Asian Journal of Research in Computer Science

Volume 17, Issue 7, Page 95-102, 2024; Article no.AJRCOS.118629
ISSN: 2581-8260

Corpus-based Approaches for Sentiment Analysis: A Review

Mahmood Umar [a*], Hauwa Ibrahim Binji [a] and Anas Tukur Balarabe [a]

[a] *Department of Computer Science, Sokoto State University, Sokoto, Nigeria.*

Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Article Information

DOI: https://doi.org/10.9734/ajrcos/2024/v17i7481

Open Peer Review History:
This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here: https://www.sdiarticle5.com/review-history/*118629*

Review Article

Received: 05/04/2024
Accepted: 06/06/2024
Published: 03/07/2024

ABSTRACT

The investigation studied the state of art on corpus-based approaches for sentiment analysis. Thus, detailing its methodologies, evaluation metrics, limitations, and future directions. The importance of sentiment analysis in fields such as marketing, customer feedback analysis, social media monitoring, financial analysis, and political science is emphasized. The methodology for corpus-based approaches in sentiment analysis includes the following key steps: data collection, preprocessing, feature extraction, and sentiment classification. The lexicon-based approaches include the corpus-based or bag of words (BOW) and dictionary (also called opinion lexicon). Evaluation of the corpus-based sentiment analysis approach is addressed through performance metrics such as accuracy, precision, recall, F1-score, and comparative analysis with other approaches including hybrid and rule-based systems. Limitations of corpus-based sentiment analysis, such as data sparsity and domain adaptation, are acknowledged, alongside potential enhancements and research directions including ensemble learning, deep learning architectures, and multimodal data integration. The conclusion emphasizes the versatility and scalability of corpus-based sentiment analysis, while ongoing research efforts aim to address its limitations and further enhance its applicability in diverse domains.

**Corresponding author: E-mail: mahmoodumar24@gmail.com;*

***Cite as:** Umar, Mahmood, Hauwa Ibrahim Binji, and Anas Tukur Balarabe. 2024. "Corpus-Based Approaches for Sentiment Analysis: A Review". Asian Journal of Research in Computer Science 17 (7):95-102. https://doi.org/10.9734/ajrcos/2024/v17i7481.*

Keywords: Corpus-based approach; sentiment analysis; NLP; feature extraction.

1. INTRODUCTION

The definition of sentiment analysis is clarified as the automatic extraction and classification of sentiments from textual data, categorizing them as positive, negative, or neutral using natural language processing (NLP) and machine learning techniques.

Sentiment analysis, also known as opinion mining, is a computational technique under Natural Language Processing (NLP) used to determine the sentiment, emotion or opinion expressed in natural language text [1,2]. It plays a crucial role in various fields such as marketing, customer feedback analysis, social media monitoring, and political analysis [3]. Sentiment Analysis is one of the most critical research areas in Natural language processing (NLP) (Badawi, Kazemi, Rezaie, & KurdiSent, 2024).

Summarily, it involves the automatic extraction and classification of sentiments from textual data, categorizing them as positive, negative, or neutral [4]. In essence, it utilizes natural language processing (NLP) and machine learning techniques to analyze the subjective information present in the text.

Sentiment classification is composed of machine learning approaches, lexicon-based approaches and hybrid-based approaches [5]. The lexicon-based approaches include the corpus-based or bag of words (BOW) and dictionary also called opinion lexicon [6].

The significance of sentiment analysis in diverse fields will be underscored, elucidating its role in extracting valuable insights from textual data, understanding public sentiment, and facilitating informed decision-making. This paper provides an overview of the corpus-based approach for sentiment analysis, focusing on its methodologies, evaluation, limitations, and future directions.

1.1 Role of Sentiment Analysis in Various Fields

Sentiment analysis, also known as opinion mining, is a valuable tool in data analysis across various domains. Its application extends to a wide range of fields, including marketing, customer service, social media monitoring, financial analysis, and political science [7]. Here are some specific applications of sentiment analysis in data analysis as highlighted by Priya Jadhav;Aditya Saha [8].

i. **Customer Feedback Analysis**: Sentiment analysis is widely used to analyze customer feedback, such as product reviews, surveys, and social media comments. By automatically categorizing sentiments expressed in customer feedback, businesses can identify areas for improvement, monitor brand perception, and tailor their products or services to meet customer needs effectively.
ii. **Market Research**: In market research, sentiment analysis helps companies gauge public opinion about their products, services, or brand reputation. By analyzing sentiments expressed in online discussions, forums, and social media platforms, businesses can identify emerging trends, assess market sentiment, and make data-driven decisions regarding product development, marketing strategies, and customer engagement.
iii. **Brand Monitoring and Reputation Management**: Sentiment analysis enables businesses to monitor their brand's online reputation by analyzing sentiments expressed in online mentions, news articles, and social media conversations. By tracking positive and negative sentiments associated with their brand, businesses can address customer concerns promptly, mitigate potential PR crises, and maintain a positive brand image.
iv. **Customer Service and Support**: Sentiment analysis is utilized in customer service to analyse customer inquiries, complaints, and feedback received through various channels, such as emails, live chats, and social media messages. By automatically categorizing the sentiment of customer interactions, businesses can prioritize and address issues more efficiently, improve response times, and enhance overall customer satisfaction.
v. **Financial Analysis**: In finance, sentiment analysis is employed to analyse sentiments expressed in financial news, social media discussions, and analyst reports to gauge market sentiment and predict stock price movements. By

monitoring investor sentiment and market sentiment indicators, financial analysts and traders can make more informed investment decisions and identify potential market trends or anomalies.

vi **Political Analysis**: Sentiment analysis is used in political analysis to analyze public sentiment towards political candidates, parties, policies, and current events. By analysing sentiments expressed in social media discussions, news articles, and public opinion polls, political analysts can assess voter sentiment, track electoral trends, and predict election outcomes.

vii **Product and Service Evaluation**: Sentiment analysis is utilized to evaluate the performance of products and services by analysing sentiments expressed in customer reviews, ratings, and feedback. By aggregating and analysing sentiment data, businesses can identify the strengths and weaknesses of their offerings, benchmark against competitors, and prioritize areas for improvement.

viii **Event Monitoring and Crisis Management**: Sentiment analysis is employed to monitor public sentiment during events, crises, or public relations campaigns. By analysing sentiments expressed in real-time social media conversations, news articles, and online discussions, organizations can assess public perception, identify potential issues or crises, and take proactive measures to address them.

2. REVIEW OF SOME RELATED WORKS

2.1 Overview of Corpus-Based Approach for Sentiment Analysis

The corpus-based approach relies on large collections of annotated text data, known as corpora, to train machine learning models for sentiment analysis [2]. It involves several key steps, including data collection, preprocessing, feature extraction, and sentiment classification.

2.2 Data Collection and Preprocessing

Data collection involves gathering text data from various sources such as social media, product reviews, and news articles. Preprocessing steps include tokenization, lowercasing, removal of stop words, punctuation, and special characters, as well as stemming or lemmatization to normalize the text [7].

2.3 Feature Extraction and Selection

Feature extraction involves transforming the preprocessed text data into numerical representations, such as bag-of-words, TF-IDF (Term Frequency-Inverse Document Frequency), or word embeddings [7]. Feature selection techniques may be employed to reduce dimensionality and improve model performance [3].

2.4 Sentiment Classification Techniques

Several machine learning algorithms, including but not limited to, Naive Bayes, Support Vector Machines (SVM) [4] Logistic Regression and Neural Networks [9], are commonly used for sentiment classification. These algorithms learn to classify text data into positive, negative, or neutral sentiment categories based on the extracted features.

3. EVALUATION OF CORPUS-BASED APPROACH

3.1 Performance Metrics for Sentiment Analysis

Performance metrics such as accuracy, precision, recall, F1-measure, confusion matrix, k folds and cross-validation and so on are used to evaluate the effectiveness of sentiment analysis models [10,11]. These metrics measure the model's ability to correctly classify sentiments and handle class imbalances.

3.2 Comparison with Other Approaches

Corpus-based approach is compared with machine learning approaches, dictionary-based approaches, rule-based approaches, and hybrid methods to assess their relative performance in sentiment analysis tasks. Corpus-based algorithms often outperform simpler approaches, especially in scenarios with complex or domain-specific language [1].

In sentiment analysis, various approaches are employed to classify the sentiment expressed in textual data. The comparison between the lexicon-based approaches (corpus-based approach in particular), machine learning approaches, rule-based systems, and hybrid methods allows us to assess their relative performance in sentiment analysis tasks, are shown in Table 1:

Table 1. Comparison between lexicon-based approaches and others

Approach	Methodology	Strengths	Weaknesses
Machine Learning Approach	• Relies on large collections of annotated text data (corpora) to train ML models for sentiment analysis [2].	• Can handle complex language patterns and adapt to different domains [2]. • Effective in capturing context and nuances in sentiment expression.	• Requires large amounts of data for training. • May suffer from data sparsity issues in specific domains [9]. • Performance heavily depends on the quality of the training corpus.
Lexicon-Based Approach	• Utilizes sentiment lexicons or dictionaries containing predefined lists of words associated with positive or negative sentiments [12] • It's classified as corpus-based and dictionary-based approaches by Raghunathan and. Kandasamy [10] Kumar, Roy, Dogra, & Kim [13] • Corpus-based approaches employs two techniques: Semantic and Statistical [10].	• Simple and interpretable. • Can handle out-of-vocabulary words. • Less computationally intensive compared to machine learning approach.	• Limited coverage of sentiment lexicons may lead to inaccuracies, especially with domain [2] • Specific or ambiguous terms. • May struggle with context-dependent sentiment expressions.
Rule-Based Systems	• Defines explicit rules or heuristics based on linguistic patterns, syntactic structures, or semantic rules. • It does sentiment analysis on the basis a set of human-created rules to identify subject, polarity, or the opinion [7]	• Allows for fine-grained control over sentiment analysis process. • Can incorporate domain-specific knowledge and linguistic rules.	• Limited scalability and generalization. • Requires manual effort to design and maintain rules. • May struggle with capturing nuances and variations in sentiment expression. • not very efficient as it will not analyse how words are combined in a sequence [7]
Hybrid Methods	• Combines multiple approaches (machine learning, lexicon-based, rule-based) to leverage their strengths and mitigate their weaknesses [7]	• Offers improved robustness and performance by combining complementary techniques. • Can adapt to diverse datasets and domains.	• Increased complexity in implementation and tuning. • Requires careful integration and coordination of different components.

Table 2. Survey on the experimental results on lexicon-based approaches for sentiment analysis

Studies	Level	Feature Extraction	Source of Data	Algorithm	Tool	Evaluation	
						Metrics	**Results**
Kandukuri & Gopal [5]	Aspect level	Tf-IDF	Email	Lexicon-based	R-Tidy text package	Positive	>100%
						Negative	-50%
Vargas-Sierra & Orts [14]	Sentence level	Polarity and intensity	Newspapers; English (Economist) and Spanish (Expansion)	Lexicon-based	Lingmotif 2 software	**IN COVID**	
						TSI-English	76%
						TSI-Spanish	81%
						TSS-English	40%
						TSS-Spanish	49%
Alves & Bekava [15]	Sentence Level	lang2vec a	News and Wikepedia	corpus-based quantitative methods: lang2vec clustering and Marsagram linear properties clustering	NLP tools:lang2vec and Marsagram tool	Language Cluster comparison:	N/A
Munnes, Harsch, Knobloch, & Vogel, [16]	Document Level	Word count,	Dictionaries: SentiWS, Rauh's German Political Sentiment Dictionary, GerVADER, GloVe Dictinary	Dictinary-based approaches: SentiWS, Rauh's German Political Sentiment Dictionary, GerVADER, GloVe algorithm, wordscores and wordfish	R package	**Maximal Correlations**	
						SentiWS	0.29
						Rauh	0.37
						GerVADER	0.32
(Heidarypur, Pahlavannezhad, & Kahani, [17]	Aspect level	TFIDF	News data from Website	Sentiment Dictionary approaches; Keyword baseline (M1) and PMI (M2)	Python Classifiers	**Accuracy**	
						Keyword baseline (M1)	37.01%
						PMI (M2)	42.01%

Each approach in Table 1 has its own strengths and weaknesses in sentiment analysis tasks. Corpus-based approach excels in capturing context and nuances but requires substantial labeled data. Lexicon-based approaches are simple and interpretable but may lack coverage and struggle with context. Rule-based systems offer fine-grained control but are limited in scalability and generalization. Hybrid methods aim to leverage the strengths of multiple approaches but require careful integration and tuning. The choice of approach depends on factors such as the availability of labeled data, the complexity of the sentiment analysis task, and the desired balance between accuracy and interpretability.

3.3 Case Studies and Experimental Results

In this section, case studies and experimental results on corpus-based approaches in real-world sentiment analysis tasks will be demonstrated and summarized in tabular form. The table showcases the performance of lexicon-based approaches especially the corpus-based approaches across different domains and datasets, highlighting their versatility and scalability.

A review of studies (in Table 2) shows that R and Python programming are the commonly used tools for sentiment classification. The review also shows that scholars design their own dictionaries for sentiment classification.

The following hint for easy understanding of the abbreviations are shown in Table 2:

Hint: (Text Sentiment Score, or TSS) and Text Sentiment Intensity (TSI)

4. LIMITATIONS AND FUTURE DIRECTIONS

4.1 Challenges in Corpus-Based Sentiment Analysis

According to Siyu Lei1 and Chu-Ren Huang, [18] challenges of sentiment analysis include the difficulty of dealing with non-standard language and semantic ambiguity. Technical problems are also posing a significant challenge to sentiment analysis, such as the lack of training data that can be used across domains and, more in general, the lack of datasets for specialized domains.

Challenges in corpus-based sentiment analysis include data sparsity, domain adaptation, context sensitivity, and the need for large annotated corpora [10]. Addressing these challenges requires advanced techniques in machine learning, NLP, and domain-specific knowledge incorporation.

4.2 Recommendations and Future Works

Future research directions include exploring ensemble learning techniques, deep learning architectures, incorporating multimodal data, merging lexical and machine learning model for improved sentiment analysis [19]. Additionally, efforts towards developing domain-specific sentiment lexicons and annotated corpora can further enhance the performance of corpus-based algorithms [20].

5. CONCLUSION

Corpus-based approach for sentiment analysis offers a powerful and versatile approach for extracting sentiment from textual data. By leveraging large annotated corpora and machine learning techniques, these algorithms can provide valuable insights into public opinion, customer sentiment, and market trends. Despite their limitations, there are ongoing research efforts that enhance the performance and applicability of corpus-based approaches for sentiment analysis in diverse domains.

DISCLAIMER (ARTIFICIAL INTELLIGENCE)

Author(s) hereby declare that generative AI technologies such as Large Language Models, etc have been used during writing or editing of manuscripts. This explanation will include the name, version, model, and source of the generative AI technology and as well as all input prompts provided to the generative AI technology

Details of the AI usage are given below:

1. ChatGPT
2. version 3.5
3. source: Google

The inputs used are: Sentiment analysis, sentiment analysis techniques, review on sentiment analysis, corpus-based techniques for sentiment analysis. Role of sentiment analysis, lexicon-based approaches for sentiment analysis.

ACKNOWLEDGEMENT

I wish to acknowledge the effort of Tertiary Education Trust Fund (TETFund) for granting me a research grant for this project with reference number: TETF/DR&D/CE/UNI/SOKOTO/IBR/2023/VOL.1.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Darwich, Mohammad & Mohd Noah, Shahrul Azman Omar, Nazlia Osman, Nurul. Corpus-Based Techniques for Sentiment Lexicon Generation: A Review. Journal of Digital Information Management. 2019;17:296-305.
2. Darwich M. Corpus-Based Techniques for Sentiment Lexicon Generation: A Review. Journal of Digital Information Management. 2019;10-34.
3. Yasen, Mais Tedmori, Sara. Movies Reviews Sentiment Analysis and Classification. 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT). Amman, Jordan: IEEE Xplore; 2019.
4. Cyril CPD, Beulah JR, Subramani N, Mohan P, Harshavardhan A, Sivabalaselvamani D. An automated learning model for sentiment analysis and data classification of Twitter data using balanced CA-SVM. Concurrent Engineering. 2021;29(4):1-10.
5. Kandukuri M, Gopal VH. Sentiment Analysis on Email Database Corpus-based Approach. Singapore Journal of Scientific Research. 2019;9(2):45-51.
6. Umar M, Ahmad NB, Zainal A. Sentiment Analysis of Student’s Opinion on Programming Assessment: Evaluation of Naïve Bayes over Support Vector Machine. International Journal of Innovative Computing. 2020;51-58.
7. Ms. Binju Saju; Ms. Siji Jose; Mr. Amal Antony. Comprehensive Study on Sentiment Analysis: Types, Approaches, Recent; 2020.
8. Priya Jadhav;Aditya Saha. Sentiment Analysis of Mobile App Reviews Using Robotic Process automation. 2023 7th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech). 2023;1-6.
9. Haisal Dauda Abubakar; Sharin Hazlin Huspi; Mahmood Umar. A Scheme of Pairwise Feature Combinations to Improve Sentiment Classification Using Book Review Dataset. International Journal of Innovative Computing. 2022;25-33.
10. Raghunathan N, Kandasamy S. Challenges and Issues in Sentiment Analysis: A Comprehensive Survey; 2023. IEEE ACCESS, 69626-69642.
11. Bordoloi M, Biswas SK. Sentiment analysis: A survey on design framework, applications and future scopes. Springer. 2023;12505–12560.
12. Umar M, Aliyu M, Modi AS. Sentiment Analysis in the Era of Web 2.0: Applications, Implementation Tools and Approaches for the Novice Researcher. Caliphate Journal of Science & Technology (CaJoST). 2022;1-9.
13. Kumar S, Roy PP, Dogra DP, Kim YG. A Comprehensive Review on Sentiment Analysis: Tasks, Approaches and Applications; 2023. Retrieved June 2024, from arXiv:2311. 11250 [cs.AI]: Available:https://arxiv.org/abs/2311.11250
14. Vargas-Sierra C, Orts MÁ. Sentiment and emotion in financial journalism: a corpus-based, cross-linguistic analysis of the effects of COVID 19. Humanities and Social Sciences Communications. 2023;1-17.
15. Alves D, Bekavac B. Analysis of Corpus-based Word-Order Typological Methods. Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023) Association for Computational Linguistics. 2023; 36-46.
16. Munnes S, Harsch C, Knobloch M, Vogel JS. Examining Sentiment in Complex texts: a Comparison of different Computational Approaches. Frontiers of Big data. 2022;1-16.
17. Heidarypur M, Pahlavannezhad MR, Kahani M. The role of corpus linguistics in sentiment analysis of Persian texts, case study: a Farsi news agency website. Journal of linguistics, philology and translation. 2023;106-121.

18. Siyu Lei1; Chu Ren Huang. Conducting Sentiment Analysis: Lei L. & Liu D. Elements in Corpus Linguistics, CUP. Springer. 2022;29(1).
19. Alnawas A, ARICI N. The Corpus Based Approach to Sentiment Analysis in Modern Standard Arabic and Arabic Dialects: A Literature Review. Journal of Polytechnic. 2018;461-470.
20. Islam T, Sheakh MA, Sadik MR, Mst, & Tahosin S. Lexicon and Deep Learning-Based Approaches in Sentiment Analysis on Short Texts. Journal of Computer and Communications. 2024; 11-34.

Peer-review history:
The peer review history for this paper can be accessed here:
https://www.sdiarticle5.com/review-history/118629