International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 04 | Apr 2025 www.irjet.net p-ISSN: 2395-0072

Evaluation Metrics for Sentiment Analysis: A Comprehensive Review and Future Directions

Palak Kaushal[1], Anita Ganpati[2]

[1] *Palak Kaushal, Himachal Pradesh University, Shimla, India*
[2]*Anita Ganpati, Himachal Pradesh University, Shimla, India*

Abstract - *Evaluation metrics are crucial for assessing the performance and reliability of sentiment analysis models in various applications. Evaluation metrics are critical for appraising sentiment analysis models performance and guaranteeing their dependability in various applications. The research thoroughly examined the classification, regression, ranking, and explainability metrics. Every measure has advantages and disadvantages that affect how well sentiment categorisation and forecasting assignments work. These measures are compared, providing insight into their effectiveness in various sentiment analysis contexts. Future studies should concentrate on fairness-driven and context-aware assessment methods to expand the reliability and interpretability of classification models.*

***Key Words*: Sentiment Analysis, Evaluation Metrics, Classification Metrics, Regression Metrics, Ranking Metrics.**

1. INTRODUCTION

Sentiment Analysis, a branch of Natural Language Processing (NLP)[1], aims to construe and assess emotions, opinions, and attitudes conveyed in textual data[2]. Sentiment analysis is becoming a crucial tool in many fields, such as business intelligence, consumer feedback analysis, and social media monitoring, due to the explosive expansion of digital material on social media, e-commerce platforms, and online reviews. Businesses utilise sentiment analysis to gain insight into public opinion, make better decisions, and improve user experience. Assessing the usefulness and dependability of sentiment categorisation models is a crucial component of sentiment analysis. A model's performance in numerous tasks is resolute by the assessment measures it uses. Different criteria for evaluation are needed depending on whether the task requires classification, regression, or ranking.

Moreover, explainability and fairness measures have become more crucial for ensuring transparency and objective judgements because advanced learning and black-box models are used more often in sentiment analysis. Classification metrics, regression metrics, ranking metrics, and explainability/fairness measurements are the four primary categories into which this study divides sentiment analysis assessment metrics. This study intends to assist researchers in choosing suitable assessment techniques depending on the nature of their sentiment analysis jobs by offering an organised summary of various measures.

2. LITERATURE REVIEW

Sentiment analysis model assessment has been deeply studied and several measures have been put out to evaluate performance on tasks involving classification, regression, and ranking. The important research that has influenced the creation of sentiment analysis assessment metrics is reviewed in this section. Most popular sentiment analysis responsibility is sentiment investigation, which is classifying text into predetermined sentiment categories. Using accuracy as the core assessment criterion, [2]used conventional learning classifiers, such as Naive Bayes, Support Vector Machines (SVM), and Maximum Entropy. Precision, recall, and F1-score have been adopted as more informative metrics because of the criticism of accuracy's shortcomings in unbalanced datasets [3].

Deep learning-based sentiment classification has further emphasized the need for robust evaluation metrics. Long Short-Term Memory (LSTMs) networks, introduced by[4]established strong performance in capturing contextual dependencies in sentiment classification. The paper[5] evaluates sentiment classification using accuracy, precision, recall, F1-score, and ROC-AUC. The ensemble model outperformed individual classifiers, achieving high F1-score and AUC, reducing misclassification, and improving sentiment detection in Arabic social media text. The paper[6] evaluates sentiment analysis models using accuracy, precision, recall, and F1-score to compare their effectiveness in recognizing emotional content. The results highlight that deep learning models beat traditional processes, achieving higher precision and recall, making them more suitable for sentiment classification tasks.

3. SENTIMENT ANALYSIS LEVELS

3.1 Document-Level Sentiment Analysis

Determines the overall sentiment of an entire document (e.g., a product review, blog post)[7].
Use Case: Classifying movie reviews as positive or negative.
Limitation: Fails to detect multiple sentiments in longer texts.

3.2 Sentence-Level Sentiment Analysis

Analyzes sentiment expressed in individual sentences[8].
Use Case: Twitter sentiment classification, headline analysis.
Challenge: Detecting sarcasm or implicit sentiment in short texts.

3. 3 Aspect-Level (or Feature-Level) Sentiment Analysis

Identifies sentiment toward specific aspects/features of a product or service within text[9].
Use Case: In a review like *"The camera is amazing, but the battery life is poor,"* aspect-level analysis can tag "camera" as positive and "battery" as negative.
Strength: Provides granular insights for businesses.

3.4 Phrase-Level Sentiment Analysis

Assigns sentiment polarity to smaller syntactic units like phrases[10].
Use Case: *"Not very good"* → negative sentiment at phrase level, though individual words may suggest otherwise.
Challenge: Requires parsing and understanding modifiers and negation.

Table 1: Shows the level of sentiment with use cases and granularity.

Level	Granularity	Use Case
Document-Level[7]	Entire document	Product reviews
Sentence-Level[8]	Individual sentence	Tweets, headlines
Aspect-Level[9]	Specific feature	Product aspect feedback
Phrase-Level[10]	Word/phrase	Negation handling

4. SENTIMENT ANALYSIS TECHNIQUES

4.1 Lexicon-Based Techniques

Use predefined dictionaries of words where each word is associated with a sentiment score (positive, negative, neutral)[11].

Types:

- *Dictionary-based*: Manually curated (e.g., SentiWordNet, NRC).
- *Corpus-based*: Scores derived from large corpora using statistical or co-occurrence methods.

Strengths: Language-agnostic, interpretable.
Limitations: Struggles with sarcasm, negation, and domain-specific terms.

4.2 Machine Learning-Based Techniques

Use traditional supervised learning algorithms to train sentiment classifiers on labeled data[2].

Common algorithms: Naive Bayes, SVM, Logistic Regression, Decision Trees.
Steps: Feature extraction (e.g., Bag-of-Words, TF-IDF) → Model training → Prediction.
Strengths: Adaptable to specific datasets.
Limitations: Requires large labeled datasets, less interpretable.

4.3 Deep Learning-Based Techniques

Automatically learn complex patterns in text using neural networks[12].

Popular architectures:

- **CNN**: Captures local word patterns and phrases.
- **RNN/LSTM/GRU**: Captures sequential dependencies in text.
- **Attention Mechanisms**: Focus on important words.

Strengths: Outperforms traditional models, captures context.

Limitations: Computationally expensive, needs large data.

4.4 Transformer-Based Techniques

Use pre-trained language models like BERT, RoBERTa, and DistilBERT fine-tuned for sentiment classification[13].

Advantage: Understand bidirectional context and semantic relationships.
Examples: BERT fine-tuned on IMDb, SST-2, Twitter sentiment datasets.
Strengths: State-of-the-art accuracy, minimal feature engineering.
Limitations: Requires high computational resources.

4.5 Hybrid Approaches

Combine lexicon-based and machine/deep learning methods to leverage the strengths of both[14].

Use Case: Lexicon helps with interpretability, ML/DL enhances accuracy.
Example: Use lexicon scores as features in an ML classifier.

Table 2: defines the technique, description and key methods.

Technique	Description	Key Methods
Lexicon-Based[11]	Uses predefined word lists	SentiWordNet, NRC
Machine Learning[15]	Supervised learning with features	Naive Bayes, SVM
Deep Learning[12]	Neural networks for sequential/semantic learning	CNN, LSTM, GRU
Transformer-Based[13]	Contextual language models	BERT, RoBERTa
Hybrid[14]	Combines lexicon + ML/DL	Ensemble, lexicon features

5. PERFORMANCE EVALUATION METRICS

5.1 Classification Metrics[16]

Accuracy

Accuracy is the most used metric in sentiment analysis and measures the proportion of correctly classified instances[17].

$$\mathbf{Acc} = (\mathbf{Tp} + \mathbf{Tn})/(\mathbf{Tp} + \mathbf{Tn} + \mathbf{Fp} + \mathbf{Fn})$$

Accuracy is useful for balanced datasets, but it is unreliable for imbalanced sentiment datasets.

Precision, Recall, and F1-Score

Precision measures how many predicted positive occurrences are positive:

$$\mathbf{Precision} = \mathbf{Tp}/(\mathbf{Tp} + \mathbf{Fp})$$

Recall (or Sensitivity) evaluates how many actual positive instances are correctly classified

$$\mathbf{Rec} = \mathbf{Tp}/(\mathbf{Tp} + \mathbf{Fn})$$

F1-Score stabilities, precision, and recall, providing a solitary presentation measure

$$\mathbf{F1-Score} = (\mathbf{2} \times \mathbf{Precision} \times \mathbf{Recall})/(\mathbf{Precision} + \mathbf{Recall})$$

F1-score is particularly important for datasets with class imbalances. F1-score is widely used in sentiment classification benchmarks[18].

ROC-AUC

ROC-AUC appraises the balance between exact positive rate (TPR) and incorrect positive rate (FPR) across different classification thresholds[19]. It is useful for comparing model performance across different decision boundaries but may not be well-suited for multi-class sentiment classification.

5.2 Regression Metrics for Sentiment Scoring[20]

Some tasks assign sentiment scores rather than discrete classes. In such cases, regression metrics are used.

Mean Squared Error (MSE) and Root Mean Squared Error (RMSE)

The mean squared variance among observed and expected sentiment evaluations is measured by MSE.

$$\mathbf{MSE}=\mathbf{1}/\mathbf{N}\ \Sigma\_\mathbf{i}\hat{}\mathbf{N} = (\mathbf{yi} - \hat{\mathbf{y}}\mathbf{i})\hat{}(\mathbf{2}\ \cdot)$$

A measure of transparent error in the same unit as original sentiment ratings is provided by RMSE, which is the square root of MSE[21].

Mean Absolute Error (MAE)

The ratio percentage differences between estimated and observed sentiment ratings are computed[22]. MAE is more aggressive against outliers associated with MSE, but it fails to compensate for substantial mistakes as well.

$$\mathbf{MAE} = \mathbf{1}/\mathbf{N}\ \sum\_\mathbf{i}\hat{}\mathbf{N} = (\mathbf{yi} - \hat{\mathbf{y}}\mathbf{i})$$

5.3 Ranking Metrics for Sentiment Ordering

Metrics for ranking evaluate algorithms that forecast sentiment rankings rather than classifications.

Spearman's Rank Correlation Coefficient

The degree to which sentiment rankings remain consistent within expected and observed levels of sentiment is assessed by Spearman's correlation[23]. When sentiment needs to be ranked instead of defined, it is helpful.

Kendall's Tau

Kendall's Tau is a ranking statistic that evaluates how well projected sentiment rankings match ground truth rankings suited for aspect-based sentiment analysis[24].

5.4 Explainability and Fairness Metrics

SHapley Additive Explanations (SHAP)

The contribution of each feature to sentiment categorisation is explained by SHAP values, which aid in the interpretation of model decisions[25]. It raises the model's integrity and openness.

Fairness Metrics

Fairness is particularly imperative when assessing views from individuals. Independent effect and chance variance are two examples of detection of bias that evaluate how well algorithms handle all sentiment groups[26].

4. COMPARATIVE ANALYSIS OF EVALUATION METRICS FOR SENTIMENT ANALYSIS

Metric Category	Metrics	Strengths	Limitations	Best Use Cases
Classification Metrics	Accuracy	Simple and intuitive	Fails with imbalanced data	Binary and multi-class sentiment classification
	Precision	Reduces false positives	May lower recall	When false positives are costly (e.g., fake reviews)
	Recall	Captures false negatives	May lower precision	Detecting negative sentiments in critical domains
	F1-score	Balances precision & recall	Can be misleading if classes are imbalanced	Sentiment models with uneven class distribution
	ROC-AUC	Measures overall classifier performance	Not ideal for highly imbalanced datasets	Evaluating binary sentiment classification models
Regression Metrics	MSE	Penalizes large errors	Sensitive to outliers	Sentiment score prediction
	RMSE	Interpretable error magnitude	Overemphasizes large errors	Evaluating continuous sentiment scores
	MAE	Less sensitive to outliers than RMSE	Can be misleading in skewed distributions	General sentiment intensity analysis
Ranking Metrics	Spearman's Correlation	Measures monotone relationship	Fails with non-monotonic trends	Rank-based sentiment evaluation
	Kendall's Tau	Captures rank correlation	Less commonly used in NLP	Sentence-level sentiment ranking
Explainability Metrics	SHAP Values	Identifies feature importance	Computationally expensive	Understanding model decisions
	Fairness Measures	Detects bias in sentiment predictions	Still evolving as a standard practice	Ensuring unbiased sentiment analysis

Fig. 1: Comparative Analysis of Evaluation Metrics for Sentiment Analysis

The comparison study demonstrates that there is not a single metric that works best for analysing sentiment. MSE/RMSE are frequently used for sentimentality rating estimates, but they are disposed to outliers, whereas F1-score is the best method for unbalanced issues in classification.

While fairness measures and SHAP parameters improve model interpretability and bias recognition, ranking metrics such as Spearman's correlation are helpful for aspect-based sentiment assessment. A hybrid evaluation strategy that incorporates many measures guarantees a more accurate evaluation. To increase sentiment analysis's accuracy and resilience, future studies need to focus on context-aware, fairness-driven, and domain-specific evaluation methods.

6. CONCLUSION

Sentiment analysis models must be evaluated using the right measures to guarantee dependability across various professions. Key assessment criteria were discussed in this research, with an emphasis on their advantages and disadvantages. Multilingual emotion, sarcasm, and socioeconomic disparity are still major issues. To increase the effectiveness and fairness of sentiment models, future studies should concentrate on context-aware, bias-mitigating, and hybrid assessment techniques.

7. REFERENCES

[1] "Christopher_D._Manning_Hinrich_Schütze_Foundations_Of_Statistical_Natural_Language_Processing".

[2] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques." [Online]. Available: http://reviews.imdb.com/Reviews/

[3] F. Sebastiani, "Machine Learning in Automated Text Categorization," 2001. [Online]. Available: http://liinwww.ira.uka.de/bibliography/Ai/automated.text.categorization.html

[4] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[5] N. Hicham, S. Karim, and N. Habbat, "Customer sentiment analysis for Arabic social media using a novel ensemble machine learning approach," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 4, pp. 4504–4515, Aug. 2023, doi: 10.11591/ijece.v13i4.pp4504-4515.

[6] N. S. I. P. and M. P. K. Kyritsis, "A Comparative Performance Evaluation of Algorithms for the Analysis and Recognition of Emotional Content', Artificial Intelligence.," *IntechOpen*, Jan. 2024.

[7] B. Pang and L. Lee, "Opinion mining and sentiment analysis," 2008.

[8] B. Liu, "Sentiment Analysis and Opinion Mining," Morgan & Claypool Publishers, 2012.

[9] M. Pontiki, H. Papageorgiou, D. Galanis, I. Androutsopoulos, J. Pavlopoulos, and S. Manandhar, "SemEval-2014 Task 4: Aspect Based Sentiment Analysis," 2014. [Online]. Available: http://alt.qcri.

[10] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis," 2005. [Online]. Available: http://www.cs.pitt.edu/

[11] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-Based Methods for Sentiment Analysis," 2011.

[12] D. Tang, B. Qin, and T. Liu, "Document Modeling with Gated Recurrent Neural Network for Sentiment Classification," Association for Computational Linguistics, 2015. [Online]. Available: http://ir.hit.edu.cn/

[13] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019. [Online]. Available: https://github.com/tensorflow/tensor2tensor

[14] O. Araque, I. Corcuera-Platas, J. F. Sánchez-Rada, and C. A. Iglesias, "Enhancing deep learning sentiment analysis with ensemble techniques in social applications," *Expert Syst Appl*, vol. 77, pp. 236–246, Jul. 2017, doi: 10.1016/j.eswa.2017.02.002.

[15] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," 2002. [Online]. Available: http://reviews.imdb.com/Reviews/

[16] Raghav Aggarwal, "https://www.searchunify.com/sudo-technical-blogs/how-to-measure-the-efficacy-of-your-sentiment-analysis-model/."

[17] "https://www.linkedin.com/advice/1/how-can-you-evaluate-sentiment-analysis-model-ygfec."

[18] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf Process Manag*, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/J.IPM.2009.03.002.

[19] "Evaluation_From_Precision_Recall_and_F-Factor_to_R (2)".

[20] "Know The Best Evaluation Metrics for Your Regression Model !"

[21] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature," *Geosci Model Dev*, vol. 7, no. 3, pp. 1247–1250, Jun. 2014, doi: 10.5194/gmd-7-1247-2014.

[22] W. Wang and Y. Lu, "Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model," in *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing, Apr. 2018. doi: 10.1088/1757-899X/324/1/012049.

[23] A. C. Leon, "Descriptive and Inferential Statistics," *Comprehensive Clinical Psychology*, pp. 243–285, 1998, doi: 10.1016/B0080-4270(73)00264-9.

[24] M. G. KENDALL, "A NEW MEASURE OF RANK CORRELATION, ," *Biometrika*, vol. 30, no. 1–2, pp. 81–93, Jun. 1938.

[25] S. M. Lundberg, P. G. Allen, and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions." [Online]. Available: https://github.com/slundberg/shap

[26] M. H. and A. N. S. Barocas, *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2021.