Cureus Journal of Computer Science

Open Access Original Article

Impact of Textual Length on Lexicon-Based Sentiment Classification for Hindi Reviews

Dhanashree S. Kulkarni [1] , Anand B. Deshpande [2] , Sonam A. Bhandurge [1] , Vania V. Estrela [3]

1. Computer Science and Engineering, Angadi Institute of Technology and Management, Belagavi, IND 2. Electronics and Telecommunication, Angadi Institute of Technology and Management, Belagavi, IND 3. Telecommunications, Universidade Federal Fluminense (UFF) Duque de Caxias, Rio de Janeiro, BRA

Corresponding author: Dhanashree S. Kulkarni, dskulkarni10@gmail.com

Received 05/11/2025
Review began 05/29/2025
Review ended 07/30/2025
Published 08/18/2025

DOI:
https://doi.org/10.7759/s44389-025-05622-w

Abstract

The performance of sentiment analysis is affected by different properties like dataset size or training corpus size, feature extraction, length of target documents, preprocessing constraints, length of the review/word count, readability, purity of data, subjectivity of data, etc. Efforts need to be taken to not only increase the performance of the algorithms but also to investigate the impact of some of the linguistic properties of the dataset on the performance. One such data property is word count. The lexicon model was experimented for effect of word count and was tested on both the single and multi-domain datasets to find the effect of word count on accuracy and other parameters by splitting into four categories, where each category or group had a different word count. The experimentation results indicated that the accuracy was significantly high for sentences that had shorter words. After a particular threshold, the accuracy starts decreasing as the number of words in the sentence goes on increasing. The statistical significance of the impact of word count on the classification accuracy clearly indicates that the impact of the review words count on the classification accuracy is statistically significant at a 99% confidence level. Less than 80 words is an ideal level for the review words count.

Categories: Algorithm Analysis, Computational Linguistics , Natural Language Processing (NLP)
Keywords: natural language processing, textual analysis, sentiment analysis, lexicon method, wordcount, artificial intelligence, message length, accuracy, sentiment classification, precision

Introduction

The progression of the Internet has given rise to blogs, forums, and social networking platforms that produce a large amount of user information on a regular basis [1]. Mining this user-generated data to discover users' likings, interests, opinions, and sentiments has become a fascinating area for researchers. However, extracting relevant opinions from web-generated information is a challenging task [2,3]. The process of mining information from different text forms such as blogs, reviews, tweets, etc., and classifying them into categories like positive, negative, and neutral based on polarity is known as Sentiment Analysis (SA) [4]. SA is an emerging research area within natural language processing that focuses on analyzing and classifying feelings and emotions expressed in text. It is sometimes called emotion mining and incorporates techniques from text analytics and computational linguistics to extract polarity and subjective information (Source: Wikipedia). Sentiments mainly relate to emotions about something that reflect what a person feels. The introduction of Unicode standards in recent years has enabled the development of webpages in Indian languages, especially Hindi. Hindi ranks as the fourth most spoken language in the world [5,6], and the number of users and contributors on social media in Hindi is rapidly increasing. There is a need to effectively mine and analyze this Hindi data, which would be especially beneficial to government organizations and product-based companies. The performance of SA depends on various factors such as dataset size, training corpus size, feature extraction, length of target documents, preprocessing constraints, review length or word count, readability, data purity, subjectivity, and more [7]. Efforts are needed not only to improve algorithm performance but also to investigate the impact of linguistic properties of datasets on results, with word count being one such important property [7,8]. Most existing research focuses on comparing the accuracy of different algorithms but does not consider the role of various data properties and settings, which may result in alterations in the accuracy. The results and the performances may differ depending on the nature of the data used and the way experimentation was done. SA has been a targeted research area for the past decade, with substantial work done on English and increasing focus on regional languages [9]. Various approaches have been applied to analyze sentiments in Hindi, with the lexicon-based method being one of them. This method uses a sentiment lexicon containing words with assigned polarity scores, indicating whether a word is positive or negative. All the scores of the words that contribute to the sentence are taken and investigated, and the combined score is calculated by summing the scores that then will reveal the polarity of the sentence [10-12]. Figure *1* shows a typical flowchart for the lexicon-based technique. As shown, preprocessing is done on the extracted Hindi sentences, which are performed to achieve dimensionality reduction. Data preprocessing is the first crucial step that deals with preparing the raw data and making it appropriate for the model for which it is being used [13].

How to cite this article
Kulkarni D S, Deshpande A B, Bhandurge S A, et al. (August 18, 2025) Impact of Textual Length on Lexicon-Based Sentiment Classification for Hindi Reviews. Cureus J Comput Sci 2 : es44389-025-05622-w. DOI https://doi.org/10.7759/s44389-025-05622-w

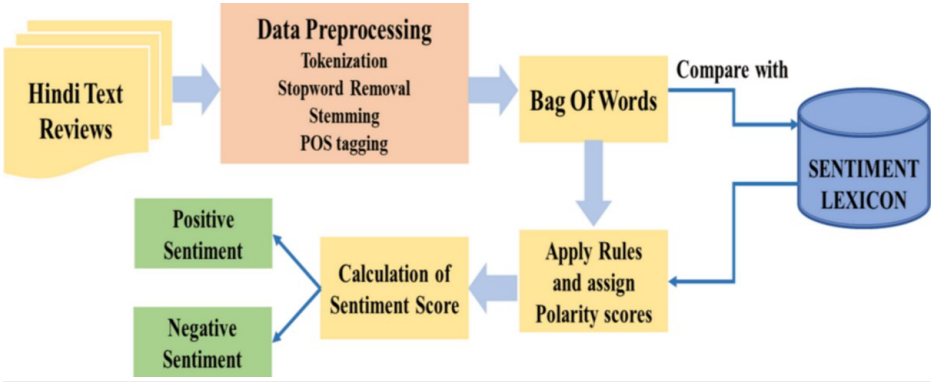

FIGURE 1: Lexicon-based technique

POS, Part of Speech

The various steps involved in data preprocessing are as follows: 1. Tokenization, 2. Stop words removal, 3. Stemming, and 4. Lemmatization. Tokenization deals with splitting the sentence into individual words. Stop words are the most common words used in the language that have no significance as such in the analysis and can be removed [14]. Hence, stop-word removal is a part of preprocessing. Stemming is generally a heuristic method of converting a particular word to its base form. Lemmatization is similar to stemming, but it reduces the word to a specific one that exists in the language. Part of Speech (POS) tagging is also done many times as the next step in preprocessing, which gives each word its corresponding POS, such as adjective, adverb, verb, noun, etc. POS tagging can be very useful when used as a feature extraction method. Therefore, this process is generally used as a part of preprocessing. After preprocessing is performed, the next step is creating a bag of words model, which is nothing but a collection of unigrams. These unigrams, or words in a sentence, are then compared against the sentiment lexicon or dictionary, which returns the respective scores [6]. It has been seen that the length of words in reviews varies widely depending on the domain. Nowadays, well-known review websites or social networking platforms enforce their users to write comments or reviews by providing them with a defined length [15]. Twitter is a well-known online micro-blogging and social networking platform that allows its users to write a message of maximum length 280 characters. However, the exact nature of the relationship between the length of textual data and classifier performance has been unclear [16]. Hence, the main objective is to analyze the effect of the length of words of a review/sentence on the accuracy of SA.

Related work

Pak and Paroubek [17] noted that sentiment classification performance dropped on Twitter data when message lengths were beyond average tweet sizes. According to Go et al. [18], short texts often lack context, making it harder to predict polarity and reducing performance in rule-based systems. Maddu and Sanapala [19] found that longer messages provided lower accuracy for Telugu due to the cumulative effect of morphological and orthographic noise. According to Joshi et al. [20], brief messages in local scripts often bear the burn of code switching, grammatical non-standardization and transliteration issues, which complicates sentiment extraction in low-resource language environments. Li et al. [21] explored how the textual quality of online criticisms affected the classification performance of the analysis of in-depth learning feelings. The analysis of feelings, particularly in the context of Hindi language texts, drew significant attention due to the growing volume of digital content produced in vernacular languages. Previous studies have emphasized the criticality of accurately determining the feeling in revisions, not only to improve the user experience, but also to inform marketing strategies adapted to the regional public. A pertinent aspect that arose from the recent discourse is the impact of the length of the message on the accuracy of feeling, as highlighted in the work of Kulkarni and Rodd [10]. Akhtar et al. [22] demonstrated that the use of a bidirectional long-term memory model significantly surpassed traditional classifiers by processing short and long reviews establishing that moderately sized review messages provided the most accurate sentiment prediction.

Materials And Methods

A simple lexicon algorithm makes use of the Hindi SentiWordNet (HSWN) as the lookup dictionary, and then calculates the polarity of the sentence. HSWN is a well-known lexicon dictionary that was developed by IIT Bombay [23]. The dictionary was built taking into consideration two different resources, the English SentiWordNet and the English Hindi WordNet Linking. The lexicon contains different Hindi words with their positive and negative scores. The words may be of different POS such as adjective, verb, adverb, or a noun. Along with the POS_TAG, positive and negative score, a synset id is also included for the words in the HSWN [24,25]. The steps and the approaches used in the investigation are as shown in Figure *2*.

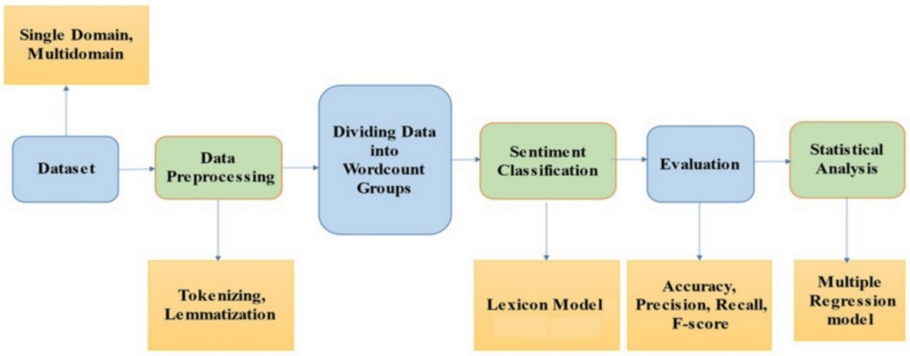

FIGURE 2: Steps and approaches used for investigation of impact of word count

In order to investigate the sensitivity or the impact of word count on the sentiment classification accuracy, the sentences in the dataset are preprocessed and then are split into several categories, and each category or group has a different word count. The four groups of word count that are formed for the investigation are sentences having word count 1-50, sentences having word count 51-80, sentences having word count 81-120, and sentences having word count >120 The next step is to perform the sentiment classification using the lexicon model and then evaluate the results considering the main parameter accuracy as well as other parameters such as recall, precision, and F-score. To verify whether the investigation being done is statistically significant, a statistical analysis is done using the multiple regression model. The lexicon model is tested on both the single and multi-domain datasets to find the effect of word count on accuracy and other parameters by considering the four word-count categories.

Results

This dataset includes 1,000 manually curated movie reviews, sourced from the Hindi news site Jagran.com and an existing Hindi movie review corpus from IIT Bombay. The dataset is balanced, containing 500 positive and 500 negative reviews. To assess the generalizability of the findings, we extended the analysis to a multi-domain dataset containing 4,000 reviews, which included movie reviews, product reviews from IIT Patna, and tourism-related reviews from IIT Bombay.

The goal was to explore how message length influences sentiment classification across domains. The data was split into training (70%), validation (15%), and testing (15%) sets using stratified sampling to preserve sentiment distribution across sets.

Implementation was carried out on the single-domain movie review dataset, and the results in terms of accuracy, precision, recall, and F-score were obtained, as shown in Figure *3*. Upon analyzing the results, it was observed that scores for some words, though present in HSWN, could not be extracted due to morphological variations. The lexicon algorithm was also tested on a multidomain review dataset, with the corresponding results shown in Figure *4*. The accuracy obtained was around 61%. The simple lexicon approach provided good results on the product domain dataset, which contributed to the overall better performance on the multidomain dataset.

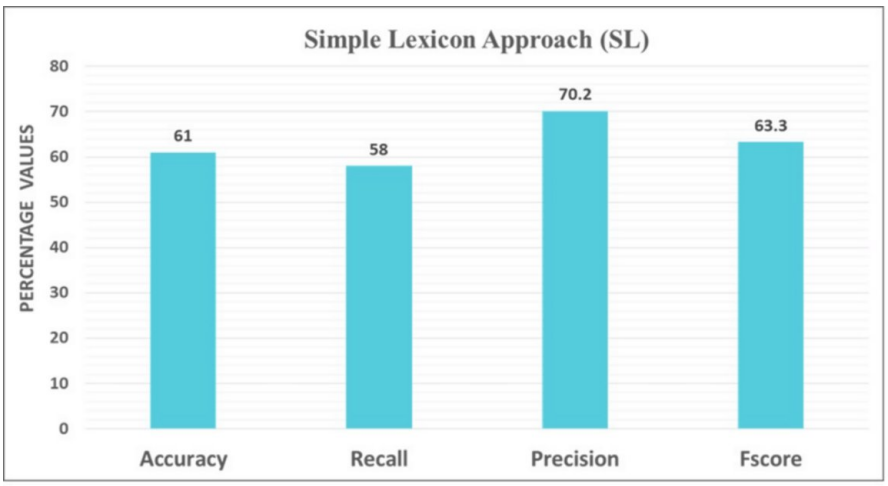

FIGURE 3: Output of using simple lexicon HSWN for single domain dataset

HSWN, Hindi SentiWordNet

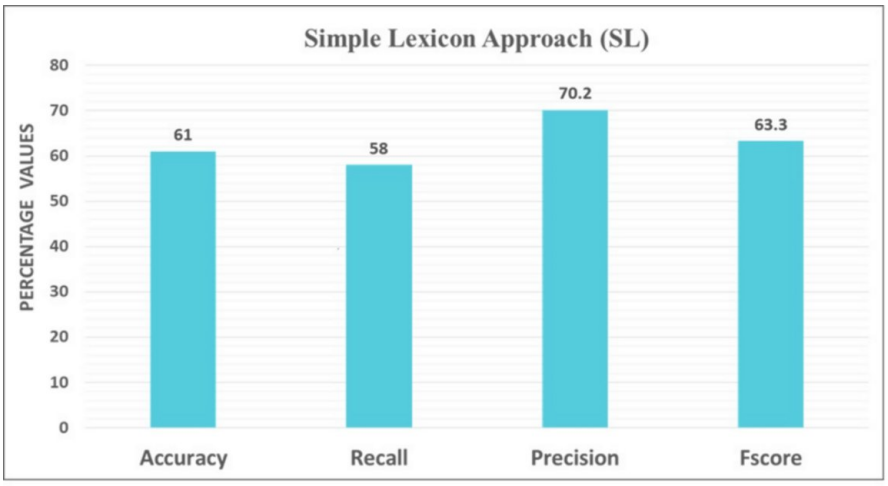

FIGURE 4: Output of using simple lexicon HSWN for multidomain dataset

HSWN, Hindi SentiWordNet

The lexicon model was tested on both single- and multi-domain datasets to evaluate the effect of word count on accuracy and other parameters by considering four word-count categories. Figure *5* shows the impact of word count on various parameters for a single-domain dataset of 1,200 sentences. The graph clearly shows that for the group with 1-50 words, the accuracy is 66.78%. As the number of words increases in the next word-count groups, accuracy slightly decreases. For sentences with more than 120 words, accuracy decreases by almost 4%, indicating that as the number of words increases, accuracy tends to decline. This may occur because reviews with fewer words tend to be more specific about the sentiment. As word count increases, noise can affect the analysis, leading to a significant decrease in accuracy. The effect of word count was also examined for other parameters. Figure *5* shows the impact on recall, precision, and F-score. The effect on recall is same as accuracy - recall decreases as word count increases. There is a slight increase in precision with higher word counts; however, since recall decreases significantly, the F-score also declines substantially. There was a significant decrease in F-score as the word count increased.

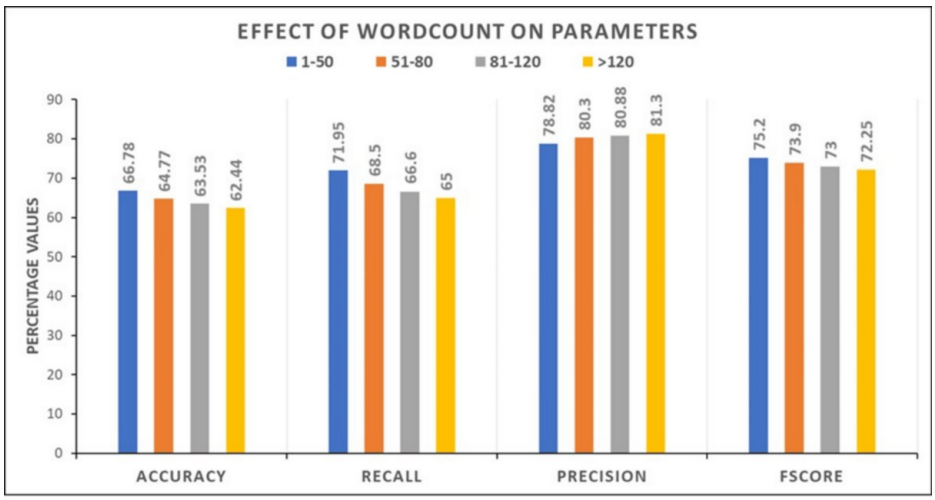

FIGURE 5: Effect of word count on accuracy, recall, precision, and F-score of lexicon model for single domain dataset

The effect of word count on various parameters of the lexicon model was also tested using a multidomain dataset containing 4,000 sentences across various word-count groups, as shown in Figure 6. A drop in accuracy by 4% was observed for the word-count group of 81-120. For sentences with more than 120 words, the accuracy was almost 10% lower compared to the first word-count group of 1-50 words. The impact of word count was also examined for three other parameters: precision, recall, and F-score, and the results have been shown. As the word count increased, recall decreased. There was no significant effect on precision as word count increased. However, due to the decrease in recall, an impact was also observed on the F-score.

A qualitative error analysis was conducted focusing on longer reviews where the model underperformed. Mixed sentiments within a single review often led to misclassification, as the lexicon-based model struggled to resolve polarity conflicts. Word sense ambiguity and negation constructs in longer texts posed additional challenges. Some errors were attributed to limitations in HSWN, where domain-specific or colloquial terms were not covered. A hybrid approach combining lexicon-based and machine learning methods could be explored and tested for handling longer reviews.

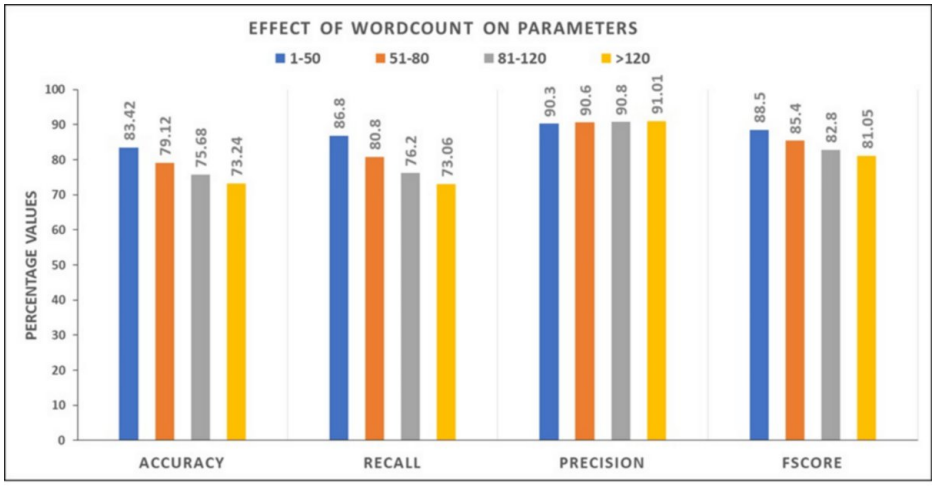

FIGURE 6: Effect of word count on accuracy, recall, precision, and F-score of lexicon model for multi-domain dataset

Discussion

Statistical significance, as the term implies, refers to the claim that an outcome observed from experimental data is unlikely to have occurred by chance and is instead likely due to a specific cause [26].

To verify the statistical significance of the impact of word count on classification accuracy, a multiple regression model was built.

In regression analysis, the null hypothesis typically assumes that each independent variable has no effect on the dependent variable [27]. The coefficient of each predictor quantifies both the magnitude and direction of its effect. A positive coefficient indicates that an increase in the predictor is associated with an increase in the dependent variable, whereas a negative coefficient implies a decrease. When multiple predictors are used, each coefficient estimates the effect of that variable while holding all others constant. The size of the coefficient for every independent variable gives the size of the effect that it is having on the dependent variable, and the sign on the coefficient whether it is positive or negative gives the direction of the effect. Hence, in regression with multiple independent variables, the coefficient tells you how much the dependent variable is expected to increase when that independent variable increases by 1, holding all the other independent variables constant [28]. Regression coefficients serve as a powerful tool for understanding relationships between variables. The null hypothesis framework helps establish whether these relationships are statistically significant, while the concept of "holding other variables constant" ensures the interpretation of each predictor is meaningful and not confounded by other factors in the model. This is especially valuable in multi-factor studies like ours, where multiple linguistic and structural variables influence sentiment classification accuracy.

The regression model is described as

$$\text{Accuracy}_i = Q_0 + Q_1\ \text{Word\_count}_i + Q_2\ \text{LexModel}_i + Q_3\ \text{Data\_size}_i$$

where

Accuracy is the dependent variable denoting the accuracy of the sentiment classification.

Word_count_i is a binary variable indicating whether the review is short or long.

LexModel_i controls for the lexicon-based sentiment analysis model used.

Data_size_i accounts for dataset size (assigned 1 when dataset size is 4,000).

Q_0 to Q_3 are the regression coefficients estimated during model fitting.

For the independent variables, a variable Word_count is introduced that denotes the level of the words count. Considering the experimentation, the short-review level is defined as having a threshold count of 80 words in each review and the long-review level having beyond 80 words. With reference to the short review level, the relative impact of the word count on the accuracy is examined. LexModel is used to control the effect of lexicon-based model. Similarly, Data_size is used to control the effect of the dataset size and are taken as 1 when the dataset size is 4,000. The regression analysis is made by using Stata software.

Table *1* shows the results of the statistical significance. From the results, it is clear that the P value for the estimate of Word_count is less than 0.0001, which implies that the impact of the word count on the sentiment classification accuracy is statistically significant at a 99% confidence level. Also, the estimate for Word_count is negative, which means that a higher word count could lower the accuracy of sentiment classification. This highlights the importance of expressing sentiments clearly using a limited number of words. Longer reviews tend to introduce noise, which negatively impacts accuracy. For Hindi sentiment analysis using lexicon-based methods, keeping the word count under 80 is ideal, as it provides sufficient information to express the reviewer’s sentiment with minimal noise. The results clearly indicate that the impact of review word count on classification accuracy is statistically significant at the 99% confidence level. Thus, less than 80 words is the recommended word count for reviews.

Parameter	Estimate	Standard Error	T-statistic	P
W-count	-2.46	0.319	-8.44	<0.0001
LexModel	12.67	0.825	15.35	<0.0001
Dsize	3.44	0.45	7.52	<0.0001

TABLE 1: Results of the statistical significance

Conclusions

This paper explores the impact of word count - a key linguistic property - within the context of low-resource languages like Hindi by investigating the lexicon-based sentiment classification approach to both single-domain and multi-domain Hindi datasets. Reviews were categorized into four distinct word-count groups to examine their effect on classification accuracy. Preliminary findings suggest that shorter reviews specifically those with fewer than 80 words tended to yield relatively higher accuracy within the lexicon-based framework. Beyond this threshold, accuracy appeared to decline, potentially due to increased textual noise in longer reviews. A multiple regression analysis further confirmed that the observed relationship between word count and classification accuracy is statistically significant at the 99% confidence level. Less than 80 words is considered to be an appropriate level for the review words count.

The results provide an indication that message length can influence sentiment classification performance in lexicon-based systems. Future research will aim to benchmark lexicon-based methods against machine learning and deep learning models to better understand how word count affects classification across architectures. Expanding the analysis to larger and more varied datasets, and exploring hybrid approaches, could further improve the robustness and applicability of findings. Additionally, future efforts may consider developing models that incorporate message length-aware objectives, utilize context-sensitive truncation strategies, and adopt adaptive tokenization schemes that adjust input length dynamically.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Dhanashree S. Kulkarni, Anand B. Deshpande

Acquisition, analysis, or interpretation of data: Dhanashree S. Kulkarni, Anand B. Deshpande, Sonam A. Bhandurge, Vania V. Estrela

Drafting of the manuscript: Dhanashree S. Kulkarni, Sonam A. Bhandurge

Critical review of the manuscript for important intellectual content: Dhanashree S. Kulkarni, Anand B. Deshpande, Sonam A. Bhandurge, Vania V. Estrela

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

1. Al Katat S, Zaki C, Hazimeh H, El Bitar I, Angarita R, Trojman L: Natural language processing for Arabic sentiment analysis: A systematic literature review. IEEE Transactions on Big Data. 2024, 10:576-594. 10.1109/tbdata.2024.3366083
2. Kulkarni DS, Rodd SF: Extensive study of text based methods for opinion mining. 2018 2nd International Conference on Inventive Systems and Control (ICISC), Coimbatore, India. 2018, 523-527. 10.1109/ICISC.2018.8399127.
3. Rathan M, Hulipalled VR, Venugopal KR, Patnaik LM: Consumer insight mining: Aspect based Twitter opinion

mining of mobile phone reviews. Applied Soft Computing. 2018, 68:765-773. 10.1016/j.asoc.2017.07.056
4. Ali HMU, Farooq Q, Imran A, El Hindi K: A systematic literature review on sentiment analysis techniques, challenges, and future trends. Knowledge and Information Systems. 2025, 67:3967-4034. 10.1007/s10115-025-02365-x
5. Bhattacharyya P, Murthy H, Ranathunga S, Munasinghe R: Indic language computing. Communications of the ACM. 2019, 62:70-75. 10.1145/3343456
6. Komal G, Buttar PK: Survey on sentiment analysis in Hindi language. International Journal of Advanced Research in Computer Science. 2017, 8:1360-1363.
7. Choi Y, Lee H: Data properties and the performance of sentiment classification for electronic commerce applications. Information Systems Frontiers. 2017, 19:993-1012. 10.1007/s10796-017-9741-7
8. Hossain MS, Babu MA, Yusuf KM: Exploring emotional dynamics: Word count effects in Facebook posts by sports shoe brands. Heliyon. 2024, 10:e39808. 10.1016/j.heliyon.2024.e39808
9. Rani S, Kumar P: A journey of Indian languages over sentiment analysis: A systematic review. Artificial Intelligence Review. 2019, 52:1415-1462. 10.1007/s10462-018-9670-y
10. Kulkarni DS, Rodd SF: Towards enhancement of the lexicon approach for Hindi sentiment analysis. IOT with Smart Systems. 2022, 445-451. 10.1007/978-981-16-3945-6_43
11. Kulkarni DS, Rodd SF: Word sense disambiguation for lexicon-based sentiment analysis in Hindi. Webology. 2022, 19:592-600. 10.14704/web/v19i1/web19042
12. Alsemaree O, Alam AS, Gill SS, Uhlig S: An analysis of customer perception using lexicon-based sentiment analysis of Arabic texts framework. Heliyon. 2024, 10:e30320. 10.1016/j.heliyon.2024.e30320
13. Rakshitha K, Ramalingam HM, Pavithra M, Advi HD, Hegde M: Sentimental analysis of Indian regional languages on social media. Global Transitions Proceedings. 2021, 2:414-420. 10.1016/j.gltp.2021.08.039
14. Jha V, Manjunath N, Shenoy PD, Venugopal KR: Sentiment analysis in a resource scarce language: Hindi. International Journal of Scientific and Engineering Research. 2016, 7:968-980. 10.14299/ijser.2016.09.005
15. Fadlil A, Riadi I, Andrianto F: Improving sentiment analysis in digital marketplaces through SVM kernel fine-tuning. International Journal of Computing and Digital Systems. 2024, 16:159-171.
16. Gkikas DC, Tzafilkou K, Theodoridis PK, Garmpis A, Gkikas MC: How do text characteristics impact user engagement in social media posts: Modeling content readability, length, and hashtags number in Facebook. International Journal of Information Management Data Insights. 2022, 2:100067. 10.1016/j.jjimei.2022.100067
17. Pak A, Paroubek P: Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). 2010, 1320-1326.
18. Go A, Bhayani R, Huang L: Twitter sentiment classification using distant supervision. CS224N Project Report. 2015, 1-6.
19. Maddu S, Sanapala VR: A survey on NLP tasks, resources and techniques for low-resource Telugu-English code-mixed text. ACM Transactions on Asian and Low-Resource Language Information Processing. 2024, 10.1145/3695766
20. Joshi P, Santy S, Budhiraja A, Bali K, Choudhury M: The state and fate of linguistic diversity and inclusion in the NLP world. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020, 6282-6293. 10.18653/v1/2020.acl-main.560
21. Li L, Goh TT, Jin D: How textual quality of online reviews affect classification performance: a case of deep learning sentiment analysis. Neural Computing and Applications. 2020, 32:4387-4415. 10.1007/s00521-018-3865-7
22. Akhtar MS, Sawant P, Sen S, Ekbal A, Bhattacharyya P: Solving data sparsity for aspect based sentiment analysis using cross-linguality and multi-linguality. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2018, 1:572-582. 10.18653/v1/N18-1053
23. Joshi A, Balamurali A, Bhattacharyya P: A fall-back strategy for sentiment analysis in Hindi: A case study. Proceedings of ICON 2010: 8th International Conference on Natural Language Processing. 2010, 1-6.
24. Pandey P, Govilkar S: A framework for sentiment analysis in Hindi using HSWN. International Journal of Computer Applications. 2015, 119:23-26. 10.5120/21176-4185
25. Kulkarni DS, Rodd SS: Sentiment analysis in Hindi—A survey on the state-of-the-art techniques. ACM Transactions on Asian and Low-Resource Language Information Processing. 2022, 21:1-46. 10.1145/3469722
26. Holmberg C: Toward a better understanding of statistical significance and p values in nursing. Nursing Forum. 2024, 2024:7263781. 10.1155/2024/7263781
27. Sufi FK: Identifying the drivers of negative news with sentiment, entity and regression analysis. International Journal of Information Management Data Insights. 2022, 2:100074. 10.1016/j.jjimei.2022.100074
28. Krugmann JO, Hartmann J: Sentiment analysis in the age of generative AI. Customer Needs and Solutions. 2024, 11:3. 10.1007/s40547-024-00143-4