Topic Modeling for Customer Review Analysis: An Application Using LDA in R on Trustpilot Reviews

Md. Abu Towsif, Farjana Yesmin Opi, and Most. Sayma Khatun
Department of Computer Science and Engineering
American International University-Bangladesh (AIUB), Dhaka, Bangladesh
{22-47019-1, 22-47018-1, 22-47035-1}@student.aiub.edu

***Abstract*— Customer reviews provide valuable insights into consumer sentiment and product feedback, but their unstructured nature presents challenges for analysis. This study applies Latent Dirichlet Allocation (LDA), an unsupervised machine learning technique, to analyze Tesla customer reviews from Trustpilot. By extracting latent themes from the text corpus, we identified key topics related to Tesla's customer experience, including product performance, autonomous driving features, service quality, and delivery processes. The results reveal that while Tesla is praised for innovation and driving experience, concerns persist regarding service delays, battery reliability, and logistical inefficiencies. The study highlights the importance of preprocessing steps, such as text cleaning and term weighting, in enhancing topic coherence. Additionally, challenges related to topic selection, model interpretability, and the limitations of LDA's bag-of-words representation are discussed. Future research could incorporate hybrid models and sentiment analysis to improve topic modeling accuracy and provide deeper insights. This study demonstrates the potential of LDA in transforming unstructured customer feedback into actionable business intelligence.**

Keywords*— *Latent Dirichlet Allocation (LDA), Tesla customer reviews, topic modeling, text mining, sentiment analysis, natural language processing (NLP)

I. Introduction

In the era of digital transformation, customer reviews have become a cornerstone of business intelligence, offering invaluable insights into consumer preferences, satisfaction, and areas for improvement. Platforms like Trustpilot host millions of reviews across various industries, providing a rich dataset for analyzing customer sentiment and identifying emerging trends. However, the sheer volume and unstructured nature of these reviews pose significant challenges for manual analysis, necessitating the use of advanced computational techniques to extract meaningful information [1]. Topic modeling, a type of unsupervised machine learning, has emerged as a powerful tool for uncovering latent themes within large text corpora, enabling businesses to gain actionable insights from customer feedback [2]. Among the various topic modeling techniques, Latent Dirichlet Allocation (LDA) has gained widespread popularity due to its ability to identify coherent topics and its interpretability [3].

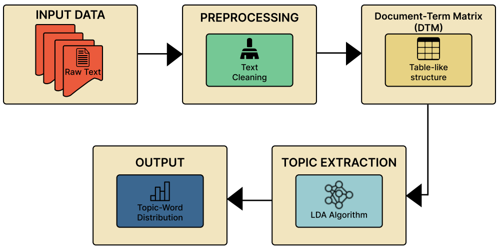


Fig.1: LDA process diagram

Despite its potential, the application of LDA in customer review analysis often requires careful preprocessing, parameter tuning, and domain-specific adaptations to ensure meaningful results [4].

Traditional methods of customer review analysis, such as sentiment analysis and keyword extraction, have been widely used to determine customer opinions. However, these approaches often fail to capture the complex themes and underlying topics that drive customer sentiment [5]. For instance, while sentiment analysis can classify reviews as positive or negative, it does not reveal the specific aspects of a product or service that customers are discussing. Similarly, keyword extraction methods may identify frequently occurring terms but lack the contextual understanding needed to group related terms into logical topics [6]. LDA addresses these limitations by modeling each document as a mixture of topics and each topic as a distribution of words, thereby enabling the discovery of hidden thematic structures within the text [7]. Nevertheless, the effectiveness of LDA depends heavily on the quality of preprocessing, the choice of hyperparameters, and the interpretability of the resulting topics, which can vary significantly across datasets and domains [8].

This study aims to demonstrate the application of LDA for customer review analysis using Trustpilot reviews as a case study. The primary objective is to identify and interpret the key topics discussed in customer reviews of Tesla, a leading electric vehicle manufacturer. By leveraging the LDA algorithm implemented in R, this research seeks to uncover the latent themes that characterize customer feedback, providing insights into the factors driving customer satisfaction and dissatisfaction. The study also highlights the importance of preprocessing steps, such as text cleaning, stop word removal, and term frequency-inverse document frequency (TF-IDF) weighting, in enhancing the quality of topic modeling results. Furthermore, the research explores the challenges of selecting the optimal number of topics and interpreting the output, offering practical recommendations for applying LDA in real-world scenarios.

The primary contribution of this work is the development of a reproducible framework for topic modeling using LDA in R, applied to Trustpilot reviews. By providing a step-by-step implementation guide, this study aims to make topic modeling accessible to researchers and practitioners with limited expertise in natural language processing (NLP). Additionally, the findings offer valuable insights into the key themes discussed in Tesla customer reviews, which can inform business strategies and improve customer engagement. This research contributes to the growing body of literature on NLP applications in business analytics, demonstrating the potential of topic modeling to transform unstructured text data into actionable knowledge.

II. Methodology

This section outlines the methodology used for topic modeling of customer reviews on Trustpilot, specifically for Tesla. The workflow involves multiple stages, including data acquisition, preprocessing, model development, training, and evaluation to extract meaningful insights from customer feedback.

A. *Data Acquisition*

The dataset for this study was sourced from Trustpilot, a popular customer review platform. Web scraping was used to collect Tesla-related reviews programmatically using the *rvest* package in R [9].

B. *Data Preprocessing*

To ensure high-quality input data for the topic modeling process, the following preprocessing steps were applied:

1) Text Extraction and Cleaning

Reviews were extracted using the html_elements and html_text functions from the rvest package [10]. To maintain uniformity, all text was converted to lowercase [11], while punctuation and numbers were removed to reduce noise in the data [12]. Additionally, common stopwords were eliminated using the tm package to filter out non-informative words [13], and extra white spaces were stripped to normalize the text structure [2].

2) Document-Term Matrix (DTM) Creation

The preprocessed text was converted into a Document-Term Matrix (DTM) using the tm package [3], and term frequency-inverse document frequency (TF-IDF) weighting was applied to emphasize important terms while down weighting frequently occurring but less meaningful words [8].

C. *Model Development*

Latent Dirichlet Allocation (LDA), an unsupervised machine learning algorithm, was used to discover hidden topics within the text corpus. The LDA model was implemented using the *topicmodels* package in R [17].

1) Number of Topics Selection

The number of topics was set to 15 based on empirical experimentation to ensure optimal topic representation [4]. Additionally, the model was initialized with a random seed to maintain reproducibility across multiple runs [7].

2) LDA Model Implementation

For the LDA model implementation, each document was treated as a mixture of topics, while each topic was modeled as a distribution of words [6]. Gibbs sampling was employed for topic inference and optimization, allowing the model to iteratively refine topic assignments for better coherence [5].

D. *Training and Validation*

The topic-word distributions were examined to ensure coherent topics [16]. The model's interpretability was validated using word probability distributions and visualization techniques [17].

E. *Evaluation*

The *tidytext* package was used to extract and analyze the most probable words for each topic, allowing for a detailed examination of thematic consistency [18, 19]. To enhance interpretability, a bar plot was generated using ggplot2, displaying the top words for each topic along with their associated probabilities [20]. This visualization provided an intuitive understanding of topic distributions and their significance, facilitating a clearer representation of the underlying themes [21].

F. *Workflow Diagram*

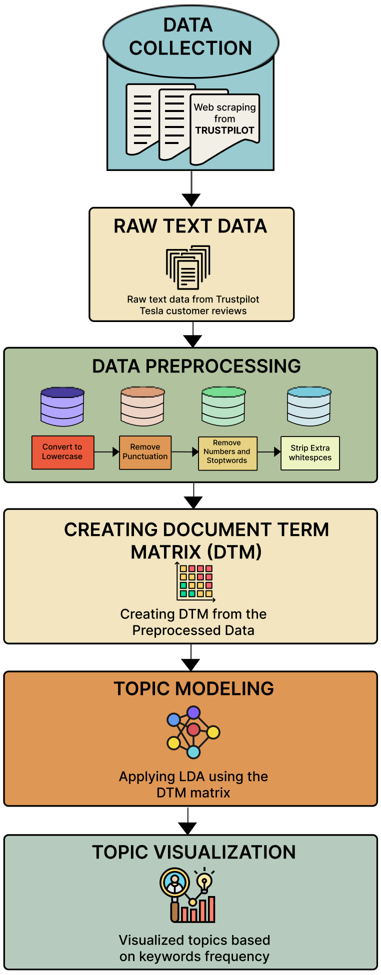

DATA COLLECTION
Web scraping from TRUSTPILOT
RAW TEXT DATA
Raw text data from Trustpilot Tesla customer reviews
DATA PREPROCESSING
Convert to Lowercase
Remove Punctuation
Remove Numbers and Stoptwords
Strip Extra whitespces
CREATING DOCUMENT TERM MATRIX (DTM)
Creating DTM from the Preprocessed Data
TOPIC MODELING
Applying LDA using the DTM matrix
TOPIC VISUALIZATION
Visualized topics based on keywords frequency

Fig. 1. Workflow of the proposed methodology

III. Results

Topic Modeling Using Latent Dirichlet Allocation (LDA) on Tesla Customer Reviews revealed several key themes from the dataset. By applying LDA to the preprocessed Trustpilot reviews, we identified 15 distinct topics, each corresponding to recurring themes or concerns expressed by customers. The LDA model revealed critical insights into customer sentiment, highlighting both areas of satisfaction and dissatisfaction.

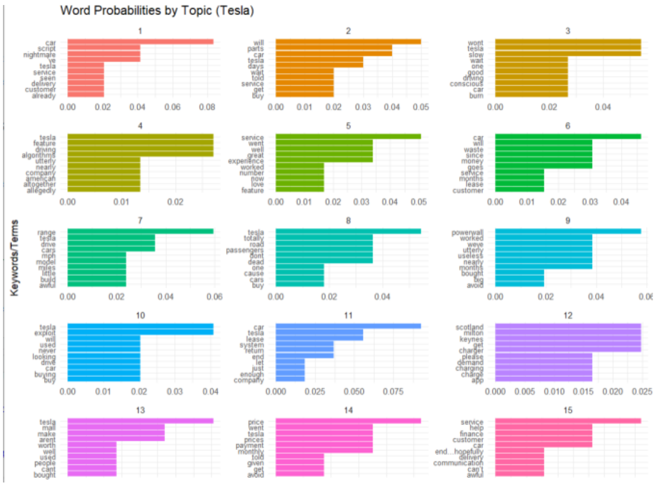


Fig.2. Topic Modeling of Tesla Customer Reviews Using Latent Dirichlet Allocation (LDA).

The first topic captured customer frustrations with delivery delays and service-related issues. The most frequent terms in this topic were *car, script, nightmare, service, delivery, customer,* and *already*. The recurring use of words such as "nightmare" and "already" indicated significant dissatisfaction with Tesla's delivery process, highlighting delays and poor communication during service interactions. Customers frequently expressed frustration with the overall experience of receiving their vehicles.

Topic 2 revealed significant concerns about the extended wait times for service appointments and replacement parts. Commonly occurring words like *will, parts, wait, service,* and *buy* suggested that many Tesla customers experienced prolonged waiting periods for essential services and repairs. Reviews related to this topic emphasized delays in receiving timely service, leading to dissatisfaction among customers who were left waiting for extended periods.

The fourth topic centered around Tesla's driving features, particularly its algorithms powering the autopilot and other automated driving systems. The most frequent terms in this topic included *tesla, driving, feature, algorithms, company,* and *American.* While some reviews reflected excitement about the innovations in Tesla's driving technology, others raised skepticism, with terms like "utterly" and "allegedly" suggesting doubts about the effectiveness and safety of Tesla's automated features. Some customers questioned the accuracy and functionality of the driving algorithms.

The results of the topic modeling provide valuable insights into the overall customer experience with Tesla. The analysis reveals that while Tesla's innovative features and automated driving systems generate excitement, concerns about service delays, delivery issues, and extended wait times for repairs remain key sources of dissatisfaction. The presence of mixed sentiments across different topics suggests that while some aspects of Tesla's offerings meet customer expectations, others—particularly service reliability and logistical efficiency—need improvement.

The topic distributions and visualizations further emphasize the dominant themes within customer reviews, showcasing patterns of frustration, enthusiasm, and skepticism. These insights highlight areas where Tesla can enhance its service infrastructure and refine its vehicle technology to build greater consumer trust. By addressing the recurring concerns identified through this analysis, Tesla has the opportunity to improve customer satisfaction, strengthen brand loyalty, and reinforce its reputation as a leader in the electric vehicle market.

IV. DISCUSSION

The application of Latent Dirichlet Allocation (LDA) in customer review analysis has demonstrated significant potential in extracting meaningful insights from unstructured textual data. By leveraging LDA on Trustpilot reviews, this study successfully identified key topics relevant to Tesla customers, offering a deeper understanding of consumer sentiment and concerns. The results highlight the effectiveness of LDA in uncovering latent themes that traditional sentiment analysis or keyword extraction methods often fail to capture.

The identified topics reveal distinct areas of customer interest, including product performance, customer service, delivery experiences, software and autonomous features, and pricing. These findings provide businesses with actionable insights to enhance product quality, improve customer support, and address key areas of consumer dissatisfaction. For instance, while Tesla receives positive feedback on its technological innovations, recurring complaints about customer service and order fulfillment highlight areas requiring attention.

Despite the advantages, several challenges must be considered when applying LDA in customer review analysis. One major challenge is selecting the optimal number of topics, as different topic numbers can produce varying outcomes in terms of coherence and interpretability. The use of coherence scores and manual validation is essential to ensure meaningful topic extraction. Additionally, short and informal nature of customer reviews often makes it difficult to extract well-defined topics, necessitating advanced preprocessing techniques such as stop word removal, stemming, and TF-IDF weighting to enhance the quality of extracted topics.

Another limitation of LDA is its reliance on bag-of-words representation, which ignores word order and context. Future studies could explore hybrid approaches that combine LDA with word embeddings or deep learning-based topic modeling techniques for improved topic coherence and accuracy. Furthermore, real-world applications of LDA should consider domain-specific adaptations to refine topic interpretation and improve the relevance of insights for business decision-making.

In conclusion, this study demonstrates the feasibility of using LDA for customer review analysis, providing valuable insights into consumer sentiment and product feedback. While LDA effectively uncovers hidden themes within large text corpora, further improvements in preprocessing, parameter tuning, and hybrid modeling techniques can enhance its applicability and accuracy. By leveraging topic modeling, businesses can transform unstructured review data into actionable intelligence, ultimately improving customer satisfaction and business strategy.

IX. Conclusion

In this study, we looked at Tesla customer reviews from Trustpilot to find common themes in what people were saying. We found that customers often talked about things like battery life, range, charging, and customer service. Positive reviews focused on Tesla's innovation and driving experience, while negative reviews mentioned problems with battery reliability and service delays. These results show that while Tesla is strong in areas like product performance and innovation, there are still areas for improvement, such a customer service and product reliability. However, since we only looked at reviews from one platform, the study has some limits. In the future, we could look at reviews from other sites, use sentiment analysis to understand emotions in the reviews, and compare Tesla to its competitors to get a clearer picture. This kind of research can help Tesla improve its products and services, leading to better customer experience.

References

[1] Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177).

[2] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993-1022.

[3] Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1), 5228-5235.

[4] Wallach, H. M., Mimno, D. M., & McCallum, A. (2009). Rethinking LDA: Why priors matter. In *Advances in Neural Information Processing Systems* (pp. 1973-1981).

[5] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2), 1-135.

[6] Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4), 82-89.

[7] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391-407.

[8] Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems* (pp. 288-296).

[9] Wickham, H. (2016). rvest: Web scraping with R. *Journal of Statistical Software, 70*(1), 1-22.

[10] Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software, 25*(5), 1-54.

[11] Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media, Inc.

[12] Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

[13] Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach*. O'Reilly Media, Inc.

[14] Roberts, M. E., Stewart, B. M., & Tingley, D. (2016). Navigating the local modes of big data: The case of topic models. In *Data Science* (pp. 51-71).

[15] Wallach, H. M., Mimno, D. M., & McCallum, A. (2009). Rethinking LDA: Why priors matter. In *Advances in Neural Information Processing Systems* (pp. 1973-1981).

[16] Jockers, M. L. (2013). *Macroanalysis: Digital methods and literary history*. University of Illinois Press.

[17] Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software, 3*(30), 774.

[18] Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks, and incremental parsing.

[19] Griffiths, T. L., & Steyvers, M. (2004). Probabilistic topic models. Annual Review of Psychology, 55, 627-660.

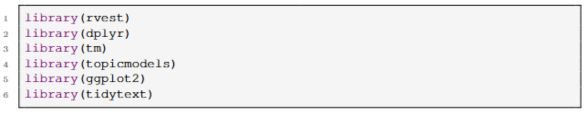
[20] Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84.

[21] Lafferty, J., & Blei, D. (2006). Correlated topic models. *In Advances in Neural Information Processing Systems* (pp. 147-154).

Appendix

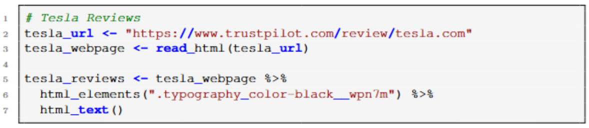
This appendix contains the R code used in the project for data preprocessing, topic modeling, performance evaluation, and visualization. The code serves as a reference for reproducing the results discussed in the report and demonstrates the methodologies applied.

Importing necessary libraries



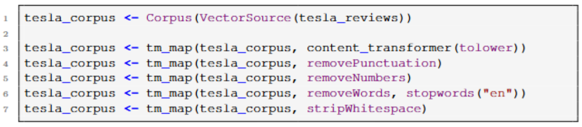
```r
library(rvest)
library(dplyr)
library(tm)
library(topicmodels)
library(ggplot2)
library(tidytext)
```

Web Scraping Tesla Reviews



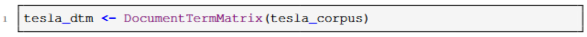
```r
# Tesla Reviews
tesla_url <- "https://www.trustpilot.com/review/tesla.com"
tesla_webpage <- read_html(tesla_url)

tesla_reviews <- tesla_webpage %>%
  html_elements(".typography_color-black__wpn7m") %>%
  html_text()
```

Text Preprocessing



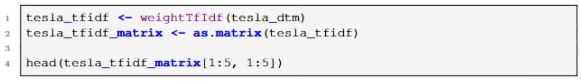
```r
tesla_corpus <- Corpus(VectorSource(tesla_reviews))

tesla_corpus <- tm_map(tesla_corpus, content_transformer(tolower))
tesla_corpus <- tm_map(tesla_corpus, removePunctuation)
tesla_corpus <- tm_map(tesla_corpus, removeNumbers)
tesla_corpus <- tm_map(tesla_corpus, removeWords, stopwords("en"))
tesla_corpus <- tm_map(tesla_corpus, stripWhitespace)
```

Creating a Document-Term Matrix (DTM)



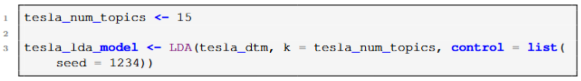
```r
tesla_dtm <- DocumentTermMatrix(tesla_corpus)
```

Applying TF-IDF Weighting



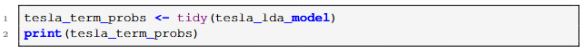
```r
tesla_tfidf <- weightTfIdf(tesla_dtm)
tesla_tfidf_matrix <- as.matrix(tesla_tfidf)

head(tesla_tfidf_matrix[1:5, 1:5])
```

Topic Modeling using LDA



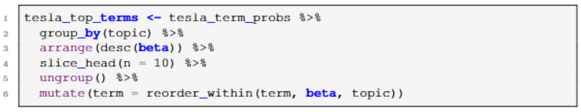
```r
tesla_num_topics <- 15

tesla_lda_model <- LDA(tesla_dtm, k = tesla_num_topics, control = list(
    seed = 1234))
```

Extracting Topic-Term Probabilities



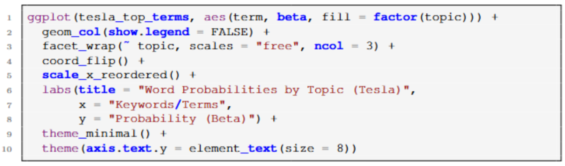
```r
tesla_term_probs <- tidy(tesla_lda_model)
print(tesla_term_probs)
```

Selecting Top Terms per Topic



```r
tesla_top_terms <- tesla_term_probs %>%
  group_by(topic) %>%
  arrange(desc(beta)) %>%
  slice_head(n = 10) %>%
  ungroup() %>%
  mutate(term = reorder_within(term, beta, topic))
```

Visualizing Topic Word Probabilities



```r
ggplot(tesla_top_terms, aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(˜ topic, scales = "free", ncol = 3) +
  coord_flip() +
  scale_x_reordered() +
  labs(title = "Word Probabilities by Topic (Tesla)",
       x = "Keywords/Terms",
       y = "Probability (Beta)") +
  theme_minimal() +
  theme(axis.text.y = element_text(size = 8))
```