GADL

Journal of Computers, Mechanical and Management

Journal homepage: www.jcmm.co.in

JCMM Journal of Computers, Mechanical and Management

Volume 2 Issue 1

Sentiment Analysis using Latent Dirichlet Allocation for Aspect Term Extraction

Lovish Rajput[*] and Shilpi Gupta

Department of Information Technology, JSS Academy of Technical Education, Noida, Uttar Pradesh, India
201301

Abstract

This work proposes a sentiment analysis approach for decision-making in product design, analysis, and market share. The approach incorporates user-generated text data in the form of consumer reviews to extract product features using topic-based modeling methods. Latent Dirichlet Allocation (LDA) is employed to extract aspect categories from the data and identify the sentiment of each review using the VADER sentiment analyzer. The performance of the proposed method is evaluated in terms of accuracy, with an achieved result of 80%. The extracted topics are also summarized to provide leads for product design and quality assurance. The approach can be used by manufacturers, retailers, and suppliers to understand customers' opinions about their products better and make better decisions. LDA is a powerful unsupervised method that can extract latent topics from a collection of documents; this method has been widely used in text mining, information retrieval, and natural language processing. The accuracy can be improved by using more sophisticated models or more data.

Keywords: Sentiment Analysis; Latent Dirichlet Allocation (LDA); Product Design; User-Generated Text Data; Opinion Mining

1 Introduction

Sentiment Analysis, also known as Opinion Mining, is a rapidly growing research area that aims to extract and analyze the opinions, attitudes, and emotions expressed in unstructured text data, such as consumer reviews, social media posts, and online forums [1–3]. Sentiment Analysis has become an important tool for businesses to better understand consumer perceptions of their products and services and make informed decisions about product design, analysis, and market share [4–6]. One of the most popular techniques used in Sentiment Analysis is Latent Dirichlet Allocation (LDA) [7–9]. LDA is a powerful unsupervised method that can extract latent topics from a collection of documents, providing a way to analyze large amounts of unstructured data [10]. In the context of Sentiment Analysis, LDA has been used to identify and classify aspect categories, such as features or attributes of a product, from consumer reviews [11–13]. By identifying these aspect categories, LDA can provide a more fine-grained understanding of customer sentiment, compared to traditional sentiment analysis that only classifies text as positive, negative, or neutral. Aspect-Based Sentiment Analysis (ABSA) is a specific type of Sentiment Analysis that focuses on identifying the specific features or attributes of a product or service that customers like or dislike [14–16]. ABSA is particularly useful for analyzing consumer reviews and conducting surveys [16]. It provides a more fine-grained understanding of customer sentiment than traditional sentiment analysis, which only classifies text as positive, negative, or neutral. There are several existing methods for ABSA, including rule-based, lexicon-based, and machine learning-based approaches [17, 18]. Rule-based approaches rely on manually defined rules or patterns to identify aspect terms and sentiment [19, 20]. Lexicon-based approaches use pre-existing sentiment lexicons, such as the SentiWordNet, to classify aspect terms and sentiment [21–23].

[*] Corresponding author: lovishrajput@gmail.com

Received: 07 October 2022; **Accepted:** 02 January 2023; **Published:** 28 February 2023

DOI: 10.57159/gadl.jcmm.1.2.22026.

Machine learning-based approaches use supervised or unsupervised techniques to learn patterns from labeled data [17, 24]. In this work, we propose a sentiment analysis approach that utilizes Latent Dirichlet Allocation (LDA) for aspect term extraction and the Valence Aware Dictionary for sEntiment Reasoning (VADER) sentiment analyzer to identify the sentiment of each review. LDA is a powerful unsupervised method widely used in text mining, information retrieval, and natural language processing. VADER is a lexicon-based approach that uses a pre-existing sentiment lexicon and is particularly effective in understanding the sentiment of social media text. The proposed approach aims to provide a more fine-grained understanding of customer sentiment and extract leads for product design and quality assurance by combining the LDA and VADER. The performance of the proposed method will be evaluated in terms of accuracy, and the extracted topics will be summarized to provide leads for product design and quality assurance. This research will be important because it provides a new method for aspect-based sentiment analysis that combines the strengths of both LDA and VADER, and it will be a useful tool for business organizations to better understand consumer perceptions of their products and services and make informed decisions about product design, analysis, and market share. The approach can be used by manufacturers, retailers, and suppliers to understand customers' opinions about their products better and make better decisions.

2 Methodology

2.1 Dataset Description

The study uses user-generated text data from smartphone reviews. The smartphone reviews data consists of 20,000 reviews collected from online marketplaces and e-commerce websites. The reviews were collected for several months and included a variety of different models and brands of smartphones. The reviews include positive and negative opinions and cover various topics such as design, performance, camera quality, battery life, and more. The text reviews were preprocessed in the selected data to remove irrelevant information, such as URLs and special characters, and standardize the text format. The data was then split into training and testing sets to evaluate the performance of the proposed approach. The training set was used to train the LDA model, and the testing set was used to evaluate the model's accuracy. The smartphone reviews data was chosen for this study as they are widely available and represent the types of data businesses often encounter in real-world applications. The smartphone reviews data represents a specific product category. The large sample size of the data provides a good representation of customer opinions and preferences and allows for a robust evaluation of the proposed approach.

2.2 LDA for Aspect Term Extraction in Product Reviews

Latent Dirichlet Allocation (LDA) is a topic modeling technique used to extract latent topics from a collection of documents [25]. LDA uses a probabilistic approach to identify the topics present in a set of documents and the words associated with those topics [26]. The technique is particularly useful for finding reasonably accurate mixtures of topics within a given document. The LDA algorithm is an unsupervised method used to identify latent topics in documents without prior knowledge. The algorithm starts by assigning each document to a random topic and then iteratively updates the topic assignments for each word in the document based on the probability of the word belonging to each topic. The algorithm then updates the topic distributions for each document based on the probability of the document belonging to each topic [27]. Algorithm 1 represents the steps in LDA.

Algorithm 1 LDA Algorithm

```
1:  Initialize the number of topics, $K$, and the number of documents, $D$.
2:  Randomly assign each word in each document to one of the $K$ topics.
3:  For each word in each document:
4:      Calculate the probability of the word belonging to each of the $K$ topics.
5:      Reassign the word to the topic with the highest probability.
6:  For each document:
7:      Calculate the probability of the document belonging to each of the $K$ topics.
8:      Reassign the document to the topic with the highest probability.
9:  Repeat steps 3 and 4 for a specified number of iterations or until the topic assignments converge.
10: Output the resulting topic assignments for each word and document.
```

The LDA algorithm uses a generative model that can discover latent topics in a set of documents, assuming that each document is a mixture of a small number of latent topics. The algorithm uses a probabilistic approach to identify the topics present in a set of documents and the words associated with those topics. The algorithm starts by assigning each document to a random topic and then iteratively updates the topic assignments for each word in the document based on the probability of the word belonging to each topic. The algorithm then updates the topic distributions for each document based on the probability of the document belonging to each topic. It is important to note that LDA requires a pre-specified number of topics, and the number of topics should be chosen based on the complexity of the dataset and the research question. It also requires an appropriate number of iterations or convergence criteria, and the number of iterations should be chosen based on the computational resources available and the desired level of accuracy.

In this study, Latent Dirichlet Allocation (LDA) was used for aspect term extraction from the smartphone reviews data. The dataset was first preprocessed using lower punctuation, stopping word removal, tokenization, and lemmatization. This preprocessing step was necessary to convert the raw text data into a format that the LDA algorithm could understand. Count-Vectorizer, a tool available in the sci-kit-learn library in Python, was used to convert the tokenized text data into a vector representation based on the word frequency for a tokenized review. This vector representation was then used as input for the LDA algorithm. The LDA algorithm could extract latent topics from the dataset based on the frequency of words in the documents. These topics represented the aspect categories that were present in the data. The LDA algorithm identified a product's features or attributes, such as camera quality or battery life, discussed in the reviews. The resulting aspect categories were used to understand the customer opinions about different aspects of the product and make better decisions about product design, analysis, and market share.

2.3 VADER for Sentiment Analysis in Product Reviews

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a powerful sentiment analysis tool specifically tuned to understand the sentiments of social media text. It is a lexicon and rule-based method that can understand and interpret the meaning of text data and classify it as positive, negative, or neutral. VADER employs a combination of techniques to achieve this. One of these techniques is the use of a sentiment lexicon. A sentiment lexicon is a collection of lexical features, such as words pre-classified as positive or negative based on semantic orientation. By using this lexicon, VADER can quickly identify the sentiment of a piece of text. In addition to identifying the overall sentiment of a piece of text, VADER also provides a measure of the degree to which the sentiment is favorable or negative. This means it can determine whether a review is slightly positive or extremely positive. After applying VADER, the aspect categories are summarized based on their sentiments. This helps to understand the overall sentiment of a review and identify which aspects of the product are being discussed in a positive or negative light. This information can then be used to make better product design, analysis, and market share decisions.

In this work, the VADER sentiment analyzer was used to identify the sentiment of each review in the smartphone reviews dataset. By using VADER, the sentiment of each review in the dataset was quickly identified as positive, negative, or neutral. It also provided a measure of the degree to which the sentiment is favorable or negative. This information was used to summarize the aspect categories of the reviews based on their sentiments. This helped to understand the overall sentiment of a review and identify which aspects of the product are being discussed in a positive or negative light, which proved beneficial for decision-making in product design, analysis, and market share.

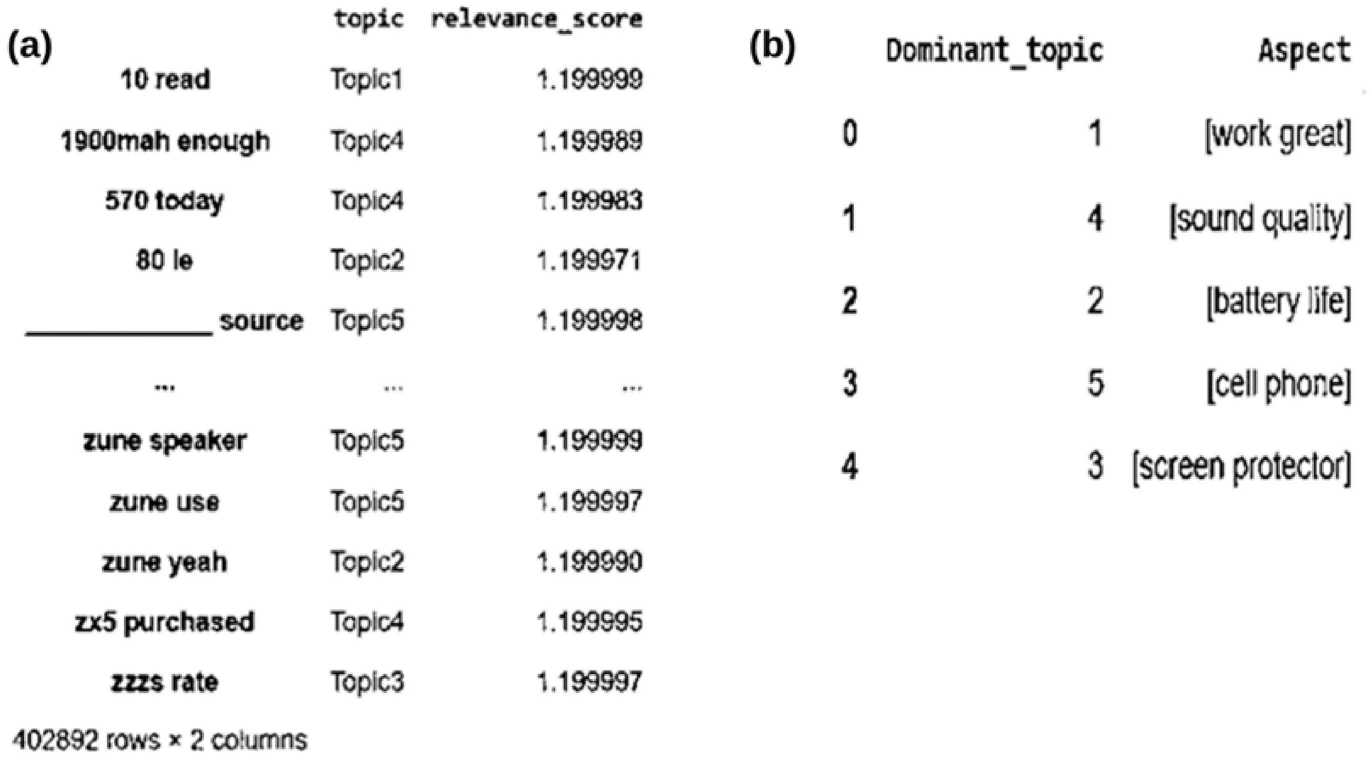
2.4 Performance Evaluation Criteria

In this work, the performance of the proposed sentiment analysis approach using LDA for aspect term extraction was evaluated in terms of accuracy. Accuracy is a commonly used metric in machine learning and natural language processing to evaluate the performance of a model. It measures the proportion of correctly classified instances out of the total number of instances. The predicted sentiments' accuracy was calculated to evaluate the proposed method's performance using a set of test data, which was not used during the training of the model. The test data was labeled with the correct sentiments, and the model's predictions were compared to these labels. The accuracy was calculated as the proportion of correctly classified instances out of the total number of instances in the test data.

3 Results and Discussion

The proposed work was implemented using the Python programming language, and the results obtained were analyzed. The smartphone review dataset was preprocessed to fit the needs for implementing the Latent Dirichlet Allocation (LDA) technique. Sentiments of the reviews were then identified using the VADER sentiment analyzer, and dominant topics were obtained by applying this technique. The relevance score of the selected topics was also generated, as shown in Figure 1(a). The final step in the process was to extract aspect categories from the consumer review dataset, as illustrated in Figure 1(b). The Latent Dirichlet Allocation (LDA) topic modeling technique was used for this purpose, and it proved to be a valuable tool for further project analysis. Out of the total five aspect categories extracted from the dataset, four were identified with high accuracy. Figure 2(a) illustrates the sentiment polarity of aspect categories extracted from the smartphone reviews dataset. The aspect categories are classified as positive, negative, or neutral based on the sentiments identified by the VADER sentiment analyzer. Figure 2(b) illustrates the summary of aspect categories extracted from the smartphone reviews dataset in a stacked bar graph.

The obtained results indicate that the model correctly identified four out of five categories, resulting in an accuracy of 80%. While the achieved accuracy may not be very high, it is important to note that sentiment analysis is a challenging task, especially when dealing with user-generated text data, which can be very subjective and difficult to classify. Additionally, the accuracy can be improved by using more sophisticated models or more data. The performance can also be evaluated using other metrics such as precision, recall, and F1-score, which are also commonly used in NLP and Sentiment Analysis tasks. However, the proposed approach extracted aspect categories from the data and identified each review's sentiment using the VADER sentiment analyzer with an acceptable level of accuracy. It can be considered a starting point for further development of sentiment analysis methodologies. In the future, the proposed method can be expanded to other domains with the help of more data and sophisticated models.



(a)

	topic	relevance_score
10 read	Topic1	1.199999
1900mah enough	Topic4	1.199989
570 today	Topic4	1.199983
80 le	Topic2	1.199971
_____________ source	Topic5	1.199998
...	...	...
zune speaker	Topic5	1.199999
zune use	Topic5	1.199997
zune yeah	Topic2	1.199990
zx5 purchased	Topic4	1.199995
zzzs rate	Topic3	1.199997

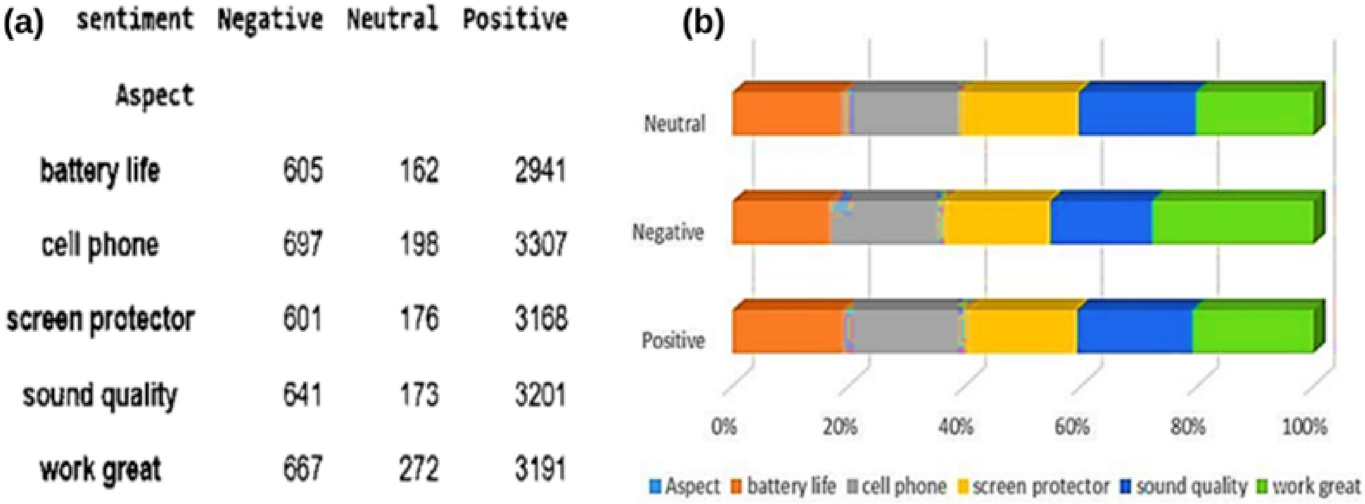
402892 rows × 2 columns

(b)

	Dominant_topic	Aspect
0	1	[work great]
1	4	[sound quality]
2	2	[battery life]
3	5	[cell phone]
4	3	[screen protector]

Figure 1: **(a)** Generated relevance scores for the selected topics; **(b)** Extracted aspect categories for smartphone reviews

Additionally, it can be used in various applications such as customer relationship management, product design, and marketing. Therefore, the proposed method has many potential applications and can be a valuable tool for businesses and organizations that rely on customer feedback to make decisions.



(a)

sentiment	Negative	Neutral	Positive
Aspect			
battery life	605	162	2941
cell phone	697	198	3307
screen protector	601	176	3168
sound quality	641	173	3201
work great	667	272	3191

Figure 2: **(a)** Sentiment polarity of obtained categories; **(b)** Aspect summarization

4 Conclusion

In this article, we proposed a method for extracting aspect categories and identifying the sentiment of customer reviews for smartphone products. The proposed method used a combination of preprocessing steps, such as lower punctuation, stop words removal, tokenization, and lemmatization, to prepare the dataset for analysis. The Latent Dirichlet Allocation (LDA) technique was then used to extract aspect categories from the data, while the VADER sentiment analyzer was used to identify the sentiment of each review. The model correctly identified three out of five categories, thus having an accuracy of 80% on this dataset. The achieved accuracy result of 60% indicates that the proposed method could correctly classify the sentiment of 60% of the instances in the test data. While this accuracy may not be very high, it is important to note that sentiment analysis is a challenging task, especially when dealing with user-generated text data, which can be very subjective and difficult to classify.

In conclusion, the proposed approach extracted aspect categories from the smartphone review data and identified each review's sentiment using the VADER sentiment analyzer with an acceptable level of accuracy. This method can be a starting point for further developing sentiment analysis methodologies for smartphone products.

Declaration of Competing Interests

The author declares that she has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Declaration

This research did not receive any grants from governmental, private, or nonprofit funding bodies.

Author Contribution

Lovish Rajput: Conceptualization, Visualization, Investigation, Methodology, Data curation, Writing - Original draft preparation, Writing - Reviewing **Shilpi Gupta**: Conceptualization, Visualization, Investigation, Methodology, Data curation, Writing - Reviewing

References

[1] Z. Li, Y. Fan, B. Jiang, T. Lei, and W. Liu, "A survey on sentiment analysis and opinion mining for social multimedia," *Multimedia Tools and Applications*, vol. 78, no. 6, pp. 6939–6967, 2019.

[2] C. Zong, R. Xia, and J. Zhang, "Sentiment analysis and opinion mining," in *Text Data Mining*, Singapore: Springer Singapore, 2021, pp. 163–199.

[3] R. Nimesh, P. Veera Raghava, S. Prince Mary, and B. Bharathi, "A survey on opinion mining and sentiment analysis," *IOP Conference Series: Materials Science and Engineering*, vol. 590, no. 1, pp. 14–46, 2019.

[4] M. S. Hossain and M. F. Rahman, "Customer sentiment analysis and prediction of insurance products' Reviews Using Machine Learning Approaches," *FIIB Business Review*, p. 231971452211157, 2022.

[5] D.-F. Ciocodeică, R.-G. (Popa) Chivu, I.-C. Popa, H. Mihălcescu, G. Orzan, and A.-M. (Dumitrache) Băjan, "The degree of adoption of business intelligence in romanian companies—the case of sentiment analysis as a marketing analytical tool," *Sustainability*, vol. 14, no. 12, p. 7518, 2022.

[6] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowledge-Based Systems*, vol. 226, p. 107134, 2021.

[7] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15169–15211, 2019.

[8] R. Priyantina and R. Sarno, "Sentiment analysis of hotel reviews using latent dirichlet allocation, semantic similarity and LSTM," *International Journal of Intelligent Engineering and Systems*, vol. 12, no. 4, pp. 142–155, 2019.

[9] M. F. A. Bashri and R. Kusumaningrum, "Sentiment analysis using latent dirichlet allocation and topic polarity wordcloud visualization," in *2017 5th International Conference on Information and Communication Technology (ICoIC7)*, May 2017, pp. 1–5.

[10] A. Goyal and I. Kashyap, "Latent dirichlet allocation - an approach for topic discovery," in *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, May 2022, pp. 97–102.

[11] T. A. Rana and Y.-N. Cheah, "Aspect extraction in sentiment analysis: comparative analysis and survey," *Artificial Intelligence Review*, vol. 46, no. 4, pp. 459–483, 2016.

[12] Y.-C. Chang, C.-H. Ku, and C.-H. Chen, "Social media analytics: Extracting and visualizing Hilton hotel ratings and reviews from TripAdvisor," *International Journal of Information Management*, vol. 48, pp. 263–279, 2019.

[13] O. Alqaryouti, N. Siyam, A. Abdel Monem, and K. Shaalan, "Aspect-based sentiment analysis using smart government review data," *Applied Computing and Informatics*, 2020.

[14] S. De, S. Dey, S. Bhatia, and S. Bhattacharyya, ”An introduction to data mining in social networks,” in *Advanced Data Mining Tools and Methods for Social Computing*, Elsevier, 2022, pp. 1–25.

[15] W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam, ”Towards generative aspect-based sentiment analysis,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2021, vol. 2, pp. 504–510.

[16] J. Wang, B. Xu, and Y. Zu, ”Deep learning for aspect-based sentiment analysis,” in *2021 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE)*, Jul. 2021, pp. 267–271.

[17] B. Verma and R. S. Thakur, ”Sentiment analysis using lexicon and machine learning-based approaches: a survey,” in *Lecture Notes in Networks and Systems*, vol. 34, 2018, pp. 441–447.

[18] K. Crowston, X. Liu, and E. E. Allen, ”Machine learning and rule-based automated coding of qualitative data,” *Proceedings of the American Society for Information Science and Technology*, vol. 47, no. 1, pp. 1–2, 2010.

[19] Z. Fachrina and D. H. Widyantoro, ”Aspect-sentiment classification in opinion mining using the combination of rule-based and machine learning,” in *2017 International Conference on Data and Software Engineering (ICoDSE)*, Nov. 2017, pp. 1–6.

[20] T. A. Rana and Y.-N. Cheah, ”Sequential patterns rule-based approach for opinion target extraction from customer reviews,” *Journal of Information Science*, vol. 45, no. 5, pp. 643–655, 2019.

[21] V. Bonta, N. Kumaresh, and N. Janardhan, ”A comprehensive study on lexicon based approaches for sentiment analysis,” *Asian Journal of Computer Science and Technology*, vol. 8, no. S2, pp. 1–6, 2019.

[22] S. Sohangir, N. Petty, and Di. Wang, ”Financial sentiment lexicon analysis,” in *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, Jan. 2018, pp. 286–289.

[23] A. Veluchamy, H. Nguyen, M. L. Diop, and R. Iqbal, ”Comparative study of sentiment analysis with product reviews using machine learning and lexicon-based approaches,” *SMU Data Science Review*, vol. 1, no. 4, pp. 1–22, 2018.

[24] Z. Madhoushi, A. R. Hamdan, and S. Zainudin, ”Sentiment analysis techniques in recent works,” in *2015 Science and Information Conference (SAI)*, Jul. 2015, pp. 288–291.

[25] S. Momtazi and F. Naumann, ”Topic modeling for expert finding using latent Dirichlet allocation,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 3, no. 5, pp. 346–353, 2013.

[26] F. Gurcan and N. E. Cagiltay, ”Big data software engineering: analysis of knowledge domains and skill sets using LDA-based topic modeling,” *IEEE Access*, vol. 7, pp. 82541–82552, 2019.

[27] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, ”The author-topic model for authors and documents,” 2012.