monitoring)

In healthcare, sentiment analysis has become instrumental in monitoring mental health by analyzing patient language across social media, digital communications, or health forums. Machine learning models and NLP techniques help detect early signs of mental health conditions, offering insights into emotional well-being and mental health patterns through online behavioral analysis (Kanojia and Joshi 2023). This capability is essential for creating proactive mental health interventions, especially as mental health needs continue to rise worldwide. For example, sentiment analysis of Twitter data has been used to gauge public mental health trends during pandemics, which supports healthcare providers and policymakers in identifying critical areas for intervention (Zunic et al. 2020).

5.2 Finance (Market sentiment tracking)

In the financial sector, sentiment analysis aids in tracking market sentiment by analyzing news articles, social media, and financial reports. Models trained on financial-specific datasets, such as tweets and news articles, predict stock movements or economic trends by capturing investor and public sentiment, thereby supporting more informed decision-making for investors and policymakers (Dickinson and Hu 2015). Additionally, research has shown that sentiment indicators derived from social media data are strongly correlated with stock price fluctuations, demonstrating the predictive power of sentiment analysis in finance (Smailovic et al. 2013). Hybrid sentiment models, combining lexicon-based and machine learning techniques, enhance the precision of market predictions, particularly under volatile economic conditions (Shayaa et al. 2018).

5.3 E-commerce (Customer feedback analysis)

In e-commerce, sentiment analysis is widely used to understand customer feedback by analyzing reviews and comments. This application helps companies to gauge product sentiment, refine marketing strategies, and improve user experiences by categorizing opinions at both the review and sentence levels. Such insights are pivotal in guiding product development and optimizing customer engagement strategies. Real-time sentiment analysis in e-commerce has been applied to Amazon reviews, allowing companies to adjust products and services dynamically based on consumer needs and preferences (Jabbar et al. 2019).

5.4 Disaster management (Crisis response sentiment analysis)

During natural disasters or crises, understanding public sentiment is crucial for effective disaster management. Sentiment analysis of social media data helps authorities gauge public reactions, misinformation spread, and immediate needs. Real-time sentiment analysis, often using lightweight architectures for rapid response, improves communication strategies and resource allocation during crisis response efforts. For instance, during the COVID-19 pandemic, sentiment analysis was utilized to assess public needs and perceptions, enabling a more adaptive response from disaster management authorities (Behl et al. 2021). This approach not only enhances situational awareness but also aids in resource prioritization and misinformation control (Hou et al. 2020).

5.5 Social media monitoring (Public opinion and trends)

Sentiment analysis has become a key tool for analyzing public opinion on social media. It is widely applied to understand public sentiments on political events, products, and social issues. By leveraging algorithms like Naïve Bayes and Support Vector Machines, companies and governments can monitor sentiment trends in real-time, guiding public relations strategies and policymaking (Sheikh and Jaiswal 2020). This application is critical in today's digitally connected world, where public sentiment on social platforms influences brand perception, election outcomes, and policy formation (Cirqueira et al. 2020).

Each of these applications demonstrates the broad utility of sentiment analysis in making informed decisions across different sectors, enhancing the responsiveness and adaptability of organizations.

6 Conclusion

This comprehensive survey has charted the development of sentiment analysis from foundational machine learning techniques to the cutting edge of deep learning, highlighting a spectrum of innovative methodologies that have enabled

sentiment analysis models to evolve significantly. Our review not only captures the technological progress but also addresses the persistent challenges that continue to shape the field, such as overfitting, domain adaptation, data sparsity, and linguistic diversity. These challenges underscore the ongoing need for models that are robust, adaptable, and capable of navigating the complexity of human emotions as expressed across various digital communications.

Our work advocates for a multidisciplinary approach, underscoring the importance of collaboration across fields like linguistics, cognitive science, and machine learning to build systems that understand nuanced sentiment. By examining key techniques, including regularization, multimodal data integration, adaptive data augmentation, and self-supervised learning, this survey provides a roadmap for advancing sentiment analysis models that generalize effectively across domains while remaining interpretable. The incorporation of these strategies is essential for developing systems that can adapt to domain-specific language and cultural contexts, particularly in applications spanning healthcare, finance, and social media monitoring.

Looking forward, this survey highlights the potential of emerging approaches-such as real-time adaptive data augmentation and multimodal data fusion-to enhance model generalization and sensitivity. By leveraging these insights, future research can contribute to the creation of sentiment analysis tools that are not only robust and efficient but also culturally inclusive and capable of accurately capturing the intricacies of human sentiment in diverse and evolving environments. This work aspires to catalyze further innovation, offering a foundation for researchers to develop the next generation of sentiment analysis systems that can effectively address both the current and future demands of an increasingly complex digital landscape.

Acknowledgements This work was partially supported by the Australian Research Council (grant numbers DP220103717, LE220100078).

Author contributions This work was a collaborative effort by the authors, each of whom contributed substantially to its conception, writing, and revision. KA played a pivotal role in data acquisition and delivering the first draft.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Data availability Not Applicable.

Declarations

Conflict of interests The authors declare that they have no Conflict of interests.

References

Abdullah T, Ahmet A (2022) Deep learning in sentiment analysis: recent architectures. ACM Comput Surv 55(8):1–37

Afriliana N, Iswari N (2022) Suryasari,"sentiment analysis of user-generated content: a bibliometric analysis,". J Syst Manag Sci 12(6):583–598

Agüero-Torales MM, Salas JIA, López-Herrera AG (2021) Deep learning and multilingual sentiment analysis on social media data: an overview. Appl Soft Comput 107:107373. https://doi.org/10.1016/J.ASOC.2021.107373

Ahmed AAA, Agarwal S, Kurniawan IGA, Anantadjaya SP, Krishnan C (2022) Business boosting through sentiment analysis using artificial intelligence approach. Int J Syst Assurance Eng Manag 13(Suppl 1):699–709

Ahmed A, Amin S (2024) A novel approach for sentiment analysis of a low-resource language using lstm. J Comput Linguist Appl 12(1):45–58

Alsayat A (2022) Improving sentiment analysis for social media applications using an ensemble deep learning language model. Arab J Sci Eng 47(2):2499–2511

Alzantot M, Sharma Y, Elgohary A, Ho B-J, Srivastava M, Chang K-W (2018) Generating natural language adversarial examples. arXiv preprint arXiv:1804.07998

An W, Tian F, Chen P, Zheng Q (2022) Aspect-based sentiment analysis with heterogeneous graph neural network. IEEE Trans Comput Soc Syst 10(1):403–412

Ansari G, Ahmad T, Doja MN (2019) Hybrid filter-wrapper feature selection method for sentiment classification. Arab J Sci Eng 44:9191–9208

Appel O, Chiclana F, Carter J, Fujita H (2016) A hybrid approach to the sentiment analysis problem at the sentence level. Knowl-Based Syst 108:110–124

Bach NX, Hai VT, Phuong TM (2016) Cross-domain sentiment classification with word embeddings and canonical correlation analysis. In: Proceedings of the 7th Symposium on Information and Communication Technology, pp 159–166

Badawi SS (2023) Using multilingual bidirectional encoder representations from transformers on medical corpus for kurdish text classification. Aro-Sci J Koya Univ 11(1):10–15

Baecchi C, Uricchio T, Bertini M, Del Bimbo A (2016) A multimodal feature learning approach for sentiment analysis of social network multimedia. Multimed Tools Appl 75:2507–2525

Barham S, Feizi S (2019) Interpretable adversarial training for text. arXiv preprint arXiv:1905.12864

Behl S, Rao A, Aggarwal S, Chadha S, Pannu H (2021) Twitter for disaster relief through sentiment analysis for covid-19 and natural hazard crises. Int J Disaster Risk Reduct 55:102101

Belghazi MI, Baratin A, Rajeshwar S, Ozair S, Bengio Y, Courville A, Hjelm D (2018) Mutual information neural estimation. In: International Conference on Machine Learning, pp 531–540. PMLR

Bengio Y, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35(8):1798–1828
Bharti P, Sagar V, Wadhwa B (2022) An analysis on sentiments using deep learning approaches. In: 2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), pp 355–360. IEEE
Bishop CM, Nasrabadi NM (2006) Pattern Recognition and Machine Learning vol. 4. Springer
Bowman S.R, Vilnis L, Vinyals O, Dai A, Jozefowicz R, Bengio S (2016) Generating sentences from a continuous space, arxiv. arXiv preprint Arxiv:1511.06349
Cambria E, Schuller B, Xia Y, Havasi C (2013) New avenues in opinion mining and sentiment analysis. IEEE Intell Syst 28(2):15–21
Caruana R (1997) Multitask learning. Mach Learn 28:41–75
Chatterjee A, Gupta U, Chinnakotla MK, Srikanth R, Galley M, Agrawal P (2019) Understanding emotions in text using deep learning and big data. Comput Hum Behav 93:309–317
Chen R, Zhou W, Li Y, Zhou H (2022) Video-based cross-modal auxiliary network for multimodal sentiment analysis. IEEE Trans Circuits Syst Video Technol 32(12):8703–8716
Chen B, Huang Q, Chen Y, Cheng L, Chen R (2018) Deep neural networks for multi-class sentiment classification. In: 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), pp 854–859. IEEE
Chen T, Hong L, Shi Y, Sun Y (2017) Joint text embedding for personalized content-based recommendation. arXiv preprint arXiv:1706.01084
Cirqueira D, Almeida F, Cakir G, Jacob A, Lobato F, Bezbradica M, Helfert M (2020) Explainable sentiment analysis application for social media crisis management in retail. In: International Conference on Computer-Human Interaction Research and Applications
Collobert R, Weston J (2008) A unified architecture for natural language processing: deep neural networks with multitask learning. In: Proceedings of the 25th International Conference on Machine Learning, pp 160–167
Corbeil J-P, Ghadivel HA (2020) Bet: a backtranslation approach for easy data augmentation in transformer-based paraphrase identification context. arXiv preprint arXiv:2009.12452
Councill IG, McDonald R, Velikovich L (2010) What's great and what's not: Learning to classify the scope of negation for improved sentiment analysis. In: Proceedings of the ACL Conference on Human Language Technologies, pp 51–59
Da Y, Bossa MN, Berenguer AD, Sahli H (2024) Reducing bias in sentiment analysis models through causal mediation analysis and targeted counterfactual training. IEEE Access 12:10120–10134. https://doi.org/10.1109/ACCESS.2024.3353056
Das R, Singh TD (2023) Multimodal sentiment analysis: a survey of methods, trends and challenges. ACM Comput Surv 55:1–38
Demotte P, Wijegunarathna K, Meedeniya D, Perera I (2023) Enhanced sentiment extraction architecture for social media content analysis using capsule networks. Multimed Tools Appl 82(6):8665–8690
Deng S, Sinha AP, Zhao H (2017) Adapting sentiment lexicons to domain-specific social media texts. Decis Support Syst 94:65–76
Diaz M, Johnson I.L, Lazar A, Piper A.M, Gergle D (2018) Addressing age-related bias in sentiment analysis. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3173574.3173986
Dickinson B, Hu W-G (2015) Sentiment analysis of investor opinions on twitter. Soc Netw 4:62–71
Ding W, Song X, Guo L, Xiong Z, Hu X (2013) A novel hybrid hdp-lda model for sentiment analysis. In: 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), vol. 1, pp 329–336. IEEE
Dong Y, Fu Y, Wang L, Chen Y, Dong Y, Li J (2020) A sentiment analysis method of capsule network based on bilstm. IEEE Access 8:37014–37020
Dong X, De Melo G (2018) Cross-lingual propagation for deep sentiment analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32
Draskovic D, Zecevic D, Nikolic B (2022) Development of a multilingual model for machine sentiment analysis in the serbian language. Mathematics 10(18):3236
Fadel R, Öz K (2020) A sentiment analysis model for terrorist attacks reviews on twitter using a hybrid lexicon-machine learning approach. In: International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp 163–170
Fei Y (2016) Simultaneous support vector selection and parameter optimization using support vector machines for sentiment classification. In: 2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS), pp 59–62. IEEE
Filatova E (2012) Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)
Frankle J, Carbin M (2018) The lottery ticket hypothesis: Finding sparse, trainable neural networks. arXiv preprint arXiv:1803.03635
Gal Y, Ghahramani Z (2016) Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning, pp 1050–1059. PMLR
Ghanem R, Erbay H (2020) Context-dependent model for spam detection on social networks. SN Appl Sci 2:1–8
Ghasemi R, Ashrafi Asli SA, Momtazi S (2022) Deep persian sentiment analysis: cross-lingual training for low-resource languages. J Inf Sci 48(4):449–462
Gkoumas D, Li Q, Yu Y, Song D (2021) An entanglement-driven fusion neural network for video sentiment analysis. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, pp 1736–1742. International Joint Conferences on Artificial Intelligence Organization
Goodfellow IJ, Shlens J, Szegedy C (2014) Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572
Gopalakrishnan V, Ramaswamy C (2014) Sentiment learning from imbalanced dataset: an ensemble based method. Int J Artif Intell 12(2):75–87
Grandvalet Y, Bengio Y (2006) Entropy Regularization
Guan M, Li F, Xue Y (2023) Enhanced syntactic and semantic graph convolutional network with contrastive learning for aspect-based sentiment analysis. IEEE Trans Comput Soc Syst 11(1):859–870
Guo H (2020) Nonlinear mixup: Out-of-manifold data augmentation for text classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp 4044–4051
Gupta CP, Kumar VR (2023) Sentiment analysis and its application in analysing consumer behaviour. In: 2023 International Conference on Emerging Techniques in Computational Intelligence (ICETCI), pp 332–337. IEEE
Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, Smith NA (2020) Don't stop pretraining: adapt language models to domains and tasks. arXiv preprint arXiv:2004.10964
Hansen LK (1990) Salamon P: neural network ensembles. IEEE Trans Pattern Anal Mach Intell 12(10):993–1001
Hasan M, Islam I, Hasan KA (2019) Sentiment analysis using out of core learning. In: 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), pp 1–6. IEEE

Hayatin N, Alias S, Lai PH, Sainin MS (2022) Sentiment analysis based on probabilistic classifier techniques in various indonesian review data. Jordanian J Comput Inform Technol **8**(3)

Hou Q, Han M, Cai Z (2020) Survey on data analysis in social media: a practical application aspect. Big Data Min Anal 3(4):259–279

Hu L, Zhang M, Li S, Shi J, Shi C, Yang C, Liu Z (2021) Text-graph enhanced knowledge graph representation learning. Front Artif Intell 4:697856

Huang F, Zhang X, Zhao Z, Xu J, Li Z (2019) Image-text sentiment analysis via deep multimodal attentive fusion. Knowl-Based Syst 167:26–37

Huddar MG, Sannakki SS, Rajpurohit VS (2020) Multi-level context extraction and attention-based contextual inter-modal fusion for multimodal sentiment analysis and emotion classification. Int J Multimed Inform Retr 9:103–112

Hutto CJ, Gilbert E (2014) Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the International AAAI Conference on Web and Social Media, pp 216–225

Iqbal F, Hashmi JM, Fung BC, Batool R, Khattak AM, Aleem S, Hung PC (2019) A hybrid framework for sentiment analysis using genetic algorithm based feature reduction. IEEE Access 7:14637–14652

Iqbal M, Karim A, Kamiran F(2015) Bias-aware lexicon-based sentiment analysis. In: Proceedings of the 30th Annual ACM Symposium on Applied Computing. https://doi.org/10.1145/2695664.2695759

Iyyer M, Wieting J, Gimpel K, Zettlemoyer L (2018) Adversarial example generation with syntactically controlled paraphrase networks. arXiv preprint arXiv:1804.06059

Jabbar J, Urooj I, JunSheng W, Azeem N (2019) Real-time sentiment analysis on e-commerce application. In: 2019 IEEE 16th International Conference on Networking, Sensing and Control (ICNSC), pp 391–396. IEEE

Jeon K, Oh J (2020) Hybrid recursive feature elimination for feature selection. In: Proceedings of the 2020 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp 1–6. IEEE

Jia R, Liang P (2017) Adversarial examples for evaluating reading comprehension systems. arXiv preprint arXiv:1707.07328

Jiang T, Sun W, Wang M (2023) Msgat-based sentiment analysis for e-commerce. Information 14(7):416

Jin L, Zhang L (2022) Discriminant variance criterion for sentiment analysis. In: 2022 26th International Conference on Pattern Recognition (ICPR), pp 3056–3062. IEEE

Jin D, Jin Z, Zhou JT, Szolovits P (2020) Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp 8018–8025

Kanojia D, Joshi A (2023) Applications and challenges of sentiment analysis in real-life scenarios. ArXiv **abs/2301.09912**

Karimi A, Rossi L, Prati A (2021) Adversarial training for aspect-based sentiment analysis with bert. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp 8797–8803. IEEE

Khan S, Shahid M (2022) Hindi/bengali sentiment analysis using transfer learning and joint dual input learning with self attention. arXiv preprint arXiv:2202.05457

Kiela D, Bhooshan S, Firooz H, Perez E, Testuggine D (2019) Supervised multimodal bitransformers for classifying images and text. arXiv preprint arXiv:1909.02950

Kim Y (2014) Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882

Kitada S, Iyatomi H (2021) Attention meets perturbations: robust and interpretable attention with adversarial training. IEEE Access 9:92974–92985

Kitada S, Iyatomi H (2023) Making attention mechanisms more robust and interpretable with virtual adversarial training. Appl Intell 53(12):15802–15817

Kobayashi S (2018) Contextual augmentation: Data augmentation by words with paradigmatic relations. arXiv preprint arXiv:1805.06201

Kristiyanti DA, Sitanggang IS (2023) Nurdiati S: feature selection using new version of v-shaped transfer function for salp swarm algorithm in sentiment analysis. Computation 11(3):56

Krogh A, Hertz J(1991) A simple weight decay can improve generalization. Adv Neural Inform Process Syst **4**

Krosuri LR, Aravapalli RS (2023) Novel heuristic-based hybrid resnext with recurrent neural network to handle multi class classification of sentiment analysis. Mach Learn Sci Technol 4(1):015033

Lai H, Yan X (2022) Multimodal sentiment analysis with asymmetric window multi-attentions. Multimed Tools Appl 81(14):19415–19428

Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R (2019) Albert: a lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942

Le Q, Mikolov T (2014) Distributed representations of sentences and documents. In: International Conference on Machine Learning, pp 1188–1196. PMLR

Li F, Yang Y, Xing E (2005) From lasso regression to feature vector machine. Adv Neural Inform Process Syst **18**

Lina X, Hamdulla A, Ablimit M, Sijie W (2024) Adversarial training for uncertainty estimation in cross-lingual text classification. In: 2024 International Joint Conference on Neural Networks (IJCNN), pp 1–7. IEEE

Liu B (2020) Sentiment analysis: mining opinions, sentiments, and emotions. Cambridge University Press, Cambridge

Liu H, Chatterjee I, Zhou M, Lu XS, Abusorrah A (2020) Aspect-based sentiment analysis: a survey of deep learning methods. IEEE Trans Comput Soc Syst 7(6):1358–1375

Liu P, Zheng X, Li H, Liu J, Ren Y, Zhu H, Sun L (2023) Improving the modality representation with multi-view contrastive learning for multimodal sentiment analysis. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 1–5. IEEE

Liu J, Li K, Zhu A, Hong B, Zhao P, Dai S, Wei C, Huang W, Su H (2024) Application of deep learning-based natural language processing in multilingual sentiment analysis. Mediterr J Basic Appl Sci https://doi.org/10.46382/mjbas.2024.8219

Liu X, He P, Chen W, Gao J (2019) Multi-task deep neural networks for natural language understanding. arXiv preprint arXiv:1901.11504

Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: a robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692

Liu Y, Lapata M (2019) Text summarization with pretrained encoders. arXiv preprint arXiv:1908.08345

Lopes V, Gaspar A, Alexandre LA, Cordeiro J (2021) An automl-based approach to multimodal image sentiment analysis. In: 2021 International Joint Conference on Neural Networks (IJCNN), pp 1–9. IEEE

Mabokela K, Çelik T, Raborife M (2023) Multilingual sentiment analysis for under-resourced languages: a systematic review of the landscape. IEEE Access 11:15996–16020. https://doi.org/10.1109/ACCESS.2022.3224136

Manikandan G, Susi E, Abirami S (2018) Flexible-fuzzy mutual information based feature selection on high dimensional data. In: 2018 Tenth International Conference on Advanced Computing (ICoAC), 237–243. https://doi.org/10.1109/icoac44903.2018.8939115

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781

Minanovic A, Gabelica H, Krstić Ž (2014) Big data and sentiment analysis using knime: Online reviews vs. social media. In: 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp 1464–1468. IEEE

Mohammad SM, Turney PD (2013) Crowdsourcing a word-emotion association lexicon. In: Proceedings of the NAACL-HLT Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp 347–355

Murayama K, Kawano S (2021) Sparse bayesian learning with weakly informative hyperprior and extended predictive information criterion. In: IEEE Transactions on Neural Networks and Learning Systems

Nagpal A, Gaur D, Gaur S (2014) Feature selection using mutual information for high- dimensional data sets. In: 2014 IEEE International Advance Computing Conference (IACC), 45–49

Nelakurthi AR, Tong H, Maciejewski R, Bliss N, He J (2017) User-guided cross-domain sentiment classification. In: Proceedings of the 2017 SIAM International Conference on Data Mining, pp 471–479. SIAM

Nitish S (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15:1

Niu L, Zheng Q, Zhang L (2021) Enhance gated graph neural network with syntactic for sentiment analysis. In: 2021 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA), pp 1055–1060. IEEE

Niven T, Kao H-Y (2019) Probing neural network comprehension of natural language arguments. arXiv preprint arXiv:1907.07355

Ombabi AH, Ouarda W, Alimi AM (2024) Improving arabic sentiment analysis across context-aware attention deep model based on natural language processing. Lang Resour Eval 1–25

Omran T, Sharef B, Grosan C, Li Y (2023) The impact of data augmentation on sentiment analysis of translated textual data. In: 2023 International Conference on IT Innovation and Knowledge Discovery (ITIKD), pp 1–4. IEEE

Oueslati O, Cambria E, Hajhmida M, Ounelli H (2020) A review of sentiment analysis research in arabic language. Future Gener Comput Syst 112:408–430

Pang B, Lee L, Vaithyanathan S (2002) Thumbs up? sentiment classification using machine learning techniques. arXiv preprint cs/0205070

Parlar T, Sarac E (2019) Iwd based feature selection algorithm for sentiment analysis. Elektronika ir Elektrotechnika 25(1):54–58

Phan KT-K, Hao DN, Van Thin D, Nguyen NL.-T (2021) Exploring zero-shot cross-lingual aspect-based sentiment analysis using pre-trained multilingual language models. In: 2021 International Conference on Multimedia Analysis and Pattern Recognition (MAPR), pp 1–6. IEEE

Plisiecki H, Lenartowicz P, Flakus M, Pokropek A (2024) Uncovering political bias in emotion inference models: implications for sentiment analysis in social science research. ArXiv https://doi.org/10.48550/arXiv.2407.13891

Prottasha NJ, Sami AA, Kowsher M, Murad SA, Bairagi AK, Masud M, Baz M (2022) Transfer learning for sentiment analysis using bert based supervised fine-tuning. Sensors 22(11):4157

Qian F, Han J, He Y, Zheng T, Zheng G (2023) Sentiment knowledge enhanced self-supervised learning for multimodal sentiment analysis. In: Findings of the Association for Computational Linguistics: ACL 2023, pp 12966–12978

Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language models are unsupervised multitask learners. OpenAI blog 1(8):9

Rastogi S (2023) Weak supervision and transformed-based sentiment analysis on multi-lingual data. In: 2023 15th International Conference on COMmunication Systems & NETworkS (COMSNETS), pp 706–712. IEEE

Rathod N, Mistry N, Talati D, Parikh M, Kore A, Kanani P (2022) Marathi social media opinion mining using xlm-r. In: 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC), pp 730–736. IEEE

Reimers N, Gurevych I (2017) Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. arXiv preprint arXiv:1707.09861

Robnik-Sikonja M, Reba K, Mozetic I (2020) Cross-lingual transfer of sentiment classifiers. arXiv preprint arXiv:2005.07456

Rozado D (2020) Wide range screening of algorithmic bias in word embedding models using large sentiment lexicons reveals underreported bias types. PLoS ONE **15**.https://doi.org/10.1371/journal.pone.0231189

Ruder S, Bingel J, Augenstein I, Søgaard A (2019) Latent multi-task architecture learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp 4822–4829

Sahar A, Ayoub M, Hussain S, Yu Y, Khan A (2022) Transfer learning-based framework for sentiment classification of cosmetics products reviews. Pak J Eng Technol 5(3):38–43

Schouten K, Frasincar F (2015) Survey on aspect-level sentiment analysis. IEEE Trans Knowl Data Eng 28(3):813–830

Shayaa S, Jaafar N, Bahri S, Sulaiman A, Wai PS, Chung YW, Piprani AZ, Al-garadi M (2018) Sentiment analysis of big data: methods, applications, and open challenges. IEEE Access 6:37807–37827

Sheikh HA, Jaiswal J (2020) Implementing sentiment analysis on real-time twitter data. J Emerg Technol Innov Res

Siddiqua A, Ahsan K (2016) Combining a rule-based classifier with weakly supervised learning for sentiment analysis. In: International Conference on Data Science and Advanced Analytics (DSAA), pp 120–128

Smailovic J, Grcar M, Lavrač N, Znidarsic M (2013) Predictive sentiment analysis of tweets: A stock market application, 77–88

Socher R, Perelygin A, Wu J, Chuang J, Manning CD, Ng AY, Potts C (2013) Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp 1631–1642

Srivastava A et al (2022) Improving sentiment analysis with neural networks. In: International Conference on Knowledge-Based Organization

Taboada M, Brooke J, Tofiloski M, Voll K, Stede M (2011) Lexicon-based methods for sentiment analysis. Comput Linguist 37(2):267–307

Tai KS, Socher R, Manning CD (2015) Improved semantic representations from tree-structured long short-term memory networks. arXiv preprint arXiv:1503.00075

Tan KL, Lee CP, Lim KM (2023) Roberta-gru: a hybrid deep learning model for enhanced sentiment analysis. Appl Sci 13(6):3915

Tan KL, Lee CP, Lim KM, Anbananthen KSM (2022) Sentiment analysis with ensemble hybrid deep learning model. IEEE Access 10:103694–103704

Tripathy A, Agrawal A, Rath SK (2016) Classification of sentiment reviews using n-gram machine learning approach. Expert Syst Appl 57:117–126

Trivedi UB, Bhatt M, Srivastava P (2021) Prevent overfitting problem in machine learning: a case focus on linear regression and logistics regression. In: Innovations in Information and Communication Technologies (IICT-2020) Proceedings of International Conference on ICRIHE-2020, Delhi, India: IICT-2020, pp 345–349. Springer

Utama P, Weir N, Basik F, Binnig C, Cetintemel U, Hättasch B, Ilkhechi A, Ramaswamy S, Usta A (2018) An end-to-end neural natural language interface for databases. arXiv preprint arXiv:1804.00401

Wang Z, Gan H-S (2023) Multi-level adversarial training for stock sentiment prediction. In: 2023 IEEE 3rd International Conference on Computer Communication and Artificial Intelligence (CCAI), pp 127–134. IEEE

Wang Y, Sun A, Han J, Liu Y, Zhu X (2018) Sentiment analysis by capsules. In: Proceedings of the 2018 World Wide Web Conference, pp 1165–1174

Wang Y, Huang M, Zhu X, Zhao L (2016) Attention-based lstm for aspect-level sentiment classification. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp 606–615

Wei J, Zou K (2019) Eda: Easy data augmentation techniques for boosting performance on text classification tasks. arXiv preprint arXiv:1901.11196

Wolpert DH (1992) Stacked generalization. Neural Netw 5(2):241–259

Wong EP, M'hiri F (2020) Analyzing language bias between french and english in conventional multilingual sentiment analysis models. https://doi.org/10.48550/arXiv.2405.06692

Xiang R, Chersoni E, Lu Q, Huang C-R, Li W, Long Y (2021) Lexical data augmentation for sentiment analysis. J Am Soc Inf Sci 72(11):1432–1447

Xue L, Constant N, Roberts A, Kale M, Al-Rfou R, Siddhant A, Barua A, Raffel C (2020) mt5: A massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv:2010.11934

Yin W, Schütze H (2016) Multichannel variable-size convolution for sentence classification. arXiv preprint arXiv:1603.04513

Ying X (2019) An overview of overfitting and its solutions. In: Journal of Physics Conference Series, vol. 1168, p 022022. IOP Publishing

Yu L, Zhang W, Wang J, Yu Y (2017) Seqgan: sequence generative adversarial nets with policy gradient. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31

Yue L, Chen W, Li X, Zuo W, Yin M (2019) A survey of sentiment analysis in social media. Knowl Inf Syst 60:617–663

Zhang P, He Z (2013) A weakly supervised approach to chinese sentiment classification using partitioned self-training. J Inf Sci 39(6):815–831

Zhang B, Xu X, Yang M, Chen X, Ye Y (2018) Cross-domain sentiment classification by capsule network with semantic rules. IEEE Access 6:58284–58294

Zhang Y, Zhang Y, Guo W, Cai X, Yuan X (2022) Learning disentangled representation for multimodal cross-domain sentiment analysis. In: IEEE Transactions on Neural Networks and Learning Systems

Zhang S, Liu H, Yang L, Lin H (2015) A cross-domain sentiment classification method based on extraction of key sentiment sentence. In: Natural Language Processing and Chinese Computing: 4th CCF Conference, NLPCC 2015, Nanchang, China, October 9-13, 2015, Proceedings 4, pp 90–101. Springer

Zhao C, Wu M, Yang X, Zhang W, Zhang S, Wang S, Li D (2024) A systematic review of cross-lingual sentiment analysis: tasks, strategies, and prospects. ACM Comput Surv 56:1–37. https://doi.org/10.1145/3645106

Zhao Q, Yang F, An D, Lian J (2024) Modeling structured dependency tree with graph convolutional networks for aspect-level sentiment classification. Sensors. https://doi.org/10.3390/s24020418

Zheng Y, Kwoh C (2011) A feature subset selection method based on high-dimensional mutual information. Entropy 13:860–901. https://doi.org/10.3390/e13040860

Zhu X, Chen Y, Gu Y, Xiao Z (2022) Sentimedqaer: a transfer learning-based sentiment-aware model for biomedical question answering. Front Neurorobot 16:773329

Zhu H, Zheng Z, Soleymani M, Nevatia R (2022) Self-supervised learning for sentiment analysis via image-text matching. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 1710–1714. IEEE

Zunic A, Corcoran P, Spasic I (2020) Sentiment analysis in health and well-being: systematic review. JMIR Med Inform 8(1):16023