

Aspect term extraction from multi-source domain using enhanced latent Dirichlet allocation

Radhika Jinendra Dhanal^{1,2}, Vijay Ram Ghorpade³

¹Department of Computer Science and Engineering, D. Y. Patil College of Engineering and Technology, Shivaji University, Kolhapur, India

²Department of Technology, Shivaji University, Kolhapur, India

³Department of Computer Science and Engineering, Bharati Vidyapeeth College of Engineering, Shivaji University, Kolhapur, India

Article Info

Article history:

Received Jan 16, 2024

Revised Feb 28, 2024

Accepted Mar 21, 2024

Keywords:

Aspect term extraction

Domain specific

Latent Dirichlet allocation

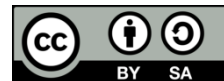
Natural language processing

Sentiment analysis

ABSTRACT

This study presents a comprehensive exploration of sentiment analysis across diverse domains through the introduction of a multi-source domain dataset encompassing hospitals, laptops, restaurants, cell phones, and electronics. Leveraging this extensive dataset, an enhanced latent Dirichlet allocation (E-LDA) model is proposed for topic modeling and aspect extraction, demonstrating superior performance with a remarkable coherence score of 0.5727. Comparative analyses with traditional LDA and other existing models showcase the efficacy of E-LDA in capturing sentiments and specific attributes within different domains. The extracted topics and aspects reveal valuable insights into domain-specific sentiments and aspects, contributing to the advancement of sentiment analysis methodologies. The findings underscore the significance of considering multi-source datasets for a more holistic understanding of sentiment in diverse text corpora.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Radhika Jinendra Dhanal

Department of Computer Science and Engineering, D. Y. Patil College of Engineering and Technology

Shivaji University

Kolhapur, India

Email: dhanallraddheka@gmail.com

1. INTRODUCTION

Automated sentiment analysis, a field within natural-language-processing (NLP), aims to computationally determine the sentiment expressed in textual data, categorizing it as positive, negative, or neutral [1]. Over the years, the integration of machine-learning (ML) and deep-learning (DL) techniques has significantly advanced the efficiency of automated sentiment analysis [2]. ML algorithms, ranging from traditional methods like support-vector-machines (SVM) [3] to more sophisticated models like random-forests [4] and gradient-boosting [5], have played a crucial role in capturing nuanced patterns in language to discern sentiments. Additionally, DL models, such as recurrent-neural-networks (RNNs) [6] and convolutional-neural-networks (CNNs) [7], have demonstrated remarkable success in learning intricate features from textual data, improving sentiment analysis accuracy. Despite the effort made in automated sentiment analysis, there are inherent challenges. One major issue is the interpretability of DL models, making it challenging to understand how decisions are reached [8]. Additionally, sentiment analysis models struggle with sarcasm [9], context-dependent sentiments [10], and domain-specific language [11], leading to inaccuracies in predictions. As these methods are applied to diverse datasets, the need for improved performance becomes crucial to ensure robust sentiment analysis across various domains. One key aspect that contributes to enhancing sentiment analysis is the consideration of aspects within the text. Aspects are

specific features or attributes associated with entities mentioned in the text, and they play a vital role in understanding the importance of sentiments. However, current research often focuses on sentiment analysis within a specific sector, neglecting the complexity introduced by multi-source data encompassing various domains.

To address these limitations, this work introduces a comprehensive dataset collected from diverse sectors, including hospitals, laptops, restaurants, cellphones, and electronics. This multi-source dataset allows for a holistic understanding of sentiments across different domains. To extract meaningful insights, an enhanced latent Dirichlet allocation (E-LDA) model is employed for topic modeling, unraveling latent topics within the review text corpus. Subsequently, an analysis of reviews pertaining to identified topics enables the extraction of aspects, providing valuable attributes for sentiment analysis. In contrast to existing approaches that focus on individual sectors [12], [13], this work pioneers a more holistic perspective by leveraging multi-source data. The proposed methodology not only broadens the scope of sentiment analysis but also tackles the challenge of aspect term extraction (ATE) from diverse domains. By employing E-LDA and analyzing reviews within topic contexts, this approach ensures an important understanding of sentiments, ultimately improving the accuracy and applicability of automated sentiment analysis across a wide range of sectors. The contribution of this work is as follows:

- Presented a multi-source dataset which comprises hospital, laptop, restaurant, cell phone and electronics dataset.
- Presented a novel technique called as E-LDA for extracting topics from review text corpus which is built on the base model of LDA.
- Presented a model for various sectors, showcasing the potential for a more refined and adaptable sentiment analysis framework.

The manuscript follows a structured organization. Section 2 provides a survey, exploring current challenges in automatic sentiment prediction and examining prevalent models for ATE in sentiment prediction. In section 3 details the methodology employed in the study. Moving to section 4, the results of E-LDA are presented, comparing with LDA and other relevant works. This section highlights the extraction of topics and the identification of frequent words (aspects) within the review text corpus. The manuscript concludes in section 5, summarizing key findings and contributions.

2. LITERATURE SURVEY

In this section, a comprehensive survey is presented, delving into the current challenges associated with automatic sentiment prediction. The exploration specifically focuses on understanding prevalent models utilized for ATE in the context of sentiment prediction. This survey aims to provide a thorough understanding of the existing works, shedding light on the issues and methodologies employed in the field of sentiment analysis. Kokatnoor and Krishnan [13], for effective analysis of sentiments and for topic-modeling using a review text dataset, a two-stage feature design operation was employed. The initial stage involved using a term-frequency inverse-document-frequency (TF-IDF) using improved trigrams estimation. The subsequent stage consisted of eliminating weak characteristics that existed in the collections which decreased the dimension of the SVM. As, the work was able to discover relevant and hidden topics within dataset through the use of improved distribution approach, the suggested approach outperformed latent-semantic-analysis (LSA) alongside hierarchical-dirichlet-process (HDP) approaches. This was achieved by combining batch-wise LDA using stochastic variational inference (SVI) through a two-stage feature design operation. By employing a two-stage feature design operation, the Batch-wise LDA using SVI was able to boost its efficiency relative to HDP alongside LSA approach by 4% and 9%, respectively, with respect to coherence scores. Wankhade *et al.* [14], primarily aimed to examine and comprehensively list the benefits and drawbacks of algorithmic classification in analyzing sentiment. According to their research, recently suggested methods were typically compared to classification utilizing SVM and naïve-bayes (NB) algorithms. Another area that was examined in the examination was the correlation among sentiment-review design and sentiment-analysis challenges. Domain dependency was revealed by this examination, which is crucial for diagnosing sentiment problems. Raghunathan and Saravanakumar [15], explored lexicon and ML-based methods for sentiment evaluation. They looked at the difficulties of four distinct kinds of sentiment categorization in this study: cross-lingual, cross-domain, small-scale, and short-term applications. In this study, they came to a conclusion that biggest challenges of cross-lingual was creating an approach which can incorporate data through multiple languages. Further, the biggest challenge of cross-domain was creating an approach for categorizing sentiments across various domains. Further the issue of small scale and short term was predicting the sentiments with less data and within real-time scenarios. Tan *et al.* [16], used conventional ML methods, which included cleaning the reviews using stop-words, standardizing it, and then representing it utilizing features dependent on frequency, like bag-of-words or TF-IDF. For categorization,

the filtered review was subsequently input into ML methods like SVM, NB, etc. They discovered that the research community has moved from natural-language-processing (NLP) to DL approaches due to the developments in NLP. Pretrained embeddings of words like fastText, word2vec and GloVe were initially used to extract aspects from reviews in DL. DL methods like CNN, gated-recurrent-unit (GRU), and long-short-term-memory (LSTM) receive input from with the data collected through these embedded data. One more finding was ensemble techniques, which involved merging estimates from various DL or ML approaches, which was also utilized in certain sentiment assessment efforts to enhance effectiveness. A few popular datasets for sentiment research included SemEval, Twitter and IMDb. The necessity for reliable language approach was underscored by the fact that, despite advancements in sentiment assessment, current approaches are prone to insufficiently structured and sarcastic texts.

Ekinci and Omurca [17], introduced concept-LDA (C-LDA), a new approach for web-based review-document analysis that relied on topic modeling to uncover hidden characteristics. Topics which included both co-occurring and semantically relevant words were obtained by enriching the review space of features using named entities and concepts and taken from Babely approach in C-LDA. By comparing the results across 10 freely accessible databases, they found that C-LDA outperformed LDA approach when it comes to topic-quality along with coherence-scores. With Concept-LDA's learnt topic accountability, it was able to accomplish the ATE with ease and accuracy. Pathan and Prakash [18], among the many unsupervised topic-modeling methods, LSA and LDA are the most popular. The suggested approach was evaluated in this study using three separate reviews information sets: movie, mobile and hotel. Specifically, they utilized LSA and LDA, in their work. Although both employed distinct depictions for the texts, each are capable of extracting topics using reviews. The experimental findings showed that both approaches were successful in ATE, however LDA performed better than LSA. Two multi-class classification methods, namely SVM and NB were used to classify the collected attributes. It was noted that the SVM approach outperformed NB approach when it came to topic-aspect classification when LDA is used. Farkhod *et al.* [19], utilized an unsupervised ML method to find the polarity of sentiments in both documents and words. The suggested approach utilizes LDA topics modeling and joint-sentiment-topic (JST) as its foundation. The study was conducted using the IMDB dataset, which included user evaluations. Following the use of the LDA approach to extract reviewing topics, the approach was used to determine polarity of sentiments at the word, document and topic levels. For displaying data, the LDavis package was employed. Experiments demonstrated that the approach was successful in topic-partition along with text and phrase sentiment categorizations demonstrating excellent accuracy in sentiment evaluation. Ali *et al.* [20], set out to discover the fundamental topics and sentiments represented by reviews found on the internet. The LDA topic modeling method yielded four topics. This study's results shed light on how management may improve the tourism experience by recognizing the behaviors of tourists and the perspectives of those in charge. But there are a few drawbacks to this study. The initial problem was that this theoretical structure mostly made utilization of rule-based sentiment evaluation methods, which were unable to identify sarcasm, irony, and possibilities. Furthermore, because the information was only used for one unique location, it would be unwise to generalize the aspect, rating and sentiment allocation results to other uses within the same domain.

3. METHOD

This section discusses the proposed architecture for the current work, providing a detailed overview of the intended approach. Subsequently, it delves into the dataset collection process, providing sources and methodologies employed to gather the required data. Following this, the discussion extends to the data preprocessing phase, which discusses the steps taken to refine and transform raw textual information into a structured and usable format. The discourse then shifts towards the application of techniques such as TF-IDF and E-LDA, discussing their roles in the context of the work. Furthermore, the discussion encompasses the process of extracting topics and aspects from the data, offering insights into the methodologies applied to discover meaningful patterns within the text corpus.

3.1. Architecture

In the adopted architecture for handling the multi-source domain dataset, as presented in Figure 1, a comprehensive data preprocessing step was implemented. Initially, the entire text corpus was converted to lowercase, and unnecessary punctuation such as commas, apostrophes, and colons were removed to ensure uniformity and consistency. Subsequently, the text was tokenized by separating each word with commas, and stop words were employed to filter out irrelevant words and part-of-speech (PoS) tags were used for extracting most important words, thereby enhancing the extraction of relevant aspects. The processed and cleaned reviews were then subjected to the TF-IDF transformation, converting the textual data into a numerical format. The transformed data was subsequently fed into the E-LDA model, where coherence

scores were evaluated. This facilitated the extraction of coherent topics and aspects from the multi-source domain dataset, providing valuable insights into the underlying patterns and themes within the text data.

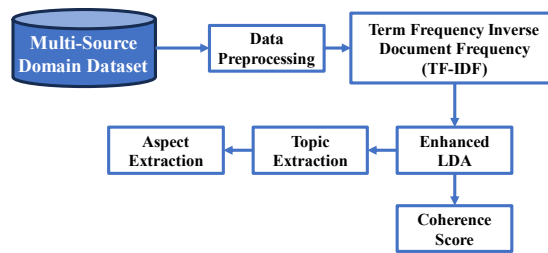


Figure 1. Architecture of proposed work

3.2. Dataset collection

The dataset comprises information gathered from various sources, including hospital [21], [22], semeval 2014 task 4 [23], and amazon [24] as presented in Figure 2. Within the hospital dataset, data has been acquired from Practo and Mouthshut using the web-scraping application programming interface (API). Specifically, the SemEval 2014 task 4 dataset contributes laptop and restaurant data exclusively. Furthermore, the Amazon dataset encompasses cell phone and electronic data. These distinct datasets have been consolidated to create a comprehensive multi-source domain dataset.

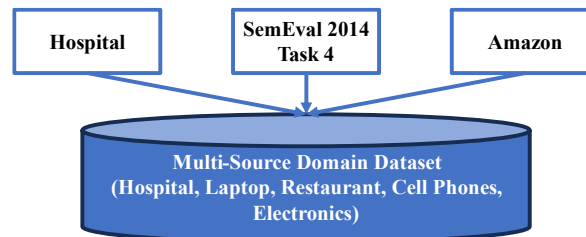


Figure 2. Dataset collection

3.3. Data preprocessing

In the initial phase of data preprocessing, a series of steps were executed to enhance the quality and uniformity of the text data. Firstly, the conversion of uppercase characters to lowercase was implemented, ensuring consistent formatting. Subsequently, the removal of punctuation marks took place, aiming to streamline the text further. Tokenization followed, breaking down the sequence of text into smaller, meaningful units. The integration of stop words was then employed to filter out commonly used words with minimal informative value. Lemmatization, the process of reducing words to their base or root form, was applied to identify underlying similarities. Additionally, PoS tagging was incorporated, assigning grammatical categories to each word, such as nouns, verbs, or adjectives. Lastly, a dictionary was employed to identify and compile a list of all unique words present in the text corpus, contributing to a comprehensive understanding of the vocabulary used.

3.4. Term frequency inverse document frequency, enhanced LDA, topic and aspect extraction

This study initially focuses on the comprehensive analysis of a multi-source dataset, undertaking meticulous data preprocessing to enhance the quality of the information. Following this preprocessing phase, the LDA is employed to extract relevant topics and assess the coherence score within the dataset. Subsequently, the TF-IDF technique is applied to the preprocessed data to convert textual information into a numerical representation. Building upon this transformation, the E-LDA model is then utilized to further extract topics and assess the coherence score, providing a comprehensive understanding of the underlying topics within the multi-source dataset. This sequential approach allows for a more refined exploration of topics and insights, contributing to the overall effectiveness of the analysis. The complete flow is presented in Figure 3.

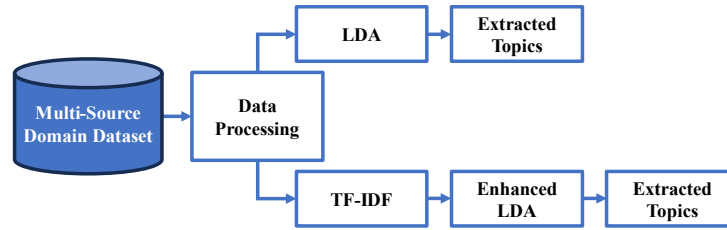


Figure 3. Topic extraction architecture

TF-IDF is a popular weighting approach often utilized in text analysis. Weighting is a crucial step in converting textual information to a number format which can be quickly and processed effortlessly by computers. The TF represents the frequency of a single word within a review. The frequency of a single word within a review directly affects its TF importance. The IDF represents the frequency of a given word across all available reviews. Words that frequently show up in multiple reviews are considered frequently observed and must be avoided. Hence, TF-IDF has to be calculated precisely. For evaluating the TF for every word in a given review, is used in (1):

$$w_{TF(W)} = \begin{cases} \text{if } 0, & \text{then } TF = 0 \\ \text{if } TF > 0, & \text{then } 1 + \log(TF(W)) \end{cases} \tag{1}$$

where, $w_{TF(W)}$ represents the weight for every word in TF and $TF(W)$ represents the frequency of words appearing in a given review. Further, the IDF is evaluated for every word in a given review using in (2):

$$IDF(W) = \log\left(\frac{|R|}{DF(W)}\right) \tag{2}$$

where, $IDF(W)$ represents the weight of IDF for each word in a given review, $|R|$ represents total reviews, and $DF(W)$ represents overall reviews which contain the W words in review. Finally, the $TF - IDF$ is calculated using in (3):

$$TF - IDF(W) = w_{TF_{W,R}} \times IDF_{W,R} \tag{3}$$

where, $TF - IDF(W)$ represents each value for W word in a given R review. After calculating the TF-IDF values for each word within the review, it is important to remove words with less TF-IDF values. The following steps are followed for removal. First, arrange the words W in decreasing order according to their TF-IDF values. Then, select a percentage value to perform the necessary calculations. Using the calculated value, the threshold is determined. Finally, using the threshold, words having having higher TD-IDF value than the threshold value are selected, while those with a lower value are removed. After the selection, the E-LDA is used. The comparison of base LDA and E-LDA is shown in Figure 4 and Figure 5 respectively.

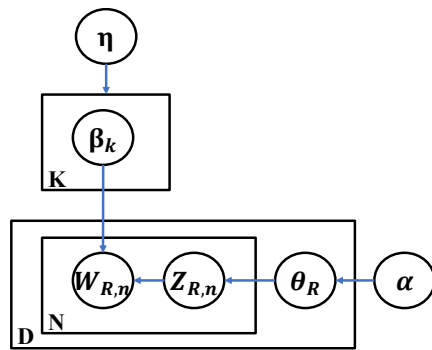


Figure 4. Base LDA

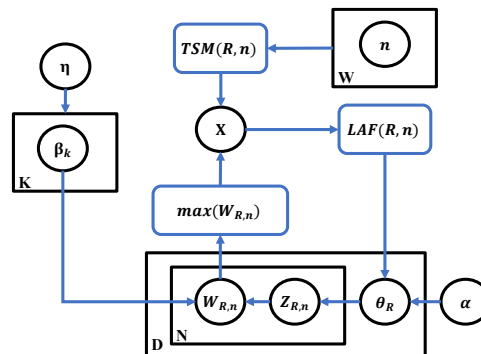


Figure 5. Enhanced LDA

LDA has become an effective method for identifying underlying patterns within reviews R . It has the ability to analyze the data and provide topic-vectors through a series of likelihood distributions. LDA proves to be an extremely efficient method in the field of modeling topics and clustering texts. In base LDA as presented in Figure 4, n represents the n^{th} position of a word within a given review R . α and η represent parameters of topic and proportion respectively. β represents a parameter that evaluates the mixture of words in every topic. K represents overall topics, θ_R represents the ratio of topics for each review. $Z_{R,n}$ denotes the matrix of reviews and words and $W_{R,n}$ denotes the likelihood distribution of words. In the LDA model, the assumption is made that the prior distribution of review topics follows a Dirichlet-distribution [19]. For a given review R , its topic distribution is represented as $\theta_R = \text{Dirichlet}(\vec{\alpha})$, where $\vec{\alpha}$ is a hyper-parameter vector with K dimensions. The topic assignment for the n^{th} word in review R , denoted as $z_{R,n}$, is calculated as $z_{R,n} = \text{multi}(\theta_R)$. Similarly, LDA presumes that the prior-distribution of words within a given topic is ruled by a Dirichlet distribution. Specifically, for any topic K , its word distribution is expressed as $\beta_k = \text{Dirichlet}(\vec{\eta})$. Finally, the observed word probability distribution for \mathcal{W}_{Rn} is $\mathcal{W}_{Rn} = \text{multi}(\beta_{z_{R,n}})$. In (1), the combined distribution of all visible and hidden variables within the LDA model can be effectively approximated through Gibbs sampling [25].

$$p(\mathcal{W}_R, z_R, \theta_R, \beta_k | \alpha, \beta) = \prod_{n=1}^N p(\theta_R | \alpha) p(z_{Rn} | \theta_R) p(\beta_k | \beta) p(\mathcal{W}_{Rn} | \beta_k) \quad (4)$$

Here, we present an enhanced LDA (E-LDA) which is presented in Figure 5 and consists of a likelihood-adjustment factor (LAF) for generating the best topics. This work considers the LAF for correcting the likelihood distribution of topics. The LAF value for each word in a given review is represented as $LAF(R, n)$. Further, the value for LAF is evaluated as in (5):

$$LAF(R, n) = TSM(R, n) \times \max(R, n) \quad (5)$$

where, $TSM(R, n)$ represents the topic significant measure for each word in the review and $\max(\mathcal{W}_{R,n})$ represents the highest likelihood value for all topics for each n^{th} position of a word in the given topic-word distribution matrix ($\mathcal{W}_{R,n}$). The coherence score for each topic with different clusters is evaluated using in (6) [26]:

$$\phi S_i(\vec{u}, \vec{w}) = \frac{\sum_{i=1}^{|\mathcal{W}|} u_i \cdot w_i}{\|\vec{u}\|_2 \cdot \|\vec{w}\|_2} \quad (6)$$

where ϕ represents the confirmation measure, S_i represents the pair of vectors u and w . \vec{u} and \vec{w} represent cosine vector similarity. After the evaluation of the coherence score, the topics that have the highest likelihood are selected using the pyLDAvis library which shows the different topics. The results achieved by the base LDA and E-LDA are discussed in the next section.

4. RESULTS AND DISCUSSION

This section provides an in-depth discussion of the results obtained through the implementation of both LDA and E-LDA models. Additionally, a comparative analysis is conducted by comparing the achieved outcomes with those of existing similar works [12], [13]. The experimentation for this study was conducted on a Windows 11 operating system, equipped with 16 GB of RAM and an Intel dual-core i7 processor. The evaluation utilized a carefully curated multi-source domain dataset. The models were developed using Python programming language, and the execution was carried out using the Anaconda integrated Python environment. Subsequently, this section presents a detailed account of the results attained by both LDA and E-LDA when configured for a hundred clusters.

The results obtained from both LDA and E-LDA for various cluster numbers were analyzed to discern their respective performances in topic modeling as presented in Figure 6 and Figure 7 respectively. It is observed that, in general, E-LDA exhibits competitive coherence scores compared to traditional LDA across a range of cluster values. Notably, at certain cluster numbers, E-LDA outperforms LDA in capturing meaningful topics within the review text corpus. The comparison reveals the importance in the efficacy of the two models, suggesting that the enhanced version, E-LDA, holds promise for improved topic modeling and aspect extraction. These findings show insights into the strengths of E-LDA in enhancing the understanding of latent structures within diverse textual datasets, paving the way for more refined sentiment analysis and aspect term extraction in multi-source domain scenarios.

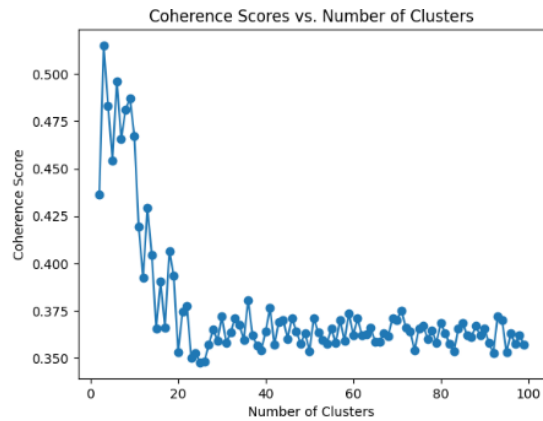


Figure 6. Coherence score for 0 to 100 clusters using LDA

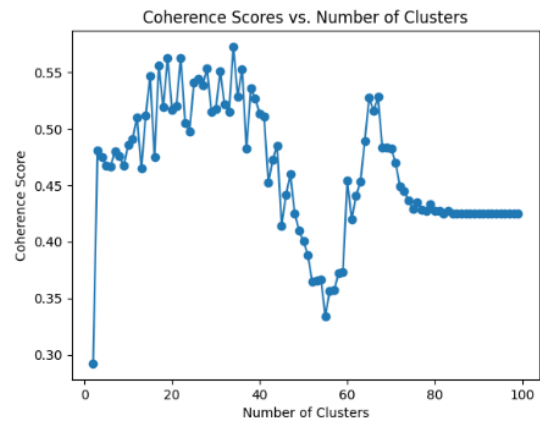


Figure 7. Coherence score for 0 to 100 clusters using E-LDA

Further, the results presented in Figure 8 indicate that both LDA and E-LDA achieved relatively high coherence scores, reflecting their effectiveness in capturing coherent and meaningful topics within the analyzed text corpus. However, it is noteworthy that E-LDA demonstrated a slightly higher coherence score (0.5727) compared to LDA (0.5356). This suggests that the enhanced version, E-LDA, exhibits an improved ability to unveil latent structures and relationships within the dataset, contributing to a more cohesive and interpretable representation of topics. The higher coherence score of E-LDA implies that it offers enhanced performance in topic modeling and aspect extraction, emphasizing its potential for more accurate sentiment analysis and important understanding of the underlying topics within the multi-source domain text corpus. In Figure 9, the results present the distribution of topics within the multi-source dataset, showcasing the count of topics associated with different domains. Notably, the most prevalent topics are related to the “Cell Phone” and “Electronics” domains, with counts of 4,997 and 4,511, respectively. This suggests that the dataset contains a substantial amount of information and reviews pertaining to these two domains. The “Laptop” domain follows closely with a count of 3,602, indicating a significant presence of laptop-related content. In contrast, topics related to “Hospital” and “Restaurant” have comparatively lower counts of 1,880 and 1,349, respectively, indicating a relatively smaller portion of the dataset dedicated to these domains. This distribution provides insights into the varied composition of the multi-source dataset, highlighting the prominence of specific domains and their corresponding topics within the overall text corpus.

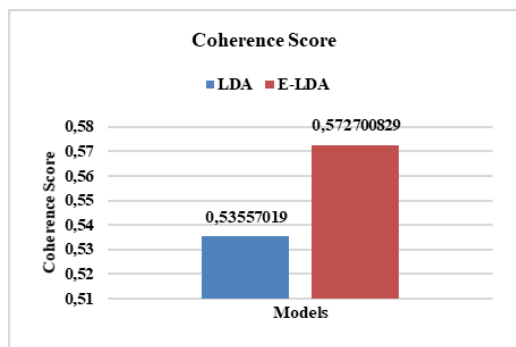


Figure 8. Highest coherence scores achieved by LDA and E-LDA

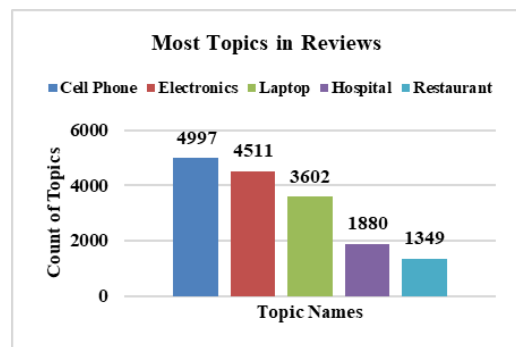


Figure 9. Most topics talked in reviews

Table 1 provides an overview of topics and their corresponding top keywords extracted from the multi-source dataset. Each topic is associated with keywords that represent key aspects within that domain. For instance, in the “Hospital” topic, keywords such as ‘doctor’, ‘staff’, and ‘treatment’ highlight aspects related to medical care. All the different topics and top keywords are presented in Table 1.

Table 1. Review, topics, and its aspect

| Topics | Top keywords |
|-------------|--|
| Hospital | ['hospital', 'service', 'doctor', 'staff', 'care', 'treatment', 'patient', 'surgery', 'excellent'] |
| Laptop | ['price', 'work', 'computer', 'laptop', 'battery', 'bought', 'sound', 'feature', 'plug'] |
| Restaurant | ['food', 'item', 'menu', 'table', 'wine', 'pizza', 'work', 'thai'] |
| Cell Phone | ['power', 'price', 'phone', 'feature', 'battery', 'bought', 'love'] |
| Electronics | ['screen', 'drive', 'amazon', 'tool', 'hold', 'line', 'camera', 'samsung', 'galaxy'] |

Table 2 presents specific reviews from the dataset along with their associated topics and aspects. For instance, the first review expresses positive experiences with a hospital, emphasizing aspects such as staff professionalism, convenient scheduling, and the overall ambiance of the place. The second review relates to a laptop, highlighting aspects of setup, installation, and the learning curve for users transitioning from a PC. The third review, associated with a restaurant, mentions an aspect related to the attentiveness of waiters, specifically pointing out a concern with the bill presentation. The fourth review, related to a cell phone, discusses aspects like the protective case, phone features, and the promised power doubling. The final review, associated with electronics, emphasizes aspects related to the functionality of a cable for connecting a camera to a TV and viewing videos.

The provided Table 3 presents a comparative analysis of various sentiment analysis models, including LDA-SV1 [13] (2020), LDA [27] (2021), LDA [12] (2023), base-LDA (proposed), and E-LDA (Proposed), across different datasets. Notably, the proposed E-LDA outperforms other models, including LDA-SV1, LDA [27], and LDA [12], in terms of coherence score when applied to a multi-source dataset comprising information from Hospital, Laptops, Restaurant, Cell Phone, and Electronics domains. E-LDA attains a high coherence score of 0.5727, indicating its superior ability to unveil coherent and meaningful topics within the diverse text corpus. In contrast, other models exhibit lower coherence scores when applied to specific datasets such as restaurant, musical instruments, mobile, automotive, IMDb, and banking. This underscores the efficacy of E-LDA in sentiment analysis across multi-source domains, showcasing its potential for understanding and improved topic modeling in diverse datasets.

Table 2. Review, topics and its aspect

| Review | Topics | Aspect |
|---|-------------|---|
| I got 4 implants from Dentzz in the last 6 months. I must say, they have an amazing set of technologically advanced equipment's and very professional staff. They scheduled all the appointments at my convenience. And all my sessions went smoothly. This place is really worth the hype. | Hospital | Staff, convenience, place |
| I like the Mini Mac it was easy to setup and install, but I am learning as I go and could use a tutorial to learn how to use some of the features, I was use to on the PC especially the right mouse click menu. | Laptop | Setup, install, learn, use, PC, Mouse click |
| The waiters were not attentive except that the bill turned up on the table before we were finished. | Restaurant | Bill |
| This is a fantastic case. Very stylish and protects my phone. Easy access to all buttons and features, without any loss of phone reception. But most importantly, its double power, just as promised. Great buy | Cell phone | Case, phone, access, loss, power, buy |
| What is there to say? I needed a cable to run from the camera to my TV and this is what was needed. It works as expected, and allows me to view my videos and pictures that are on my camera directly on my television. | Electronics | Cable, run, camera, TV, view, videos |

Table 3. Comparative study

| Models | Dataset | Coherence score |
|---------------------|---|-----------------|
| LDA-SV1 [13] | Restaurant | 0.513 |
| LDA [27] | Musical instruments | 0.36402 |
| | Mobile | 0.41770 |
| | Automotive | 0.45847 |
| | Restaurant | 0.46909 |
| | IMDb | 0.357489 |
| LDA [12] | Banking | 0.3919 |
| Base-LDA (proposed) | Multi-source (hospital, laptops, restaurant, cell phone, and electronics) | 0.5355 |
| E-LDA (proposed) | Multi-source (hospital, laptops, restaurant, cell phone, and electronics) | 0.5727 |

5. CONCLUSION

This work addresses the challenges in sentiment analysis by introducing a multi-source domain dataset encompassing diverse sectors such as hospitals, laptops, restaurants, cell phones, and electronics.

The utilization of this extensive dataset allowed for a comprehensive exploration of sentiment across various domains. The proposed E-LDA model demonstrated superior performance in topic modeling and aspect extraction, as presented in its higher coherence score compared to traditional LDA and other existing models. The analysis of extracted topics and aspects from reviews in different domains shed light on sentiments and specific attributes that influence opinions. The findings underscore the importance of considering multiple sources and domains for a more holistic understanding of sentiment in text data. Furthermore, the incorporation of aspects derived from the E-LDA model enhances the granularity of sentiment analysis by capturing domain-specific areas. This work contributes to the advancement of sentiment analysis methodologies, particularly in the context of multi-source datasets. Moving forward, the insights gained from this study can inform the development of more effective sentiment analysis models, facilitating improved decision-making in various sectors. The methodology employed, including data preprocessing, topic modeling, and ATE, serves as a valuable framework for future research in sentiment analysis across diverse domains. For future work, the work can be extended to extract the sentiments of the reviews, i.e., neutral, positive or negative. Overall, this work provides a significant step towards enhancing the accuracy and applicability of sentiment analysis in real-world scenarios.




REFERENCES

- [1] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," *Social Network Analysis and Mining*, vol. 11, no. 1, p. 81, Dec. 2021, doi: 10.1007/s13278-021-00776-6.
- [2] I. H. Sarker, "Machine learning: algorithms, real-world applications and research directions," *SN Computer Science*, vol. 2, no. 3, p. 160, May 2021, doi: 10.1007/s42979-021-00592-x.
- [3] C. P. D. Cyril, J. R. Beulah, N. Subramani, P. Mohan, A. Harshavardhan, and D. Sivabalaselvamani, "An automated learning model for sentiment analysis and data classification of Twitter data using balanced CA-SVM," *Concurrent Engineering*, vol. 29, no. 4, pp. 386–395, Dec. 2021, doi: 10.1177/1063293X211031485.
- [4] I. Setiawan *et al.*, "Utilizing random forest algorithm for sentiment prediction based on Twitter data," in *Proceedings of the First Mandalika International Multi-Conference on Science and Engineering 2022, MIMSE 2022 (Informatics and Computer Science)*, Dordrecht: Atlantis Press International BV, 2022, pp. 446–456. doi: 10.2991/978-94-6463-084-8_37.
- [5] S. Singh, D. Krishnan, P. Sehgal, H. Sharma, T. Surani, and J. Singh, "Gradient boosting approach for sentiment analysis for job recommendation and candidate profiling," in *2022 IEEE Bombay Section Signature Conference (IBSSC)*, IEEE, Dec. 2022, pp. 1–6. doi: 10.1109/IBSSC56953.2022.10037443.
- [6] L. Kurniasari and A. Setyanto, "Sentiment analysis using recurrent neural network," *Journal of Physics: Conference Series*, vol. 1471, no. 1, p. 012018, Feb. 2020, doi: 10.1088/1742-6596/1471/1/012018.
- [7] T. Diwan and J. V. Tembhurne, "Sentiment analysis: a convolutional neural networks perspective," *Multimedia Tools and Applications*, vol. 81, no. 30, pp. 44405–44429, Dec. 2022, doi: 10.1007/s11042-021-11759-2.
- [8] I. H. Sarker, "Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions," *SN Computer Science*, vol. 2, no. 6, p. 420, Nov. 2021, doi: 10.1007/s42979-021-00815-1.
- [9] Y. Y. Tan, C.-O. Chow, J. Kanesan, J. H. Chuah, and Y. Lim, "Sentiment analysis and sarcasm detection using deep multi-task learning," *Wireless Personal Communications*, vol. 129, no. 3, pp. 2213–2237, Apr. 2023, doi: 10.1007/s11277-023-10235-4.
- [10] L. Deng, B. Liu, Z. Li, J. Ma, and H. Li, "Context-dependent multimodal sentiment analysis based on a complex attention mechanism," *Electronics*, vol. 12, no. 16, p. 3516, Aug. 2023, doi: 10.3390/electronics12163516.
- [11] T. Kincl, M. Novák, and J. Přebil, "Improving sentiment analysis performance on morphologically rich languages: Language and domain independent approach," *Computer Speech & Language*, vol. 56, pp. 36–51, Jul. 2019, doi: 10.1016/j.csl.2019.01.001.
- [12] B. Ogunleye, T. Maswera, L. Hirsch, J. Gaudoin, and T. Brunson, "Comparison of topic modelling approaches in the banking context," *Applied Sciences*, vol. 13, no. 2, p. 797, Jan. 2023, doi: 10.3390/app13020797.
- [13] S. Kokatnoor and B. Krishnan, "A two-stepped feature engineering process for topic modeling using batchwise LDA with stochastic variational inference model," *International Journal of Intelligent Engineering and Systems*, vol. 13, no. 4, pp. 333–345, Aug. 2020, doi: 10.22266/ijies2020.0831.29.
- [14] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5731–5780, Oct. 2022, doi: 10.1007/s10462-022-10144-1.
- [15] N. Raghunathan and K. Saravanakumar, "Challenges and issues in sentiment analysis: a comprehensive survey," *IEEE Access*, vol. 11, pp. 69626–69642, 2023, doi: 10.1109/ACCESS.2023.3293041.
- [16] K. L. Tan, C. P. Lee, and K. M. Lim, "A survey of sentiment analysis: approaches, datasets, and future research," *Applied Sciences*, vol. 13, no. 7, p. 4550, Apr. 2023, doi: 10.3390/app13074550.
- [17] E. Ekinici and S. İ. Omurca, "Concept-LDA: incorporating babelify into LDA for aspect extraction," *Journal of Information Science*, vol. 46, no. 3, pp. 406–418, Jun. 2020, doi: 10.1177/0165551519845854.
- [18] A. F. Pathan and C. Prakash, "Cross-domain aspect detection and categorization using machine learning for aspect-based opinion mining," *International Journal of Information Management Data Insights*, vol. 2, no. 2, p. 100099, Nov. 2022, doi: 10.1016/j.ijmei.2022.100099.
- [19] A. Farkhod, A. Abdusalomov, F. Makhmudov, and Y. I. Cho, "LDA-based topic modeling sentiment analysis using topic/document/sentence (TDS) model," *Applied Sciences*, vol. 11, no. 23, p. 11091, Nov. 2021, doi: 10.3390/app112311091.
- [20] T. Ali, B. Omar, and K. Soulaïmane, "Analyzing tourism reviews using an LDA topic-based sentiment analysis approach," *MethodsX*, vol. 9, p. 101894, 2022, doi: 10.1016/j.mex.2022.101894.
- [21] Practo, "Practo | book doctor appointments online, order medicine, diagnostic tests, consult," *Www.Practo.Com*, 2017, [Online]. Available: <https://www.practo.com/>
- [22] "Consumer reviews on movies, cars, bikes, mobile phones, music, books, airlines, restaurants, hotels and more - MouthShut.com." Accessed: Feb. 13, 2024. [Online]. Available: <https://www.mouthshut.com/>
- [23] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androulopoulos, and S. Manandhar, "SemEval-2014 task 4: aspect based sentiment analysis," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2014, pp. 27–35. doi: 10.3115/v1/S14-2004.




- [24] J. Ni, J. Li, and J. McAuley, "Justifying recommendations using distantly-labeled reviews and fine-grained aspects," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 188–197. doi: 10.18653/v1/D19-1018.
- [25] K. S. Arun and V. K. Govindan, "A hybrid deep learning architecture for latent topic-based image retrieval," *Data Science and Engineering*, vol. 3, no. 2, pp. 166–195, Jun. 2018, doi: 10.1007/s41019-018-0063-7.
- [26] S. Syed and M. Spruit, "Full-text or abstract? examining topic coherence scores using latent dirichlet allocation," in *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, Oct. 2017, pp. 165–174. doi: 10.1109/DSAA.2017.61.
- [27] A. F. Pathan and C. Prakash, "Unsupervised aspect extraction algorithm for opinion mining using topic modeling," *Global Transitions Proceedings*, vol. 2, no. 2, pp. 492–499, Nov. 2021, doi: 10.1016/j.gltp.2021.08.005.

BIOGRAPHIES OF AUTHORS



Radhika Jinendra Dhanal    is an academican from Kolhapur, Maharashtra, with a passion for Computer Science and Information Systems. She is currently affiliated with DY Patil College of Engineering and Technology, situated in Kasaba Bawada, Kolhapur, where she has been contributing to the field of education since 2003. Her academic journey led her to pursue a Master's in Engineering (M.E.) from Shivaji University in 2012, specializing in her area of interest. Her commitment to academic excellence and a deep understanding of Computer Science and Information Systems laid the foundation for her successful career. With a rich academic background, a wealth of experience, and a commitment to advancing the field of Computer Science, Miss Radhika Dhanal stands as a noteworthy figure in the academic landscape of DY Patil College of Engineering and Technology, Kolhapur. Her journey continues to inspire both students and fellow educators alike. She can be contacted at email: dhanallraddheka@gmail.com.



Vijay Ram Ghorpade    is an academic researcher from Shivaji University. He is the Professor/Principal of Bharti Vidyapeeth's College of Engineering, Kolhapur. The author has contributed to research in topics: Wireless sensor network and Quality of service. The author has an H-index of 7, co-authored 24 publications receiving 146 citations. He can be contacted at email: vijayghorpade@hotmail.com.