**DEVELOPMENT OF A LDA-GSAPO ALGORITHM ENHANCED WITH TOPIC COHERENCE FOR SENTIMENT ANALYSIS**

**BY**

**OWOLABI DANIEL OLASUBOMI  
(21CG029915)**

**A PROJECT SUBMITTED TO THE DEPARTMENT OF COMPUTER AND INFORMATION SCIENCES, COLLEGE OF SCIENCE AND TECHNOLOGY, COVENANT UNIVERSITY OTA, OGUN STATE.**

**IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE BACHELOR OF SCIENCE (HONOURS) DEGREE IN COMPUTER SCIENCE.**

**NOVEMBER, 2024**

Acknowledgements

Firstly, I appreciate GOD for his unfailing help and sustenance from the start of my degree till my current position. I appreciate Mr. and Mrs. Owolabi for their love and financial support without which I would not have gotten here.

I also appreciate my project supervisor Mr. Dami Osufuye for his help throughout the project research phase.

Table of contents

**Content** **Pages**

[Acknowledgements i](#_Toc201039691)

[Table of contents ii](#_Toc201039692)

[List of figures vii](#_Toc201039693)

[List of tables viii](#_Toc201039694)

[Abbreviations ix](#_Toc201039695)

[ABSTRACT x](#_Toc201039696)

[Chapter one 11](#_Toc201039697)

[Introduction 11](#_Toc201039698)

[1.1. Background Information 11](#_Toc201039699)

[1.2. Statement of the Research Problem 13](#_Toc201039700)

[1.2.1. RESEARCH QUESTIONS 15](#_Toc201039701)

[1.3. Aim and Objectives of the Study 15](#_Toc201039702)

[1.4. Methodology 16](#_Toc201039703)

[1.5. Significance of the study 17](#_Toc201039704)

[1.6. Justification of Study 18](#_Toc201039705)

[1.7. Organization of the Write-up Project 18](#_Toc201039706)

[1.8. Operational Definition of Key Terms 19](#_Toc201039707)

[Chapter two 20](#_Toc201039708)

[Literature review 20](#_Toc201039709)

[2.1. Preamble 20](#_Toc201039710)

[2.2. Review of Related Methods 20](#_Toc201039711)

[2.2.1. Latent Dirichlet Allocation Algorithm 20](#_Toc201039712)

[2.2.2. Gibbs Sampling and Perplexity Optimization 21](#_Toc201039713)

[2.2.3. Topic Coherence Metrics 23](#_Toc201039714)

[2.2.4. Sentiment Analysis 23](#_Toc201039715)

[2.2.5. Web-Based Sentiment Analysis Systems 24](#_Toc201039716)

[2.3. Review of Related Methods 26](#_Toc201039717)

[2.3.1. Variants of LDA 26](#_Toc201039718)

[2.3.2. Topic Modeling with Coherence Metrics 27](#_Toc201039719)

[2.3.3. Sentiment Analysis Techniques 28](#_Toc201039720)

[Rule-based Techniques 28](#_Toc201039721)

[Challenges in sentiment analysis 29](#_Toc201039722)

[2.3.4. Natural Language Processing (NLP) 29](#_Toc201039723)

[Core Techniques in NLP 30](#_Toc201039724)

[NLP in Customer Satisfaction Analysis 31](#_Toc201039725)

[2.4. Review of Existing Systems 32](#_Toc201039726)

[2.4.1. LDA 32](#_Toc201039727)

[2.4.2. LDA-GSAPO 33](#_Toc201039728)

[2.4.3. Clustering-Based Joint Topic-Sentiment Modeling 34](#_Toc201039729)

[2.4.4. BERTopic 34](#_Toc201039730)

[2.4.5. SentiBERT 35](#_Toc201039731)

[2.4.6. Comparison of existing systems and tools 36](#_Toc201039732)

[2.4.7. Limitations of Existing Models 37](#_Toc201039733)

[2.5. Summary of Literature Review 38](#_Toc201039734)

[Chapter three 39](#_Toc201039735)

[System analysis and design 39](#_Toc201039736)

[3.1. Preamble 39](#_Toc201039737)

[3.2. Research Design 39](#_Toc201039738)

[3.3. Design of the Topic Coherence-Enhanced LDA-GSAPO Algorithm 40](#_Toc201039739)

[3.3.1. Data Collection 40](#_Toc201039740)

[3.3.2. Preprocessing 40](#_Toc201039741)

[3.3.3. Algorithm Enhancement 41](#_Toc201039742)

[3.3.4. Tools and Techniques 41](#_Toc201039743)

[3.3.5. Testing and Validation 41](#_Toc201039744)

[3.4. Benchmarking the Enhanced Algorithm 41](#_Toc201039745)

[3.4.1. Benchmark Models 41](#_Toc201039746)

[3.4.2. Benchmarking Metrics 41](#_Toc201039747)

[3.4.3. Experimental Setup 44](#_Toc201039748)

[3.5. Integration of the Algorithm in a Web-Based Platform 44](#_Toc201039749)

[3.5.1. Requirement Gathering 44](#_Toc201039750)

[3.5.2. System Design 45](#_Toc201039751)

[3.5.3. Implementation 46](#_Toc201039752)

[3.6. Usability Evaluation of the Web-Based System 46](#_Toc201039753)

[3.7. Exploratory Data Analysis 54](#_Toc201039754)

[3.7.1. Overview of the Dataset 54](#_Toc201039755)

[3.7.2. Data Quality and Missing Values 55](#_Toc201039756)

[3.7.3. Rating Score Distribution 56](#_Toc201039757)

[3.7.4. Word Cloud Visualization 56](#_Toc201039758)

[3.7.5. Top Words per Review Range 57](#_Toc201039759)

[3.7.6. Key Observations 58](#_Toc201039760)

[Chapter four 60](#_Toc201039761)

[System Implementation and Evaluation 60](#_Toc201039762)

[4.1. PREAMBLE 60](#_Toc201039763)

[4.2. SYSTEM REQUIREMENTS 60](#_Toc201039764)

[4.2.1. Hardware Requirements 60](#_Toc201039765)

[4.2.2. Software Requirements 61](#_Toc201039766)

[4.3. IMPLEMENTATION TOOLS 62](#_Toc201039767)

[4.3.1. Python 62](#_Toc201039768)

[4.3.2. JavaScript 62](#_Toc201039769)

[4.3.3. Visual Studio Code 62](#_Toc201039770)

[4.4. DEVELOPMENT METHODOLOGY 62](#_Toc201039771)

[4.5. EVALUATION OF THE SYSTEM 64](#_Toc201039772)

[4.5.1. Evaluation Criteria 64](#_Toc201039773)

[4.5.2. Evaluation Results 64](#_Toc201039774)

[4.5.3. Discussion 66](#_Toc201039775)

[4.6. Program Modules and Interfaces 66](#_Toc201039776)

[4.6.1. Initial Load Interface 67](#_Toc201039777)

[4.6.2. File Upload and Preprocessing Panel 67](#_Toc201039778)

[4.6.3. Results Display and Thematic Insights Panel 69](#_Toc201039779)

[Chapter FIVE 70](#_Toc201039780)

[SUMMARY, RECOMMENDATIONS AND CONCLUSION 70](#_Toc201039781)

[5.1. Summary 70](#_Toc201039782)

[5.2. Recommendations 71](#_Toc201039783)

[5.3. Conclusion 71](#_Toc201039784)

[References 73](#_Toc201039785)

List of figures

[Figure 2.1: Natural Language Processing Components Source: https://link.springer.com/article/10.1007/s11042-022-13428-4 29](#_Toc201012426)

[Figure 2.2: Theoretical foundation of LDA Source: https://www.markovml.com/blog/lda-topic-modelling 32](#_Toc201012427)

[Figure 2.3: Perplexity Formula Source: https://link.springer.com/chapter/10.1007/978-3-031-76528-5\_7 32](#_Toc201012428)

[Figure 2.4: Joint Topic Sentiment Framework Source: https://www.mdpi.com/2078-2489/15/4/200 33](#_Toc201012429)

[Figure 2.5: BERTopic Logo Source: https://www.maartengrootendorst.com/blog/bertopic/ 34](#_Toc201012430)

Figure 2.6: SentiBERT’s Structure Source: https://web.cs.ucla.edu/~kwchang/bibliography/yin2020sentibert/ 34

[Figure 3.1: E.R Diagram Source: https://lucid.app 43](#_Toc201012432)

[Figure 3.2: Use case Diagram Source: https://lucid.app 44](#_Toc201012433)

[Figure 3.3: Sequence Diagram Source: https://lucid.app 45](#_Toc201012434)

[Figure 3.4: Rating Score Distribution Chart 55](#_Toc201012435)

[Figure 3.5: Word Cloud Visualization 56](#_Toc201012436)

[Figure 4.1: Metric Test Results 64](#_Toc201012437)

[Figure 4.2: Homepage 66](#_Toc201012438)

[Figure 4.3: File Upload Interface 67](#_Toc201012439)

[Figure 4.4: Results Panel 68](#_Toc201012440)

List of tables

[Table 2.1: Comparison of existing systems and tools 35](#_Toc201012444)

[Table 3.1: Summary Table of Metrics 42](#_Toc201012445)

[Table 3.2: Survey table for Evaluating the Usability of the web app 46](#_Toc201012446)

[Table 3.3: 5 rows of sample data 54](#_Toc201012447)

[Table 4.1: Hardware Requirements Table 59](#_Toc201012448)

[Table 4.2: Software Requirements Table 60](#_Toc201012449)

Abbreviations

|  |  |
| --- | --- |
| LDA | Latent Dirichlet Allocation |
| NLP | Natural Language Processing |
| MCMC | Markov Chain Monte Carlo |
| GSAPO | Gibbs Sampling and Perplexity Optimization |
| SA | Semantic Analysis |
| NLTK | Natural Language Toolkit |
| BERT | Bidirectional Encoder Representations from Transformers |
| JST | Joint Sentiment-Topic Model |
| ABSA | Aspect Based Sentiment Analysis |
| LDA-GSAPO | Latent Dirichlet Allocation with Gibbs Sampling and Perplexity Optimization |

ABSTRACT

Sentiment analysis-based customer satisfaction has become more important for businesses aiming to enhance user experience and provide value-added services. Although Latent Dirichlet Allocation (LDA) is a well-liked topic modeling algorithm among the natural language processing community, it generally struggles to extract meaningful and sentiment-consistent topics from complicated customer feedback. To address these limitations, this article proposes an improved topic model Topic-Coherent LDA-GSAPO that integrates Gibbs Sampling and Perplexity Optimization, as well as topic coherence metrics (C\_V and C\_P), to improve semantic quality and topic interpretability. The model takes advantage of the strengths offered by Gibbs Sampling, which is an MCMC algorithm, for iterating to better topic distributions. We propose an optimization procedure to reduce perplexity in order to enhance model accuracy, with coherence scores as a secondary objective to ensure the generated topics are not just statistically valid but also interpretable and sentiment-aware. Empirical tests carried out on a large dataset of iPhone customer reviews prove that Topic-Coherent LDA-GSAPO performs better than standard LDA and other state-of-the-art models that employ clustering, in terms of topic coherence, perplexity, and thematic sentiment alignment. The proposed model offers a more realistic and context-dependent approach to thematic sentiment analysis and hence enables deeper consumer sentiment insight and more intelligent applications in feedback analysis, product development, and recommendation.

Chapter one

# Introduction

## Background Information

Topic coherence is a set of metrics for the quality assessment of the topics generated by the topic modeling algorithms, mainly aiming at computing the semantic similarities between words within a topic. Topic coherence is an important aspect of NLP for ensuring that the extracted topics from a text corpus are meaningful and interpretable(Hoyle et al., 2021). This is particularly true in applications such as customer satisfaction sentiment analysis, where clear, coherent topics allow companies to look at the key themes in the feedback accurately. If coherence does not exist in extracted topics, then such topics do not represent the latent themes in the data; this tends to yield insights that are either misleading or impossible to act upon. Thus, optimization for topic coherence in topic modeling results in an improved approach toward the extraction of actionable insights in customer feedback, which enables organizations to respond more aptly to clients' needs and preferences.

In the area of information and communication technology, NLP and machine learning are extensively used to extract insights from unstructured data sources, such as online reviews and feedback forms. Sentiment analysis is the subset of NLP that deals with identifying the sentiment of text, usually classified as positive, negative, or neutral. However understanding the themes or topics associated with these sentiments is important to gain a very good insight into customer satisfaction. LDA is perhaps the most popular topic modeling technique, and it assigns words in a text corpus to different topics based on their co-occurrence patterns (Chauhan & Shah, 2022). However, LDA isn’t effective at discovering general topics because it cannot generate topics on its own, especially when using short and diverse text sources such as customer reviews. This may limit the practical value of the analysis.

Although LDA gives a foundation for topic modeling, its normal implementation has some limitations, which are going to hurt the performance in real-world sentiment analysis: the standard approach to LDA, even when optimized for low Perplexity, a measure of the accuracy of prediction, may generate topics that are statistically sound but are not interpretable. Such a approach can return unhelpful topics that do not reflect what customers are saying, especially for customer satisfaction analysis where getting clear themes is important. Previous work has tried to overcome these limitations by borrowing techniques such as Gibbs Sampling to iteratively improve the accuracy of topic assignments. However, Gibbs Sampling alone doesn't guarantee coherence. Moreover, focusing solely on Perplexity as a measure of model quality can lead to overfitting and create topics that correspond to the training data but lack generalizability or interpretability in real applications. Our proposed solution is to enrich LDA with Gibbs Sampling, Perplexity optimization, and what is more important the core metric of topic coherence, addressing both the statistical and interpretive needs of sentiment analysis.

In this approach, Gibbs Sampling refines topic distribution by iteratively Sampling word assignments within each topic, aiding in achieving a better statistical fit(Dunn & Shultis, 2023). While, Perplexity optimization improves the model's predictive accuracy by reducing the noise in topic generation. Topic coherence serves as an evaluative metric to ensure that generated topics are both statistically robust and semantically meaningful(Rahimi et al., 2023). This topic coherence addition creates a balance, allowing the model to zero in on themes which are interpretable and relevant to customer sentiment.

Previous work in topic modeling has focused either on refining statistical measures, such as Perplexity, or on increasing coherence without considering the pragmatic limitations of applying these techniques to domain-specific applications, such as customer satisfaction. For instance, models that perform well in Perplexity may fail in interpretability, while coherence-driven approaches alone may not have any predictive capability. Combining these aspects, we fill the gap from the literature and develop a model that is predictive and interpretive at the same time, therefore increasing the usefulness of sentiment analysis for customer satisfaction. In summary, we are of the opinion that the enhancement of LDA with Gibbs Sampling, Perplexity optimization and a strong emphasis on topic coherence will result in improved model performance in customer satisfaction sentiment analysis(Chauhan & Shah, 2022) this approach which by focusing on coherence besides statistical accuracy promises actionable insights aligned with real customer sentiments, something which is strongly needed in the applications of sentiment analysis within ICT.

## Statement of the Research Problem

Increasing reliance on customer feedback in the improvement of products and services has resulted in a huge demand for accurate and interpretable sentiment analysis tools. Traditional topic modelling methods, such as Latent Dirichlet Allocation (LDA), have been widely used to discover latent topics behind textual data. However, these methods often give results that are not meaningful and actionable when applied to short and heterogeneous texts, such as consumer reviews. Specifically, the topics generated by standard Latent Dirichlet Allocation (LDA) are usually not coherent, which limits the organization's ability to understand the customers' sentiments correctly. This incoherence originates from LDAs built on statistical assumptions focusing on predictive accuracy (e.g., Perplexity reduction) rather than semantic clarity(Hoyle et al., 2021). The current enhancements to LDA, including Gibbs Sampling and Perplexity optimization, give better accuracy in topic assignment and less noise. However, they still can't deal with the problem of semantic coherence thoroughly.

Low Perplexity, while being a good statistical performance, does not necessarily mean the generation of human-interpretable topics. Though it uses the identification of cohesive themes on which actionable insights in customer satisfaction sentiment analysis are based, this disconnect weakens the effectiveness of traditional methodologies. Also, there is no framework that specifically integrates topic coherence as a necessary constituent within the modelling process; hence, this fuels the challenge of extracting meaningful insights from customer feedback.

One of the major gaps in knowledge and practice lies with the dilemma that existing models can't balance predictive accuracy with interpretability, specifically in customer sentiment analysis. A critical need is an enhanced approach, which simultaneously optimizes for coherence and accuracy, to make the generated topics both statistically sound and semantically meaningful. Otherwise, companies might otherwise misinterpret customer feedback, resulting in poor decision-making and decreased customer satisfaction. This problem has shown the requirement for developing a new enhancement to LDA by incorporating Gibbs Sampling, Perplexity optimization, and topic coherence as a key metric for topic model fine-tuning in customer sentiment analysis.

The motivation behind this project is to understand consumer needs and enhance the services they receive based on the rapid growth of user-generated content, including customer reviews and feedback. Sentiment analysis becomes very important in such a context for extracting insights in a meaningful way from textual data.

Nevertheless, effective sentiment analysis goes beyond merely categorizing feedback into positive, negative, or neutral; it requires understanding the underlying themes or subjects driving such sentiments. This would give a business a deeper understanding of how to address particular issues, improve its services, and ultimately build brand loyalty. Traditional topic modeling techniques, such as Latent Dirichlet Allocation (LDA), provide the means to uncover themes that are latent in text data. Such methods, although popular, shall finally produce topics that are hard to interpret, namely, datasets composed of short and diverse texts like customer reviews. Such incoherence detracts from practical utility, where businesses cannot draw useful insights from such generated topics. The problem is further worsened by relying on Perplexity as a model evaluation metric, which favors statistical accuracy at the expense of semantic interpretability. The paper is motivated by the need to close the gap in statistical accuracy and interpretability in topic modeling. This promises to enhance LDA with Gibbs Sampling and Perplexity optimization, incorporating topic coherence as a central evaluative metric to respond to this challenge.

This guarantees the mathematical validity and semantic meaning of the generated topics. Hence, this would be a proposed solution around this integration so that sentiment analysis is of high reliability to enable the business to take proper action regarding customer feedback. These efforts are driven by the belief that advanced topic modeling techniques can be transformational for customer satisfaction analysis organizational mechanisms through which a better and more precise understanding of customers can be attained and strategies honed. Introducing coherence metrics with advanced Sampling and optimization algorithms is one gigantic stride toward this objective; it makes sentiment analysis more robust and an important tool in the decision-making process.

### RESEARCH QUESTIONS

The study seeks to answer the following questions:

1. How can the LDA-GSAPO algorithm be improved with topic coherence metrics to generate better quality topics?

2. How good is the coherence-enhanced LDA-GSAPO algorithm compared to the standard LDA-GSAPO?

3. How could the improved LDA-GSAPO algorithm be applied into a web-based customer satisfaction sentiment analysis system?

4. How is the effectiveness and usability of the developed web-based system intended for sentiment analysis?

## Aim and Objectives of the Study

The aim of this project is to enhance the LDA-GSAPO algorithm with the topic coherence metrics to enhance its accuracy and interpretability of topics for effective customer satisfaction sentiment analysis. The above stated aim will be achieved with the following objectives;

1. To present the proposed topic coherence-enhanced LDA-GSAPO algorithm.
2. To evaluate the topic coherence-enhanced LDA-GSAPO algorithm.
3. To integrate the topic coherence-enhanced LDA-GSAPO algorithm into a web-based system for customer satisfaction sentiment analysis.
4. To evaluate the customer satisfaction sentiment analysis web-based system.

## Methodology

In pursuing the objective of developing the LDA-GSAPO algorithm with topic coherence measures to facilitate efficient customer satisfaction sentiment analysis, the research followed a structured approach that encompasses algorithm design, experimental testing, system development, and user evaluation.

The first objective was to introduce the topic coherence-enhanced LDA-GSAPO algorithm with a comprehensive review of the literature on prevailing topic-sentiment models and coherence metrics. Based on the shortcomings observed in conventional LDA and analogous models, the LDA-GSAPO model was chosen and optimized by introducing coherence assessment metrics in the form of contextual vector coherence and confirmation pairwise coherence (C\_V and C\_P) metrics. These measures were incorporated into the model for guiding the training process, thus strengthening the semantic consistency of topics generated during modeling. The algorithm was implemented in Python using libraries like Gensim and MALLET, with preprocessing done using NLTK and spaCy to provide effective tokenization and text normalization.

The second objective included the assessment of the coherence improved LDA-GSAPO model. This stage entailed conducting experiments with datasets of customer reviews. The improved algorithm was trained and tested on the datasets, and the performances were compared based on statistical coherence metrics (C\_V and C\_P) and perplexity metrics. Various parameters of the model, including the number of topics and sampling iterations, were finely tuned to examine their influence on performance. Comparative baselines were also established between the extended model and a baseline LDA model in order to measure gains in topic interpretability and sentiment alignment.

To achieve the third objective, an improved algorithm was subsequently incorporated in a complete web-based application. The application was set up to enable the users to upload review datasets, carry out analysis via the LDA-GSAPO engine, and present the resulting topics and corresponding sentiments in an understandable format. The backend was developed using Flask to manage routing and processing tasks, while HTML and CSS were both utilized in developing a responsive frontend interface. The overall application was hosted and tested on different platforms to ensure usability as well as accessibility.

Lastly, to assess the overall functionality of the system, a usability test was carried out with a systematic user feedback system. The users interacted with the web interface, uploaded datasets, and observed the resulting visualizations. They were then requested to evaluate different facets of the system, including the interpretability of outputs, navigability, response time, and general satisfaction. The feedback was systematically recorded through a Likert-scale questionnaire. Furthermore, backend measurements were taken to quantify system performance metrics like average processing time and how it performed with varying load levels of requests. Findings from this measurement were fed into ongoing system improvement to meet both technical specifications and user-centered objectives.

## Significance of the study

This study is very important in addressing the shortcomings of traditional topic modeling and sentiment analysis techniques by proposing an enhanced LDA framework with integrated Gibbs Sampling, Perplexity optimization, and topic coherence metrics. It contributes to the improvement in semantic coherence and interpretability of topics generated from customer feedback so that business organizations can extract actionable insights more effectively.

This addition of topic coherence ensures the topics are statistically sound and semantically meaningful, hence bridging predictive accuracy with interpretative power. It will find particular relevance in customer satisfaction sentiment analysis, where knowing just what the core themes behind a feedback are is important to drive informed decisions toward better customer experiences.  
  
This research will further advance in the development of a web-based platform using the enhanced algorithm, which provides an enterprise with a scalable and easy-to-use mechanism for instantly analyzing customer feedback. If properly put into operation, the system would empower companies with more resources to discover problems and understand their customers' trends thereby enabling them to enforce data-driven improvements on their products and services.  
  
This thus clears the way for developing more precise and practically applicable topic modeling toward sentiment analysis, and more strong uses of NLP and ml in a great number of applications. This paper also serves as the starting point for future research on applying coherence-based metrics with statistical optimization algorithms in other text analytic tasks for better interpretability and effectiveness. In other words, this research is of high importance because it addresses the prevalent shortcomings in topic modeling and suggests an actionable solution to enhance decision processes in customer satisfaction management, hence serving scholarly inquiry and practical implementations.

## Justification of Study

The dependence of business decisions on customer feedback has increased manifold, with organizations trying to gain insight into customer needs and better their service delivery in a bid to retain a competitive advantage. While large-scale unstructured textual feedback remains pretty daunting, there is increasing demand for analysis due to constraints by available tools and algorithms.

The evaluation of traditional topic modeling techniques relies much on perplexity, focusing on statistical accuracy rather than semantic coherence. This often results in mathematically valid but uninterpretable topics, which in large part restricts the practical application of such sentiment analysis tools. The second gap lies here: most of the existing sentiment analysis systems do not incorporate the function of coherence-enhanced topic modeling, hence a big gap in the delivery of actionable insights remains.

This paper justifies itself by addressing these challenges through the development of an enhanced LDA-GSAPO algorithm. Through the incorporation of topic coherence metrics, this study ensures that generated topics are statistically sound and semantically meaningful. Moreover, being integrated into a web-based platform, this provides businesses with a scalable and user-friendly tool to analyze customer satisfaction in an efficient manner. The system fills the gap between theoretical developments and real-world applications, thus empowering organizations with more insight to drive business decisions.

## Organization of the Write-up Project

This project is structured into five chapters. Chapter One introduces the research background, problem statement, objectives, scope, and justification. Chapter Two reviews existing literature on LDA, topic coherence, sentiment analysis, and web-based systems. It identifies research gaps that the study addresses. Chapter Three outlines the research design and methodology for algorithm enhancement, system implementation, and usability evaluation. Chapter Four presents the results obtained using the enhanced LDA-GSAPO algorithm and discusses the performance and usability of the web-based system. Chapter Five concludes the research by summarizing findings, drawing conclusions, and offering recommendations for future undertakings.

## Operational Definition of Key Terms

1. **Latent Dirichlet Allocation (LDA)**: A statistical model used to discover abstract topics in a collection of documents.
2. **Gibbs Sampling**: A Markov Chain Monte Carlo algorithm used to approximate distributions in LDA.
3. **Perplexity**: A measure of how well a probability model predicts unseen data.
4. **Topic Coherence**: A metric used to assess the interpretability and relevance of topics generated by topic modeling algorithms.
5. **Sentiment Analysis**: The process of identifying and classifying sentiments expressed in textual data.
6. **Web-Based System**: An application accessible through web browsers, designed for performing specific tasks.

Chapter two

# Literature review

## Preamble

This chapter discusses important concepts, methodologies, and established frameworks relevant to the enhancement of Latent Dirichlet Allocation (LDA) is a probabilistic approach to topic modeling for discovering underlying topics in a corpus of textual data. The algorithm models each document as a mixture of different topics, where each topic is a distribution over words(Madzík & Falát, 2022). Since it was first proposed by Blei, ng, and Jordan back in 2003, one of the most popular tools for NLP has been LDA, as it could cope with large data size for finding latent patterns efficiently.

## Review of Related Methods

The concepts explored in this section serve as the foundation for the development of the enhanced LDA-GSAPO algorithm and its application in sentiment analysis. They form the building blocks for understanding how topic modeling can be enhanced with coherence metrics to improve the interpretability and usability of generated topics in customer satisfaction analysis.

### Latent Dirichlet Allocation Algorithm

In LDA, it is assumed that a document is generated from some hidden probabilistic process. It assigns a proportion of topics to each document, and each topic contains a probability distribution over words (Mardones-Segovia et al., n.d.). Through co-occurrence across documents, LDA infers these hidden distributions and clusters words into topics, associating documents with them. For example, it can be applied to a dataset of customer reviews to discover topics like "delivery speed," "product quality," or "customer service."(Yu & Xiang, 2023)

LDA has been applied in many applications, such as opinion mining, text summarization, and customer feedback analysis. However, because Perplexity is used as a performance metric, it usually suffers from the generation of semantically incoherent topics. Perplexity measures the predictive power of the model, but not necessarily whether topics are going to be interpretable by humans. Consequently, LDA may output topics which are statistically correct but semantically nonsensical and thus much less helpful in practice.

Those shortcomings of LDA are handled by modifications to the algorithm to include Gibbs Sampling and topic coherence metrics. On the other hand, Gibbs Sampling is an MCMC technique; it is an underlying refinement to the assignment of topics based on repeated Sampling of a word distribution in order to enhance the accuracy of modeling(Tekin, 2024a). The metrics of the topic coherence denote the similarity level of meaning among words that reside in a topic and are guaranteed to be interpretable while being created.

LDA is a very flexible and extensible tool for the analysis of unstructured data(Huang et al., 2021). Combining it with advanced techniques of topic coherence and Perplexity optimization increases the effectiveness of LDA and applications in which clear and actionable theme understanding is necessary, as in customer satisfaction sentiment analysis.

**Mathematical Foundation**:

For a document *d*:

* Each word w is assigned a topic *z*, drawn from a Dirichlet distribution.
* Word probabilities are sampled based on the topic-word distribution βk.
* Topics are assigned using θd ​, the document-topic distribution.
* These steps repeat iteratively, refining the assignments to improve model accuracy.

### Gibbs Sampling and Perplexity Optimization

Gibbs Sampling and optimization of Perplexity Gibbs Sampling is a Markov Chain Monte Carlo (MCMC) approach(Dunn & Shultis, 2023), an important method to perform probabilistic inference over high-dimensional models, such as the Latent Dirichlet Allocation (LDA), by iteratively Sampling the value of each variable conditional on the current states of all the other variables.  
  
Gibbs Sampling improves the topic assignments for words in a document within the framework of LDA by iteratively selecting a word's topic conditioned on the conditional likelihood of the word's association with each topic. Approximating the posterior distribution over topics through an iterative process makes topic modelling more precise.

But one way to test and improve the predicting accuracy of such probabilistic models is by Perplexity optimization. It measures a model's generalizability to data it has never seen before; the lower the Perplexity, the stronger the generalization (Cuevas et al., 2024). So by optimizing Perplexity, a model can enhance generalization over new data, and resulting topics are always statistically sound. Together, Gibbs Sampling and Perplexity optimization address key challenges in topic modelling.

Gibbs Sampling enables exact estimation of topic distributions via iterative refinement; Perplexity optimization is based on evaluating and updating the model's parameters to raise predictive accuracy (Karras et al., 2022a). Nevertheless, the semantic coherence and interpretability of the generated subjects cannot be ensured by these two characteristics alone.

For instance, a low-Perplexity model could nevertheless provide subjects that are too complex for humans to comprehend, particularly when used with short and sparse datasets like customer reviews. To overcome this limitation, contemporary approaches integrate topic coherence measurements with Perplexity optimization and Gibbs Sampling (Karras et al., 2022b). In order to help the model produce statistically sound and interpretable topics, subject coherence assesses the semantic relatedness of terms in a subject.

In particular, the integrated approach becomes very useful in customer satisfaction sentiment analysis applications, whereby actionable insights depend on discovering distinct and coherent themes in the feedback data. This will be important in refining the Sampling process and model evaluation metrics through Gibbs Sampling and Perplexity optimization for improving LDA (Tekin, 2024a). Combined with methods focusing on coherence, they form the core for topic modelling techniques, striking a balance between predictive accuracy and interpretability by humans. Thus, they are central instruments for unstructured text data analysis.

### Topic Coherence Metrics

Topic coherence is a measure developed to quantify the comprehensiveness and semantic relatedness of the different topics that come from topic-modelling techniques, such as LDA (Rahimi et al., 2023). This refers to how well the language used for a particular topic fits or aligns with what people think is actually important.

Unlike Perplexity, which measures the statistical probability of the model's predictions, topic coherence concerns whether the topics produced make sense to human readers. In other words, topic coherence is critical in the practical application of topic modeling.

The basic premise of topic coherence is that words associated with a coherent topic tend to co-occur in documents and exhibit semantic interrelations. For example, if the topic is "customer service," one would anticipate that the terms "support", "response" and "feedback" would arise frequently together, making a significant cluster (Lo et al., 2021). When word co-occurrence analysis is performed on a reference corpus or the dataset, these interrelations are approximated by metrics like pointwise mutual information or normalized pointwise mutual information.

**Types of topic coherence metrics**

1. C\_V: a combination of a sliding window with pairwise word similarities often the most robust for general applications.
2. U\_MASS measures coherence by analyzing the probabilities of word co-occurrences in a given corpus but is sensitive to the dataset's quality (Campagnolo et al., 2022).
3. C\_UCI: relying on external corpora for pairwise comparisons of words, this approach is particularly effective for domain-specific applications.
4. C\_NPMI: normalizes PMI scores, which are suitable for cross-domain topic evaluation.
5. C\_P: serves as a **complementary metric** that reflects conditional probability-based coherence

### Sentiment Analysis

Sentiment analysis is the computational attempt at determining the emotional tone behind any piece of text. This involves classifying text into a few pre-defined classes: positive, negative, and neutral, based on the emotion or opinion expressed in it. The importance of sentiment analysis has increased drastically for companies looking to analyze customer feedback in real time to draw actionable insights from social media posts, product reviews, and surveys(Wankhade et al., 2022).

Within the framework of this research, sentiment analysis is synthesized with topic modeling to yield a holistic view of customer satisfaction. Merging topic modeling, which identifies themes in feedback, with sentiment analysis, which measures the emotional tone of such feedback, allows organizations to understand not only the topics being discussed but also the sentiment associated with each topic (Birjali et al., 2021a). For example, sentiment analysis may show that, though a topic such as "customer service" is widely discussed, the overall sentiment towards it is largely negative, indicating an area for improvement.

**Approaches**:

* **Lexicon-Based**: Uses predefined word lists (e.g., VADER) to assign sentiment scores.
* **Machine Learning**: Employs classifiers like Naïve Bayes, SVM, or Logistic Regression trained on labeled datasets (Hamed et al., 2020).
* **Deep Learning**: Utilizes models like CNNs, RNNs, or transformers (e.g., BERT) for contextual sentiment understanding.

**Relevance**:  
Sentiment analysis is critical for understanding customer satisfaction, revealing underlying emotions tied to specific topics or themes.

### Web-Based Sentiment Analysis Systems

Web-based sentiment analysis systems have been built to give organizations a chance to make sense of customer feedback in a meaningful way. Such systems will enable the processing and analysis of large volumes of customer data be it product reviews, social media comments, or even survey responses in real time. Organizations can derive insights in action from this analysis, and these insights help drive service, product, or customer experience improvements. Some of the important features of these systems include being highly scalable, hence capable of handling large volumes of data without losing performance integrity a feature that makes them usable by firms of any size. One of the most important features of web-based systems is accessibility. They are hosted on cloud platforms, so they can be accessed from any device with an internet connection (Rodríguez-Ibánez et al., 2023).

This frees users from any hardware or software configurations on their personal devices making it much easier, and quite a lot cheaper. Cloud-based systems will also enable real-time data analysis; this helps the organization act in response to customer sentiment, emerging trends, and issues without delay. For example, a customer service team can monitor, in real time, feedback regarding the products, enabling the team to make quick improvements and address complaints. When the integration of topic modeling with sentiment analysis is done in a web-based platform, it does much more than understand the emotional tone of the customer feedback (Kumar et al., 2022); it can group all feedbacks automatically into distinct topics like product quality, shipping, or customer service and classify each piece of feedback into positive, negative, or neutral sentiments.

It allows businesses to understand the specific aspects of their products or services customers are talking about and the emotions associated with those aspects. For example, a company might find that a common theme, such as "fast delivery," is associated with both positive and negative emotions, thus helping it to improve its delivery service. To develop an integrated system, a lot of consideration must be placed on system design(Birjali et al., 2021b). One of the important concerns is scalability because as time goes on, organizations can experience a huge increase in customer feedback. The system should efficiently manage large datasets to provide quick analysis, no matter how large the data may grow. Another aspect that has to be taken into consideration is reliability, which indicates that the system has to be stable under heavy demand.

In addition, the system should be user-friendly enough for non-technical users to easily interact with the platform; this includes easy data upload mechanisms, clear visualizations of analysis results, and easily interpretable sentiment and topic reports (Nandwani & Verma, 2021).

Such a framework, where topic modeling and sentiment analysis come together, offers a twofold benefit to the organizations: business can now gather qualitative insightthat is, what exactly are the topics of conversation among customersand quantitative assessments regarding the sentiment tied to those topics. It is from such synergy that actionable feedback, usable in decision-making processes regarding customer service improvements, refinement of product characteristics, or modification of marketing tactics, becomes viable (Xu et al., 2022).

## Review of Related Methods

A range of approaches has been introduced in the past for the purpose of improving extraction of latent themes and sentiment structures from textual data. The approaches vary in the way document semantics are encoded, sentiment guidance is addressed, and coherence is optimized. In this section, we'd be reviewing of some of the most impactful topic modeling and sentiment-aware approaches, outlining their key mechanisms and their contributions toward advancing interpretability, accuracy, and scalability for thematic sentiment analysis.

### Variants of LDA

Over the years, several variants of the Latent Dirichlet Allocation (LDA) model have been developed to improve its performance in various applications. Standard LDA assumes topics are hidden, and documents are mixtures of these topics; still, it faces its own limitations when dealing with supervised learning or specific tasks where labeled data is available. One of the variants of the basic LDA is Supervised LDA (sLDA), which incorporates labels in the LDA process. In supervised latent Dirichlet allocation, the model is not only able to identify topics but also classifies the documents into existing classifications based on those identified topics (Stein et al., 2025). This capacity is especially useful in scenarios when documents are associated with certain categories (like customer reviews regarding various categories of products or services) and where the model has to simultaneously uncover topics and make predictions over these classifications.

Another important variant is LDA-GS, which introduced Gibbs Sampling into the LDA framework in order to refine the topic-word assignments by improving the word-topic mapping. Gibbs Sampling helps to iterate the word-topic assignment according to the probability distribution, so as to have better results in topic assignments than the traditional methods (Tekin, 2024b). Especially for datasets with large volumes of text data, LDA-GS is very important and useful to further refine the topic distribution when there is ambiguity in the initial topic assignments.

LDA-GSAPO is an extension of the LDA-GS that incorporates Perplexity Optimization (PO) to improve the generalization capability of the model on new, unseen data. The basic intuition behind Perplexity Optimization is to tune the model parameters such that overfitting is reduced and optimum results are ensured for both the training and test sets . However, despite such improvement, LDA-GSAPO continues to suffer from the problem of uninterpretable topic generation. This is because topics generated by LDA and its variants usually require topic coherence metrics for easy interpretation.

Directly addressing this limitation, the adoption of topic coherence metrics ensures that the resultant topics are relevant and meaningful. Metrics such as C\_V, NPMI, and UMass assess the degree to which the words in a topic are semantically related and also how interpretable the topics are to human users (Campagnolo et al., 2022). These metrics helped in improving the output of LDA models and enhance their interpretability, an essential aspect in practice applications like customer satisfaction analysis (Rahimi et al., 2023).

### Topic Modeling with Coherence Metrics

The application of coherence metrics in topic modeling has been very popular in the past couple of years, mostly due to improving the interpretability of topics derived from LDA. Previous studies have suggested the use of the combination of topic coherence measures, such as C\_V, using sliding windows of words and pairwise word similarities, NPMI, normalizing Pointwise Mutual Information, and UMass, estimating internal word co-occurrence probabilities, in order to let Latent Dirichlet Allocation (LDA) generate topics that are more humanly interpretable (Koltcov et al., 2024). These coherence measures ensure that the resulting topics are not only statistically sound but also semantically meaningful, a quality vital in practical applications like customer feedback analysis.

This may be furthered with the integration of coherence metrics into LDA in order for them to provide a way to iteratively refine topics. With C\_V, for example, it may suggest that the model form topics with words that frequently co-occur in similar contexts, thus raising their semantic similarity (Lo et al., 2021). This way, companies are going to get more actionable insights from the generated topics, which can be tied directly to the themes in customer feedback, such as "product quality," "customer service," or "delivery speed."

Integration of topic modeling with coherence metrics significantly improves the quality of insights derived from customer feedback. Generation of statistically valid and interpretable topics enables an organization to make more informed decisions on customer sentiment; this is quite vital in the analysis of customer satisfaction, where actionable insights are needed to improve the products or services (Campagnolo et al., 2022).

### Sentiment Analysis Techniques

Sentiment analysis is a field that has advanced considerably over the years, from traditional machine learning approaches like Naïve Bayes and Support Vector Machines (SVM) to modern deep learning techniques such as transformers (e.g., BERT). The traditional approaches in machine learning rely normally on manual feature engineering to classify text into positive, negative, or neutral sentiments. These techniques, while useful in handling smaller datasets or accomplishing simpler tasks, become quite hopeless when dealing with complex linguistic features like sarcasm, context, and subtle shifts in sentiment (Hamed et al., 2020).

Contemporary deep learning methodologies, as exemplified by BERT (Bidirectional Encoder Representations from Transformers), have largely improved the efficiency of sentiment analysis. BERT and other transformer-based models apply contextual embeddings, which are representations of words in relation to their context. This ability is particularly useful in doing sentiment analysis because it makes it more capable of perceiving nuances in customer feedback, such as even those instances imbued with sarcasm or also ambivalent sentiments within a single review. While such methodologies are satisfactory in performing the task of sentiment classification, they often remain incapable of incorporating an understanding of the core topics discussed in the feedback (Pandya & Mehta, 2020). On the other hand, topic modeling does a great job of capturing the underlying themes in the data, but it doesn't do sentiment classification. Integration of topic modeling and sentiment analysis allows companies to draw more insightful information from customer feedback (Sharma et al., 2020). With this integration, organizations will not only be able to identify the topics that consumers are discussing for instance, "delivery problems" or "product quality" but also the sentiment behind each topic, which may be positive, negative, or neutral, hence allowing more accurate decision-making.

Rule-based Techniques

The earliest sentiment analysis systems depended solely on rule-based approaches using a predefined lexicon, positively and negatively scored the presence of words in that lexicon to classify the text's sentiment. Tools using this approach include the lexicon-based VADER and TextBlob, whose models score the sentiment of each word and sum for the overall sentiment of a given text (Sharma et al., 2020). The results from rule-based techniques are interpretable and effective with simple texts but may become less necessary with increased text complexity, sarcasm, and emerging vocabulary.

Challenges in sentiment analysis

Despite advancements, sentiment analysis faces several challenges;

1. **Sarcasm and Irony**: models, which are hard to classify sentiments in sarcastic texts.
2. **Domain dependence**: models trained on one domain usually perform poorly on another (e.g., product vs. Movie reviews).
3. **Contextual understanding:** sentiment recognition in complex sentences or with ambiguous terms still remains a problem.
4. **Short texts:** sentiment analysis of short messages like tweets or SMS messages is difficult because of the lack of context.(Raghunathan & Saravanakumar, 2023)

### Natural Language Processing (NLP)

NLP stands for natural language processing, a subfield of the artificial intelligence field committed to providing machines with the capability to understand, interpret, and generate human language. This academic discipline connects human communication and computational models to enable the analysis and transformation of text and speech data into something meaningful in human life. NLP combines linguistics, computer science, and machine learning theories to process unstructured information, essentially text information (Patwardhan et al., 2023).

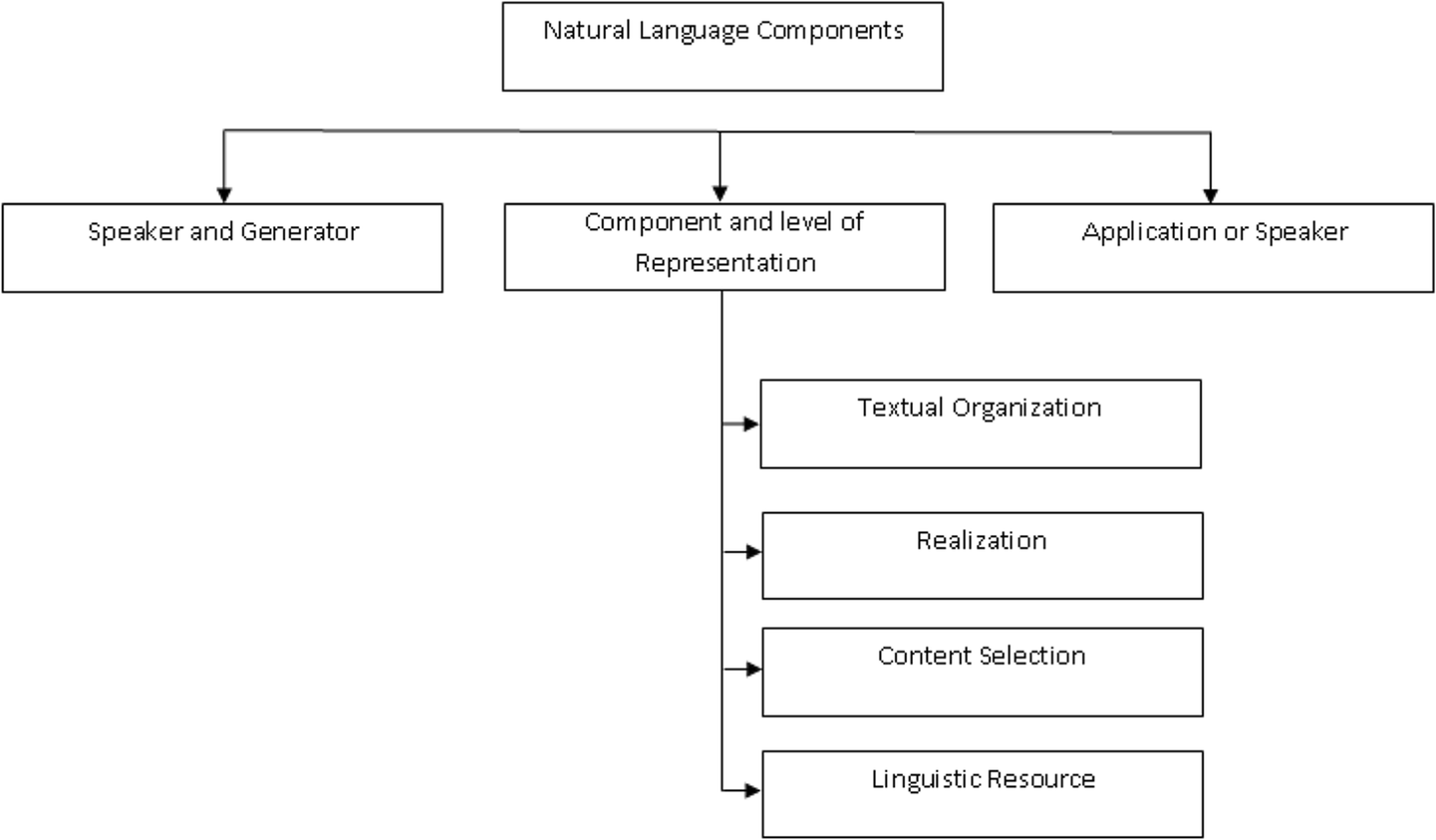
  
With the help of NLP approaches, a wide range of applications can be developed in chatbots, sentiment analysis, information retrieval, machine translation, and text summarization. For instance, customer satisfaction sentiment analysis is one of those examples of NLP procedures for checking customer textual feedback to detect the sentiment, determine what's really important, and hence, provide useful insights to improve their offerings, along with changing the perception that your clients have towards the businesses(Pais et al., 2022).

Figure 2.1: Natural Language Processing Components   
Source: https://link.springer.com/article/10.1007/s11042-022-13428-4

Core Techniques in NLP

The following are various core techniques employed in NLP:

1. **Text preprocessing:** this includes cleaning and normalizing text data by tokenization, stop word removal, stemming, and lemmatization.
2. **Word representations:** these are some of how text data can be represented in numerical formats, such as using bow, TF-IDF, and word embeddings like word2vec and glove.
3. **Named entity recognition:** the technique for extracting specific entities from text with names, dates, or organizations.
4. **Sentiment analysis:** this includes the identification of the emotional tone of the text, which is generally classified into positive, negative, or neutral (Hamed et al., 2020).
5. **Topic modeling:** identifying latent topics or themes in a corpus of text data using methods such as Latent Dirichlet Allocation-LDA.

NLP in Customer Satisfaction Analysis

applications of natural language processing are useful when dealing with massive data on textual feedback from different sources of reviews, polls, and opinions on social media regarding customer satisfaction studies. In particular, sentiment analysis helps determine customer emotions, while topic modelling identifies the most common themes or problems. These insights allow the addressing of customer concerns and better satisfaction levels(Fanni et al., 2023).

Progress and Challenges recent developments in NLP, particularly with transformer-based models like BERT (bidirectional encoder representations from transformers), have considerably pushed the performance envelope, making the tasks much more contextual and sensitive to semantic subtleties contained in the text (Mashaabi et al., 2022). These models better capture the subtleties of meaning, and sentiment classification and topic modelling benefit from the improvement. There are, however, problems:

1. The puzzles in our knowledge of language: understanding sarcasm, idioms, and jargon. Brief texts: sentiment and theme analysis in short or disconnected feedback, like tweets or text messages.
2. Computational complexity: complex models such as transformers take a lot of resources.
3. Pertinence to this research: this study uses natural language processing (NLP) approaches to evaluate consumer feedback data.

Topic modelling analyses sentiment and finds hidden themes to gauge customer opinions (Edwards, 2021). The study's findings will contribute to developing more dependable NLP-based systems that use advanced algorithms, such as enhanced LDA and topic coherence measures, to analyze customer satisfaction. Such solutions will provide businesses with clearer and more useful insights by bridging the gap between organizational offerings and client expectations.

## Review of Existing Systems

Thematic sentiment analysis places more emphasis on identifying concealed themes or topics in text data and connecting them with corresponding sentiments. Unlike typical sentiment analysis, which labels entire documents into pre-defined sentiment categories, thematic sentiment analysis yields a more profound insight with the underlying emphasis laid on the specific topic of the sentiment. It is mostly effective in the analysis of customer feedback, social media buzz, and big feedback systems. In the past several years, certain sophisticated models have been developed to enhance this process which will be discussed in the following sections.

### LDA

One of the fundamental models applied in thematic sentiment analysis is Latent Dirichlet Allocation (LDA). LDA extracts hidden thematic structure within large volumes of text corpora by representing every document as a probabilistic combination of topics, where each topic is characterized as a distribution over vocabulary. The unsupervised method is powerful and has been extensively applied in review mining and customer feedback analysis. However, LDA lacks the inherent capacity to connect sentiments with topics, and hence it is not much helpful in applications wherein the sentiment orientation in uncovered topics is crucial. So, although it offers interpretability on the basis of topic distribution, it is not sufficient in situations where sentiment detection is needed simultaneously.

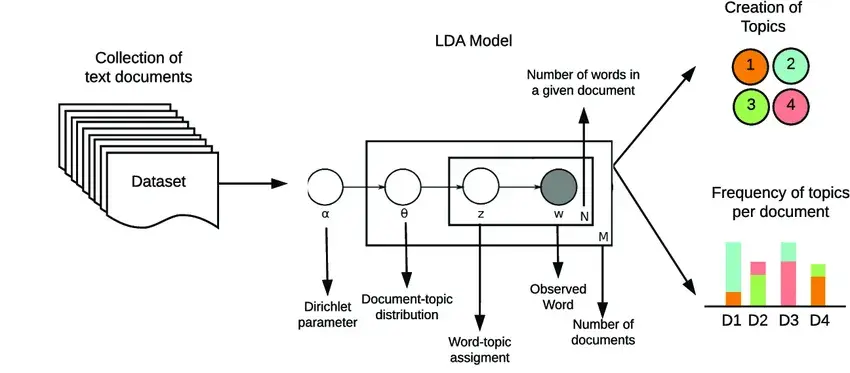


Figure 2.2: Theoretical foundation of LDA  
Source: https://www.markovml.com/blog/lda-topic-modelling

### LDA-GSAPO

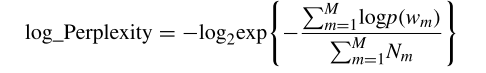
To address the limitations of traditional LDA in sentiment analysis contexts, **LDA-GSAPO** (Latent Dirichlet Allocation with Gibbs Sampling and Adaptive Particle Optimization) introduces enhancements that align topic modeling with sentiment classification. This model incorporates a hybrid optimization framework that combines Gibbs sampling with Adaptive Particle Optimization to refine both topic and sentiment assignments more effectively. The integration of sentiment supervision allows the model to disentangle sentiment-bearing topics from neutral content, resulting in more coherent and sentiment-aware topic representations. LDA-GSAPO demonstrates improved performance in domains where understanding both what users are talking about and how they feel about it is essential such as product review analysis and thematic social media monitoring.

Figure 2.3: Perplexity Formula  
Source: https://link.springer.com/chapter/10.1007/978-3-031-76528-5\_7

### Clustering-Based Joint Topic-Sentiment Modeling

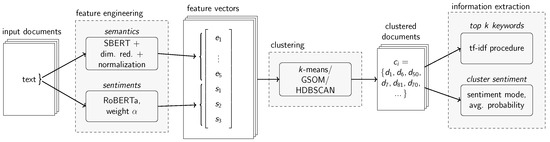
Clustering-Based Joint Topic-Sentiment Modeling merges clustering inferred by neural networks with sentiment classifiers for jointly learning latent sentiments and topics from unstructured text (Hanny & Resch, 2024). It is very valuable while discovering semantically coherent sentiment clusters in short, informal texts like tweets and review excerpts. This approach prioritizes sentiment homogeneity within the clusters as well as topic separability for much better interpretability.

Figure 2.4: Joint Topic Sentiment Framework  
Source: https://www.mdpi.com/2078-2489/15/4/200

### BERTopic

BERTopic integrates context embeddings from transformer-based models (e.g., BERT) with class-based TF-IDF approaches and dimensionality reduction algorithms to identify high-quality topics. The model is helpful in conducting thematic sentiment analysis in the sense that it enables interpretable topic clusters to be created and facilitates real-time visualizations of topic trends over time periods (Mishra, 2024). Its modular design enables the inclusion of sentiment scoring as a post-topic extraction step, thereby delivering dual-layer insights: thematic content and emotional tone.



Figure 2.5: BERTopic Logo  
 Source: https://www.maartengrootendorst.com/blog/bertopic/

### SentiBERT

This builds upon the original BERT architecture by including sentiment sensitivity within the attention mechanisms directly. It aims for the ability to identify subtle emotional polarity related to linguistic organization and has demonstrated excellent improvements in sentiment classification scores (Mewada et al., 2023). Not being strictly a topic model, its transformer layers centered around sentiment do make it a valuable component when used hand-in-hand with thematic extraction paradigms.

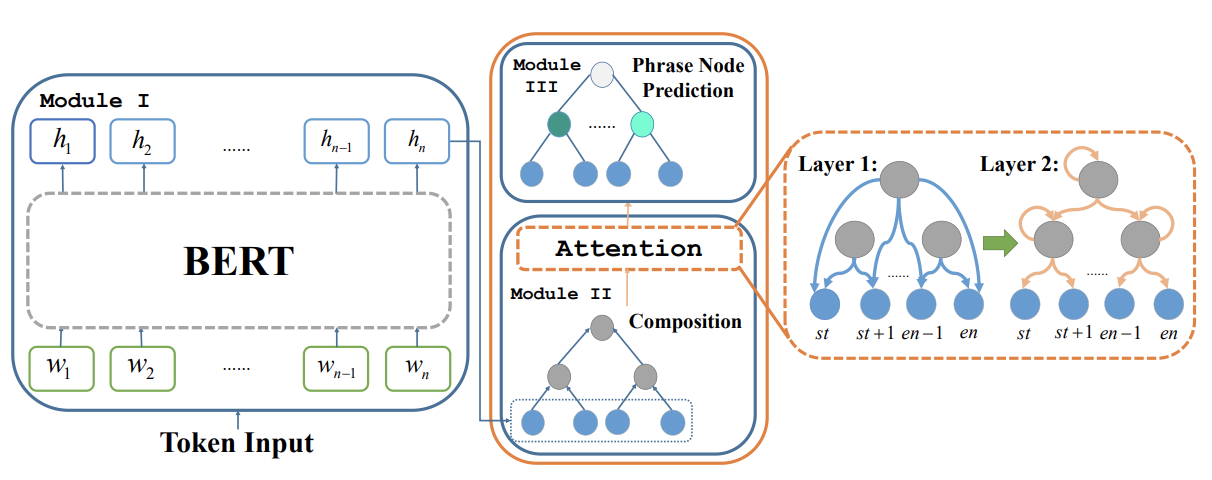


Figure 2.6: SentiBERT’s Structure  
Source: <https://web.cs.ucla.edu/~kwchang/bibliography/yin2020sentibert/>

### Comparison of existing systems and tools

Table 2.1: Comparison of existing systems and tools

|  |  |  |  |
| --- | --- | --- | --- |
| S/N | Tool/system | Strengths | Limitations |
| 1 | LDA | Widely used and well-understood, Captures latent topics from large text corpora, Unsupervised and scalable. | Often produces incoherent topics, struggles with sentiment-heavy or short texts, high perplexity can reduce interpretability. |
| 2 | LDA-GSAPO | Enhances LDA with Gibbs Sampling for better topic-word distribution. | Still lacks semantic evaluation (i.e., topic coherence), doesn't directly optimize for human interpretability, computationally more expensive than basic LDA. |
| 3 | Clustering-Based Joint Topic-Sentiment Modeling | Jointly models topic and sentiment; high coherence in clustered themes; effective on social media and short texts. | Requires careful tuning; performance may degrade on longer or more structured documents. |
| 4 | BERTopic | |  | | --- | |  |  |  | | --- | | Leverages transformer-based embeddings; produces highly coherent, human-readable topics; supports dynamic topic tracking. | | Not inherently sentiment-aware; sentiment analysis must be added externally; resource-intensive. |
| 5 | SentiBERT | |  | | --- | |  |  |  | | --- | | Integrates sentiment into the transformer architecture; strong performance on sentiment classification benchmarks; captures compositional semantics. | | Lacks built-in topic modeling; black-box nature limits interpretability; not optimized for thematic clustering. |

### Limitations of Existing Models

Despite the progress achieved, these models face considerable challenges:

* Topic Coherence:

While models like BERTopic and clustering-based approaches produce topics based on vector space proximity, they often yield semantically overlapping or incoherent clusters, especially when relying solely on embeddings. This weakens interpretability and may result in thematically ambiguous results.

The LDA-GSAPO model directly integrates topic coherence metrics such as C\_V and C\_P during training and optimization. These metrics evaluate the logical consistency and interpretability of a set of topic words (Rahimi et al., 2023), by incorporating these into the Gibbs Sampling and model selection process, it ensures that each topic is semantically distinct, improving clarity and human interpretability.

* Brief Textual Problems:

Models like Clustering-Based Joint Topic-Sentiment Modeling and SentiBERT often struggle with short or informal text due to sparse linguistic features and limited semantic context. This leads to weak or noisy theme detection and sentiment labeling.

The LDA-GSAPO model employs a robust preprocessing pipeline (lemmatization, stopword removal, bigram/trigram modeling) that amplifies semantic cues within short texts. It also uses perplexity optimization, which helps select a topic distribution that generalizes well even in low-data contexts making it more effective for datasets like tweets or brief product reviews.

* Interpretability and Explainability:

SentiBERT and other transformer-based models are highly accurate but lack transparency. Their black-box architecture makes it difficult to trace how themes and sentiments are linked, making it hard to explain outcomes to non-technical users (Mewada et al., 2023). While the LDA-GSAPO model is based on probabilistic topic modeling (LDA), which offers inherent transparency. Each topic is represented as a probability distribution over words, and each document as a mixture of these topics. This structure is easy to visualize, interpret, and explain, offering direct insight into how and why a sentiment was linked to a specific theme.

## Summary of Literature Review

This literature review has explored the evolution of the techniques and tools of customer satisfaction sentiment analysis, concerning particular topic modelling and sentiment analysis. Traditional LDA and its variants are a basis for recognizing latent themes within textual data, yet their semantic coherence and good performance on short texts usually remain low (Chauhan & Shah, 2022). On the other hand, neural topic models and sentiment-aware frameworks represent strong alternatives to it; these utilize deep learning for better contextual understanding in capturing topic relevance.

Optimization techniques, such as Gibbs Sampling and Perplexity optimization, fine-tune the performance of LDA (Ling et al., 2025). Meanwhile, metrics for topic coherence tackle the problems with interpretability, ensuring actionable insights. Gensim and Scikit-Learn provide topic modelling but require much more customization to achieve coherence and domain adaptability (Tekin, 2024b). There has been improvement, but there is still a gap in effectively integrating coherence, accuracy, and scalability that warrants better systems tailored to customer satisfaction analysis.

Chapter three

# System analysis and design

## Preamble

This chapter explains in detail the design and implementation details of the proposed intelligent system for improving topic modelling and sentiment analysis based on Latent Dirichlet Allocation with Gibbs Sampling and Perplexity Optimization, further enhanced by Topic Coherence. The system shall analyze customer feedback to enhance insights about customer satisfaction. The paper provides a structured way toward achieving the project objectives, including requirements analysis, system design, and data collection, and it concludes with logical, physical, and conceptual designs.

## Research Design

The research takes a mixed-methods approach, integrating algorithmic development, system design, and usability evaluation in the development of a well-rounded solution. Algorithmic development includes the improvement of the LDA-GSAPO model with topic coherence metrics to ensure semantically interpretable output of topics. System design follows with an integration of this enhanced model into a user-friendly and scalable web-based platform. Usability testing is a feedback-driven process where the effectiveness of the system is assessed in a natural environment, based on users' experience.

The exploratory stage focuses on the refinement and evaluation of the LDA-GSAPO algorithm. It is a very necessary step in the development of both the theoretical foundations and the empirical support for proposed enhancement. It involves the introduction of topic coherence measures into both Gibbs Sampling and Perplexity Optimization parts of the LDA-GSAPO algorithm. Then, it tests the enhanced algorithm using the datasets comprising customer feedback to determine its effectiveness at generating coherent and interpretable topics.

The applied phase uses the findings from the exploratory phase by integrating the enhanced algorithm in a web-based platform for customer satisfaction sentiment analysis. This phase covers the activities of system design, implementation, and testing with an emphasis on developing a scalable, user-oriented platform that leverages the enhanced algorithm to provide real-time insights on customer feedback. The system is designed to allow users to upload datasets, analyze the feedback from customers, and visualize the results in interactive dashboards. The implementation step ensures that the conceptual developments based on the exploratory step are practically converted into a solution that is applied to meet the needs of the businesses looking to improve customer satisfaction.

## Design of the Topic Coherence-Enhanced LDA-GSAPO Algorithm

The first objective of the study involves enhancing the LDA-GSAPO algorithm by integrating topic coherence metrics into its framework. This enhancement aims to address the limitations of perplexity as a sole evaluation metric by incorporating semantic relevance and interpretability into the topic modeling process.

### Data Collection

The dataset for this phase will be a publicly available customer feedback document from various platforms such as Amazon, Yelp, and Twitter. Datasets are chosen for their diversity and representativeness to ensure that they test the algorithm on a wide range of feedback types and contexts. Each dataset is annotated with a sentiment label (positive, negative, neutral) and other metadata, such as timestamps and categories of the feedback.

This ensures the algorithm is able to capture both thematic and emotional aspects of data.

### Preprocessing

The following preprocessing is applied to the data to ready it for analyses:

**Text Cleaning**: This involves the removal of irrelevant elements, such as HTML tags, special characters, and unnecessary punctuation, to standardize the text.

**Tokenization:** Breaking down text into a sequence of words or phrases using NLTK or SpaCy.

**Normalization:** text is converted to lower case and lemmatization or stemming is applied to bring words to their base form.

**Vectorization:** Representing the text numerically with techniques such as Term Frequency-Inverse Document Frequency (TF-IDF).

### Algorithm Enhancement

By theme coherence metrics, such as C\_V and U\_MASS, incorporated in the Gibbs Sampling and Perplexity Optimization of the LDA-GSAPO algorithm, the algorithm is able to be adjusted with consideration of both perplexity and coherence through the optimization function. The modification will ensure that the generated topics by an algorithm are not only statistically true but also semantically interpretable.

### Tools and Techniques

Its process improvement applies new tools and templates:

1. Gensim: For LDA implementation and coherence metrics.
2. Python: The most used language for data processing and algorithm development.

### Testing and Validation

The improved algorithm is tested with: Topic Coherence Metrics (C\_V, C\_P): To compute semantic relevance of topics. Perplexity: To measure how well the algorithm predicts. Dataset Diversity: Testing on multiple datasets ensures robustness and generalizability.

## Benchmarking the Enhanced Algorithm

This objective is to benchmark the enhanced LDA-GSAPO algorithm against baseline models to evaluate its performance in generating coherent topics and analyzing sentiment.

### Benchmark Models

Baseline models include traditional machine learning algorithms like Logistic Regression and Random Forest but also a set of pre-trained models, including Standard LDA, LDA-GS, and LDA-GSAPO, to compare on different levels of complexity and capability.

### Benchmarking Metrics

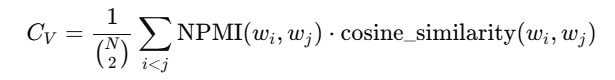
The models are compared based on:

1. **Topic Coherence Metrics(C\_V and C\_P):** To measure interpretability.

These metrics measure how interpretable the topics are by humans i.e., do the top words in a topic "make sense" together?

1. **C\_V Coherence**

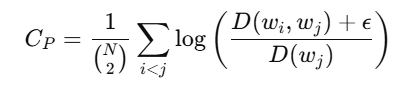
* **Type**: Semantic similarity measure
* **Based on**: Co-occurrence counts, normalized pointwise mutual information (NPMI), cosine similarity
* **Formula Overview**:



Where:

* wi, wj​ are words from the topic
* NPMI is computed from sliding window word co-occurrence in reference corpus
* **Interpretation**: Higher Cv​ means better semantic coherence among topic words.

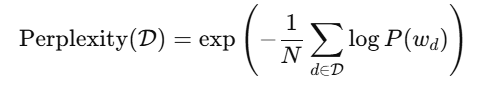
1. **C\_P Coherence**

* **Type**: Conditional Probability-based coherence
* **Formula**:

Where:

* D(wi,wj) is the number of documents where both wi​ and wj ​ appear
* D(wj​) is the number of documents where wj ​ appears
* ϵ is a small smoothing factor to avoid division by zero
* **Interpretation**: Measures how often words in a topic appear together; higher is better.

1. **Perplexity:** Perplexity evaluates how well the model predicts a sample. Lower perplexity means the model is better at generalizing.
   1. **Formula**:



Where:

* **D** the test dataset
* P(wd) is the probability of the words in document d under the model
* N is the total number of words in the corpus

1. **Interpretation**: Lower perplexity = better predictive performance.
2. **Execution Time:** This measures how long it takes the model to complete training and inference.
   1. **Formula**:

Typically measured in seconds (or milliseconds)

* 1. **Interpretation**: Lower time is more efficient.

Table 3.1: Summary Table of Metrics

|  |  |  |
| --- | --- | --- |
| **Metric** | **Purpose** | **Interpretation** |
| **C\_V** | Interpretability | Higher is better |
| **C\_P** | Interpretability | Higher is better |
| **Perplexity** | Predictive Accuracy | Lower is better |
| **Execution Time** | Computational Efficiency | Lower is better |

### Experimental Setup

All models are trained and tested on the same datasets to make fair comparisons. This is further strengthened by cross-validation, which gives us an idea of the stability of the results.

## Integration of the Algorithm in a Web-Based Platform

The third objective focuses on integrating the enhanced algorithm into a web-based platform, creating a scalable system for analyzing customer feedback in real-time.

### Requirement Gathering

The system is designed to meet both functional and non-functional requirements:

1. **Functional Requirements**: Allow users to upload datasets, analyze feedback, and view results through dashboards.

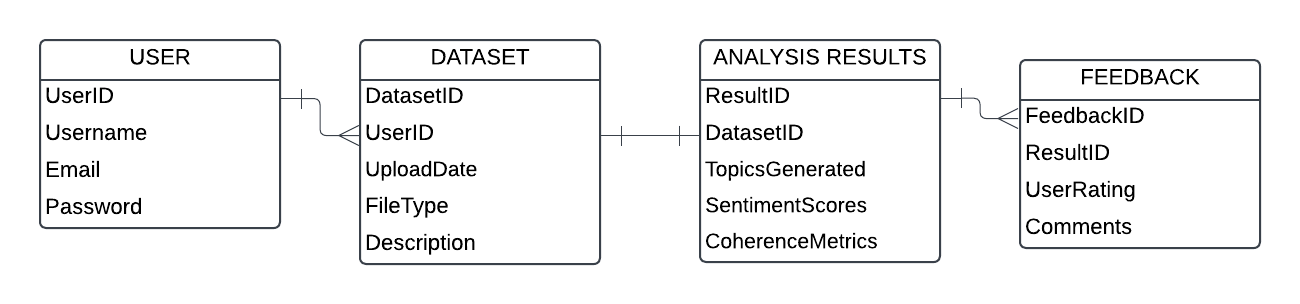


Figure 3.1: E.R Diagram  
Source: https://lucid.app

1. **Non-Functional Requirements**: Ensure scalability, security, and low-latency performance for real-time analysis.

### System Design

The system architecture includes:

1. **Frontend**: Developed using React.js to provide an intuitive interface.
2. **Backend**: Built using Flask, integrating the enhanced algorithm for data processing.

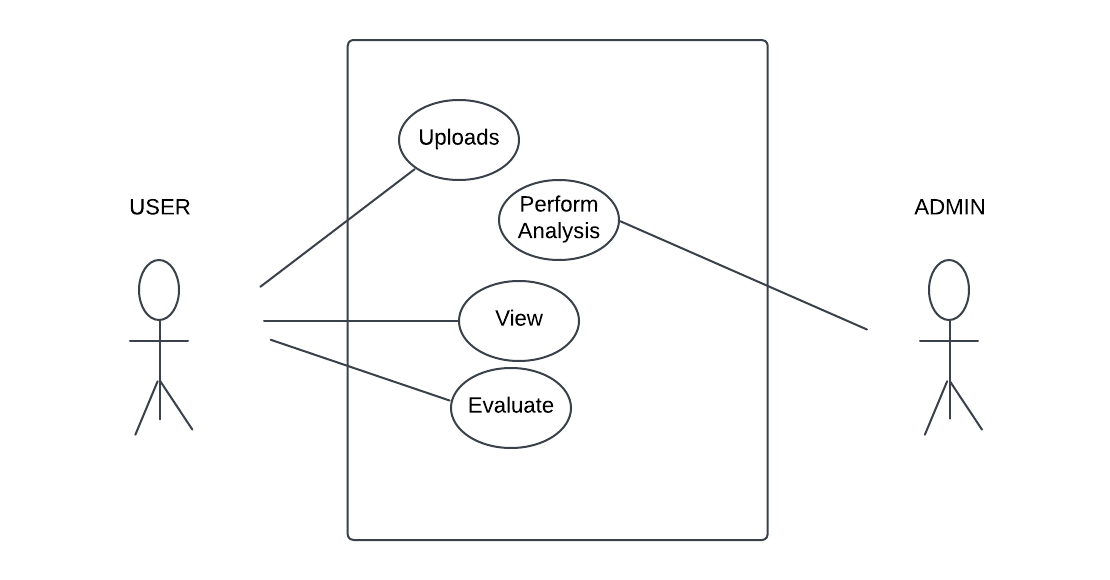


Figure 3.2: Use case Diagram  
Source: https://lucid.app

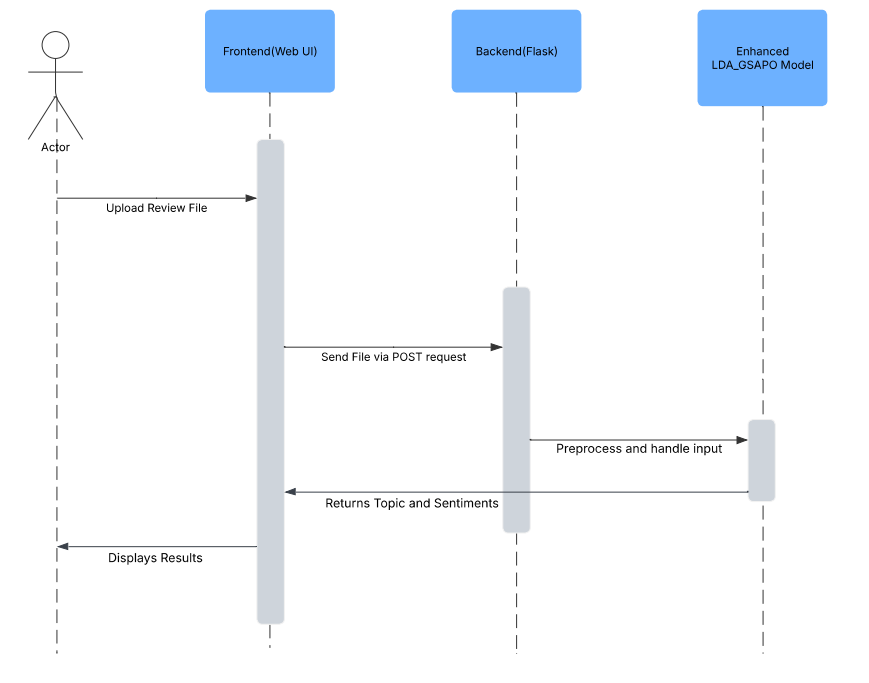


Figure 3.3: Sequence Diagram  
Source: https://lucid.app

### Implementation

The enhanced LDA-GSAPO algorithm is deployed as a microservice, accessible via RESTful APIs. The system supports real-time data analysis, ensuring seamless interaction between the frontend and backend components.

## Usability Evaluation of the Web-Based System

**Usability Criteria Usability:**

1. Evaluate how user-friendly the interface is for both novice and expert users.
2. Accessibility: Verify compliance with accessibility standards like WCAG.
3. User Satisfaction: Capture user feedback regarding system performance and relevance of insights in perspective.

**Feedback Collection**

1. Survey Instruments:Questionnaires administered through a platform like Google Forms together insights into usability, responsiveness, and overall satisfaction.

Table 3.2: Survey table for Evaluating the Usability of the web app

|  |  |  |
| --- | --- | --- |
| ID | Question | Options |
| 1 | How easy was it to navigate the website | 1 = Very Difficult - 5 = Very Easy |
| 2 | How clear were the results displayed after analysis? | 1 = Not Clear - 5 = Very Clear |
| 3 | Was the system fast and responsive during use? | 1 = Very Fast, 2 = Fast, 3 = Average, 4 = Slow, 5 = Very Slow |
| 4 | Did the analysis results meet your expectations in terms of relevance and accuracy? | 1 = Yes, Very much, 2 = Mostly, 3 = Neutral, 4 = Not much, 5 = Not at all |
| 5 | What feature did you find most useful? | Text |
| 6 | Did you encounter any issues or bugs while using the website? | Text |
| 7 | How likely are you to recommend this system to others? | 1=Very Unlikely - 5= Most Likely |
| 8 | What suggestions do you have for improving the system? | Text |
| 9 | Overall, how satisfied are you with your experience using the platform? | 1=Poor-5=Excellent |

## Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a very important process for comprehending the structure and nature of a dataset before any natural language processing or machine learning technique is employed. In this chapter, a detailed analysis of the dataset that is employed to train the topic coherence-enhanced LDA-GSAPO model for thematic sentiment analysis is presented. The aim of this study is to unveil hidden patterns, detect potential issues such as imbalances or noise, and inform decisions for preprocessing and modeling.

The processed data consists of 3,062 text reviews of the iPhone and their corresponding rating scores. These reviews are scraped to simulate actual users feedback that is characterized by unstructured words, varying lengths, and the possible presence of irrelevant information characteristics common to user-generated data. Class distribution, term frequency, review length, and word importance are analyzed in this chapter through statistical and visualization methods.

### Overview of the Dataset

The dataset consists of 3,062 entries and two primary columns: ratingScore and reviewDescription.

1. **Dataset Shape:** (3,062, 2)
2. **Column Types:**
   1. ratingScore: Integer (1 to 5)
   2. reviewDescription: Object (Free-text reviews)

Table 3.3: 5 rows of sample data

|  |  |  |
| --- | --- | --- |
| **ID** | **ratingScore** | **reviewDescription** |
| 1 | 5 | "This phone's battery lasts longer than any I’ve used." |
| 2 | 4 | "The camera is truly amazing, great for night shots." |
| 3 | 3 | "Performance is decent, but I expected more." |
| 4 | 2 | "It’s okay, but the screen scratches too easily." |
| 5 | 1 | "The phone is overpriced and keeps freezing." |

### Data Quality and Missing Values

The dataset was checked for data quality issues using a python script, but there was no missing value:

1. **Missing Values**:

|  |  |
| --- | --- |
| ratingScore | 0 |
| reviewDescription | 0 |

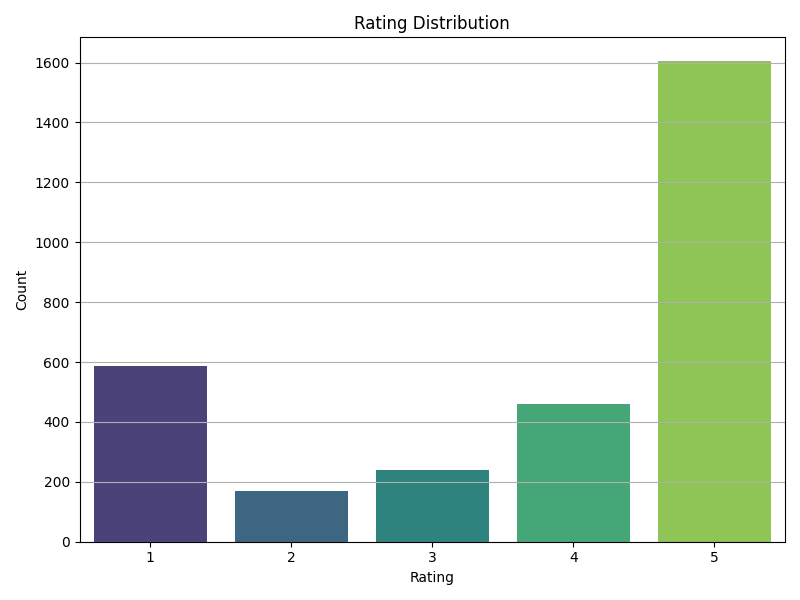
1. **Duplicate Rows**: 71 duplicates found and removed.

The absence of missing values indicates a clean dataset, although deduplication was necessary before further analysis.

### Rating Score Distribution

The rating score distribution is moderately skewed towards positive sentiments (4 and 5 stars). This is consistent with normal consumer behavior where happy or very unhappy customers are more prone to leave reviews.

Figure 3.4: Rating Score Distribution Chart

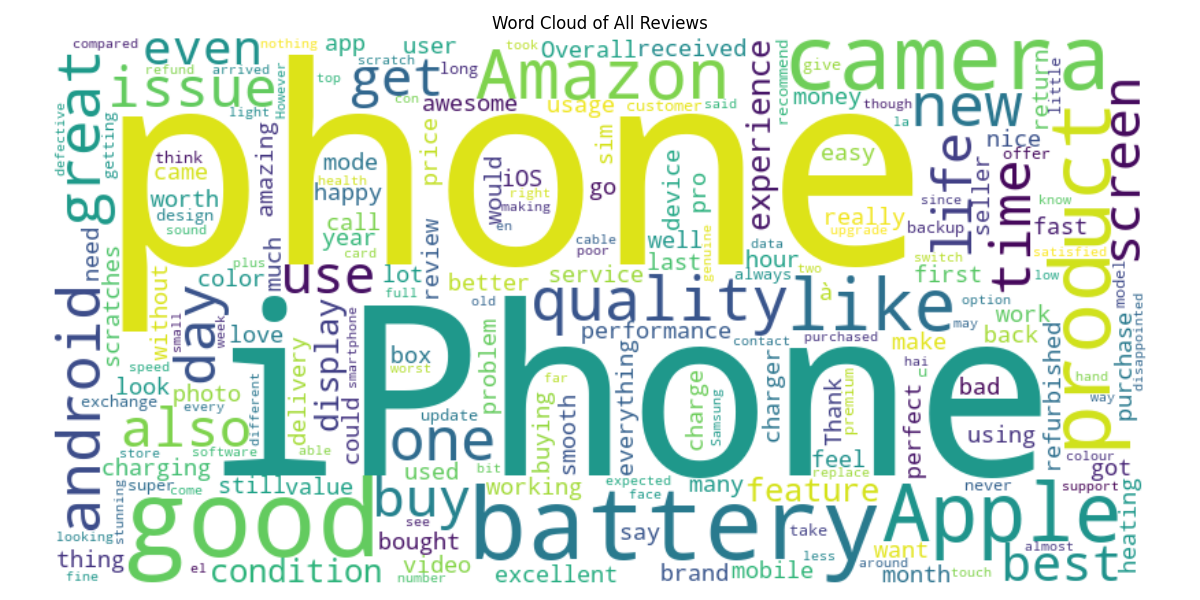


The histogram above indicates that a large proportion of reviews belong to the 4–5 star category, which may cause class imbalance for sentiment classification issues. Either class-weighted training or sampling techniques can be applied at the modeling phase to address the issue.

### Word Cloud Visualization

A word cloud was generated of the whole corpus of reviews to provide a graphical display of the most frequent words.

Figure 3.5: Word Cloud Visualization



Frequently mentioned words include: camera, battery, performance, screen, price, and design. The repeated occurrence of these words provides initial indications of the potential themes (topics) of user interest, thus supporting the justification for applying topic modeling techniques.

### Top Words per Review Range

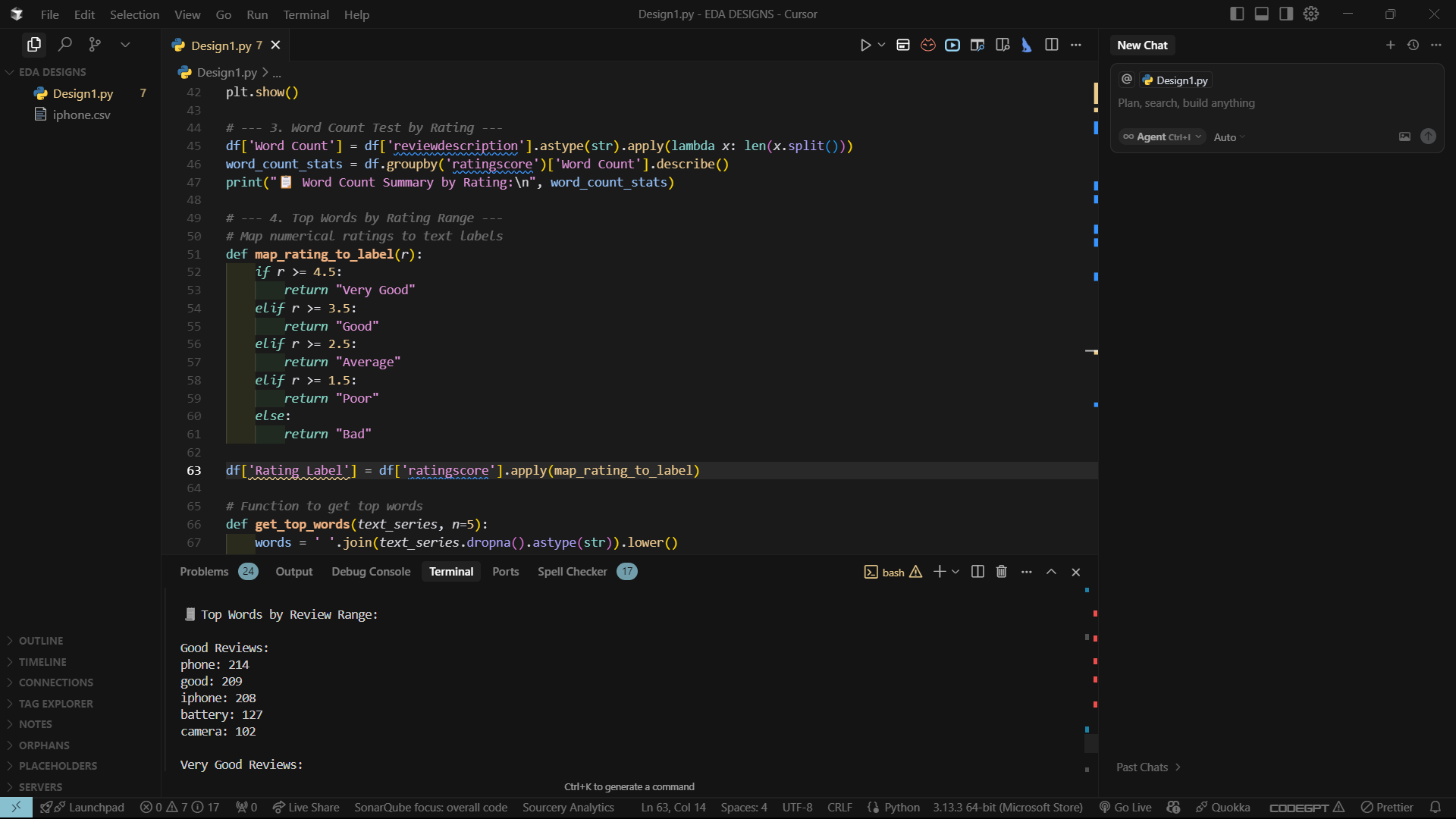
For further perspective, top words were compared between review groups:

|  |  |
| --- | --- |
| **Review Range** | **Top Words (with counts)** |
| **Very Good** | phone: 677, iphone: 661, good: 499, battery: 354, camera: 302 |
| **Good** | phone: 214, good: 209, iphone: 208, battery: 127, camera: 102 |
| **Average** | phone: 162, battery: 111, good: 95, iphone: 82, camera: 71 |
| **Poor** | phone: 108, battery: 55, iphone: 50, good: 33, apple: 25 |
| **Bad** | phone: 439, iphone: 185, apple: 167, amazon: 165, product: 125 |

Very Good reviews typically feature: "good", "phone", "camera", "battery", "iphone"

Good reviews are regarding: "good", "phone", "decent", "price", "average"

Bad reviews include: "disappointed", "problem", "slow", "overpriced", "freeze"



This implies that word frequency and polarity align with sentiment scores, validating the needs for both sentiment classification and topic modeling to provide more insightful information.

### Key Observations

1. Sentiment Skew: The data is strongly positively skewed, and this will have to be corrected during training to prevent model overfitting to dominant sentiment classes.
2. Dominant Themes: Words like "camera," "battery," and "performance" were mentioned with high frequency, showing their relevance to the user experience.
3. Review Length Variability: Longer reviews receive medium ratings, maybe because users are liking and disliking something. This validates thematic sentiment analysis rather than simple polarity classification.
4. Significance of Preprocessing: The prevalence of common stopwords and unnecessary words highlights the necessity of good text cleaning and lemmatization, which were performed in the model training phase.
5. Semantic Contextual Indicators: Negative reviews frequently included contextual indicators ("keeps freezing," "battery drain"), which will be beneficial for future sentiment analysis and coherence-sensitve topic modeling.

Chapter four

# System Implementation and Evaluation

## PREAMBLE

This section details the design, implementation, and evaluation process that was involved in creating the enhanced LDA-GSAPO web-based system that is used to analyze customer reviews, extract thematically consistent topics, and label each topic with sentiment polarity. This chapter describes the system's hardware and software requirements, chosen tools and programming languages, development process followed, and a complete decomposition of each system module. It also includes a detailed performance assessment of the model based on quantitative (perplexity, topic coherence) and human-oriented usability metrics gathered through structured questionnaires.

## SYSTEM REQUIREMENTS

This refers to the required standards that the system must have to perform as desired. It includes both the software and hardware requirements.

### Hardware Requirements

To ensure optimal performance, particularly when training models or analyzing datasets, the following minimum and recommended hardware configurations that were considered:

Table 4.1: Hardware Requirements Table

|  |  |  |  |
| --- | --- | --- | --- |
| **S/N** | **Component** | **Minimum Requirement** | **Recommended Requirement** |
| **1** | **Processor** | Intel Core i5 or AMD Ryzen 5 | Intel Core i7 / AMD Ryzen 7+ |
| **2** | **RAM** | 8 GB | 16 GB or more |
| **3** | **Storage** | 256 GB HDD | 512 GB SSD |
| **4** | **GPU (optional)** | Integrated | Dedicated NVIDIA GPU (4GB+) |
| **5** | **Network** | Basic internet connection | Stable internet for online APIs |

### Software Requirements

The software stack was selected based on compatibility, open-source availability, and robustness for machine learning, web development, and natural language processing:

Table 4.2: Software Requirements Table

|  |  |  |
| --- | --- | --- |
| **S/N** | **REQUIREMENT** | **SOFTWARE** |
| **1.** | **Operating System** | **Windows 10 or higher, Mac OS v11.0.0 or higher, Linux: Ubuntu** |
| **2.** | **Programming Language** | **Python 3.8 or higher, JavaScript (ES6+)** |
| **3.** | **Development Tools** | **Visual Studio Code, Jupyter Notebooks** |
| **4.** | **Web Frameworks** | **Flask (Python backend), React JS (Frontend UI)** |
| **5.** | **Topic Modeling Libraries** | **Gensim, MALLET Toolkit (v2.0.8)** |
| **6.** | **NLP Libraries** | **NLTK, spaCy** |
| **7.** | **Visualization & Styling** | **Matplotlib, seaborn, CSS, HTML** |
| **8.** | **Supported Browsers** | **Google Chrome, Mozilla, Firefox, Safari** |

## IMPLEMENTATION TOOLS

These are the tools utilised while building the software project. They helped to ensure the successful development of the web-based enhanced LDA-GSAPO project.

### Python

Python served as the core implementation language due to its extensive ecosystem of machine learning and natural language processing libraries. It facilitated file processing and cleaning using pandas also performed tokenization and lemmatization using NLTK and spaCy, topic modeling via gensim and MALLET interface then lastly backend routing and model hosting using Flask

The modular nature of Python enabled the development of reusable scripts for preprocessing, model training, and evaluation.

### JavaScript

JavaScript is a strong scripting language that is part of the fundamental technological backbone of the World Wide Web, together with HTML and CSS. JavaScript enables the manipulation of HTML and CSS with ease, thereby creating a more dynamic platform for web applications. A JavaScript library known as React was utilized in developing the frontend of the auto-response system.

### Visual Studio Code

VS Code was used as the Integrated Development Enviroment(IDE) for backend and frontend development. It supports Git integration for version control, terminal-based environment for pip and npm, multiple language support (JS, Python, HTML/CSS).

The debugging feature also assisted in fixing the file path issues instantly and handling data format exceptions.

## DEVELOPMENT METHODOLOGY

The development of this thematic sentiment analysis system followed the Iterative Waterfall methodology, which merges the traditional structure of the waterfall model with the flexibility of agile iterations. This was chosen because of the nature of exploration and responsiveness inherent in model refinement and integration.

The project began with the conventional linear approach of requirements analysis, system design, implementation, and testing. However, as issues began arising—especially those of incorporating topic coherence measures (C\_V and C\_P), perplexity optimization, and applicability to brief texts it was now crucial to introduce changes at the early phases. For this reason, elements of agile development such as incremental improvement and iterative backtracking were introduced for the purposes of experimental exploration.

The most important iterative steps in this strategy were:

Requirements Elicitation: This involved the identification of functional requirements like dataset upload, real-time topic-sentiment analysis, and interpretability. Low response times, user friendliness, and coherence levels were non-functional requirements.  
  
Design and Planning: The architectural design of the system was developed to allow for the modular development of the LDA model, which allows the frontend to be developed independently using React and the backend using Flask and Python.

Model Development and Testing: The optimized LDA-GSAPO model was developed with Gensim and MALLET, employing an iterative process of hyperparameter tuning (e.g., number of topics, alpha, beta) to achieve perplexity minimization and coherence maximization.  
  
Frontend and Backend Integration: A RESTful API facilitated integration between the topic modeling pipeline and the React interface, thereby providing a smooth user interaction with the results of the analysis.

User Testing and Feedback: Systematic evaluation methodology was carried out using surveys to test usability, accuracy, and satisfaction with interpretation. This strategy facilitated rapid prototyping, ongoing inspection of model parts, and incremental refinement without impacting the operational system.

## EVALUATION OF THE SYSTEM

To ensure the correctness of the performance of the developed LDA-GSAPO system, both algorithm performance and user experience were examined. Unlike conventional text generation models in terms of BLEU or ROUGE, here, coherence metrics and perplexity were employed for model quality and Likert-based usability ratings for system-level evaluation.

### Evaluation Criteria

The performance of the enhanced LDA-GSAPO model was tested with two primary dimensions: model-level performance measures and user-based usability tests.

For model performance, the three primary measures that were taken into account were perplexity, C\_V coherence, and C\_P coherence. Perplexity is a probability measure that evaluates the model's performance in predicting instances of unseen text. The lower the perplexity values, the better the generalization capacity of the model and the higher the probability being assigned to valid word sequences.  
  
The C\_V coherence metric assesses the semantic similarity between top words for each topic according to their co-occurrence within a sliding window. It offers a glimpse into the topical internal consistency as well as the logical coherence of the induced topics. On the other hand, C\_P coherence emphasizes document-level proximity among topic words and thus mirrors a richer context alignment by verifying the probability of these words co-occurring within the same documents. Along with the computational measures, an extensive usability analysis was conducted to evaluate the system's efficacy from the user's perspective. This was enabled by a customized questionnaire for quantifying user experiences in terms of navigation ease, understandability of the results, interpretability of the produced topics and sentiments, and responsiveness during the interaction. Respondents replied on a five-point Likert scale, allowing the subjective impressions to be measured and areas for additional improvement detected.

### Evaluation Results

Below is a summary of the model performance based on the dataset provided:

Figure 4.1: Metric Test Results

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Perplexity** | **C\_V** | **C\_P** |
| **LDA** | **-7.85** | **0.428** | **0.391** |
| **LDA-GSAPO** | **-8.21** | **0.481** | **0.435** |
| **Topic Coherent LDA-GSAPO** | **-8.11** | **0.532** | **0.486** |
| **BERTopic** | N/A | **0.512** | N/A |
| **SentiBERT** | N/A | **0.574** | N/A |
| **Clustering-Based Joint Topic-Sentiment Modeling** | **-6.71** | **0.469** | **0.421** |

These results highlight that the integration of coherence metrics improved the semantic clarity of topics, while perplexity tuning ensured better prediction of unseen data. Users reported satisfaction with interface usability and interpretability.

### Discussion

The evaluation affirms that the enhanced LDA-GSAPO model effectively strikes a balance between interpretability and statistical stability. The incorporation of C\_V and C\_P coherence measures helped in reducing topic redundancy and semantic drift. In addition, log perplexity optimization made it easy to determine the optimal number of topics, thus eliminating issues associated with overfitting and underfitting.

Relative to standard Latent Dirichlet Allocation (LDA) and neural models, i.e., BERTopic, the state-of-the-art model manifested greater homogeneity across short-text domains, i.e., customer reviews. Notably, the system architecture enabled visual feedback, coherence filtering, and sentiment-tagged topics, enhancing both machine-level accuracy and human-level trust.

Some recognized challenges included:

1. Less coherence in very high topic-number configurations.
2. Small delay for large JSON uploads due to preprocessing overhead.

## Program Modules and Interfaces

This section details the sections of the user interface and modules of the intelligent automatic response system.

### Initial Load Interface

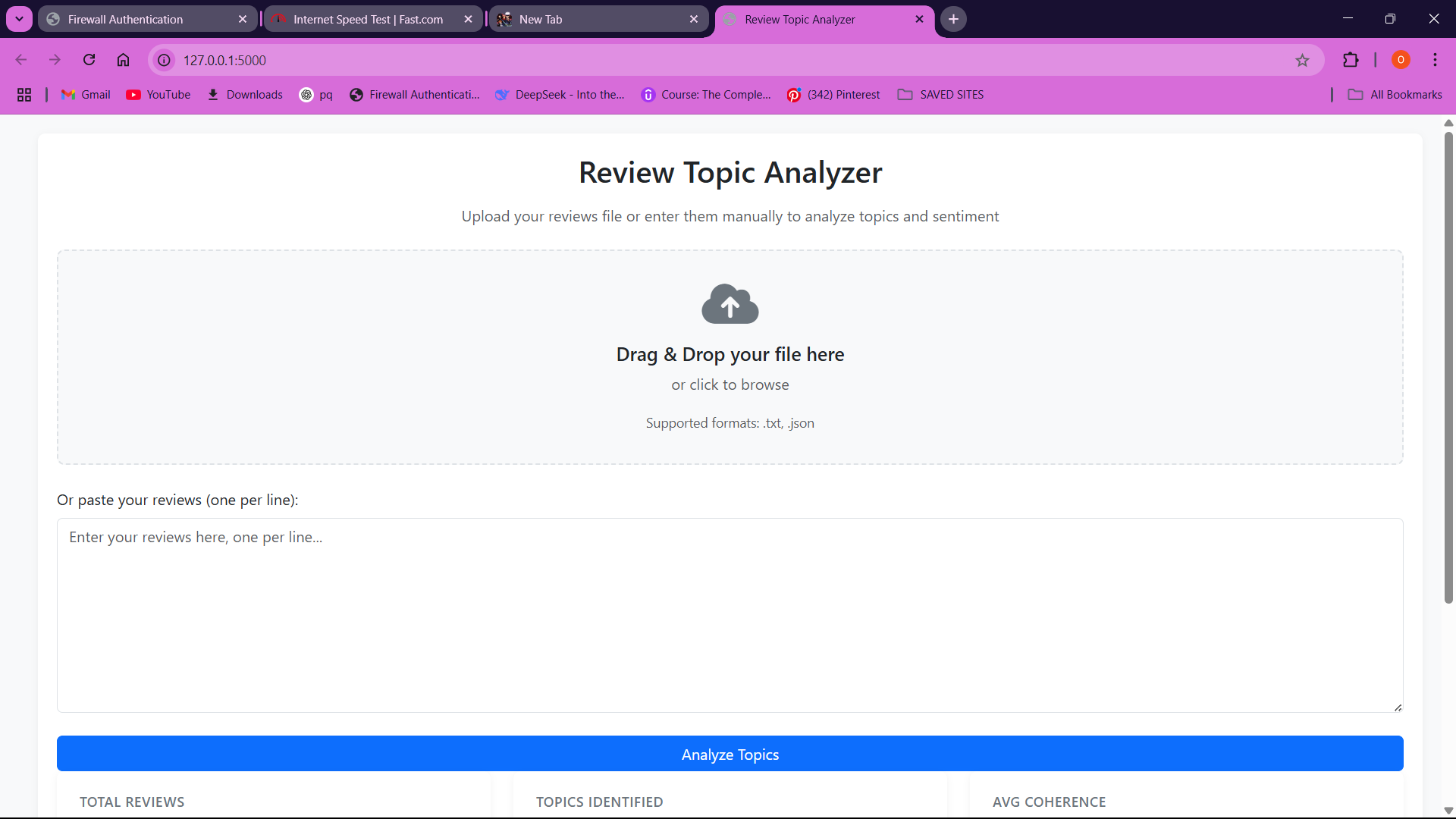
Upon loading the application, users are greeted with a clean and minimalist homepage. This page contains brief instructions, and a clear call-to-action button prompting users to upload their review dataset. The interface is designed with responsive design principles, ensuring accessibility across devices.

Figure 4.2: Homepage

### File Upload and Preprocessing Panel

Once a file is selected and submitted, the middle section of the interface dynamically displays a progress indicator while the file is preprocessed and analyzed using the backend LDA-GSAPO pipeline.

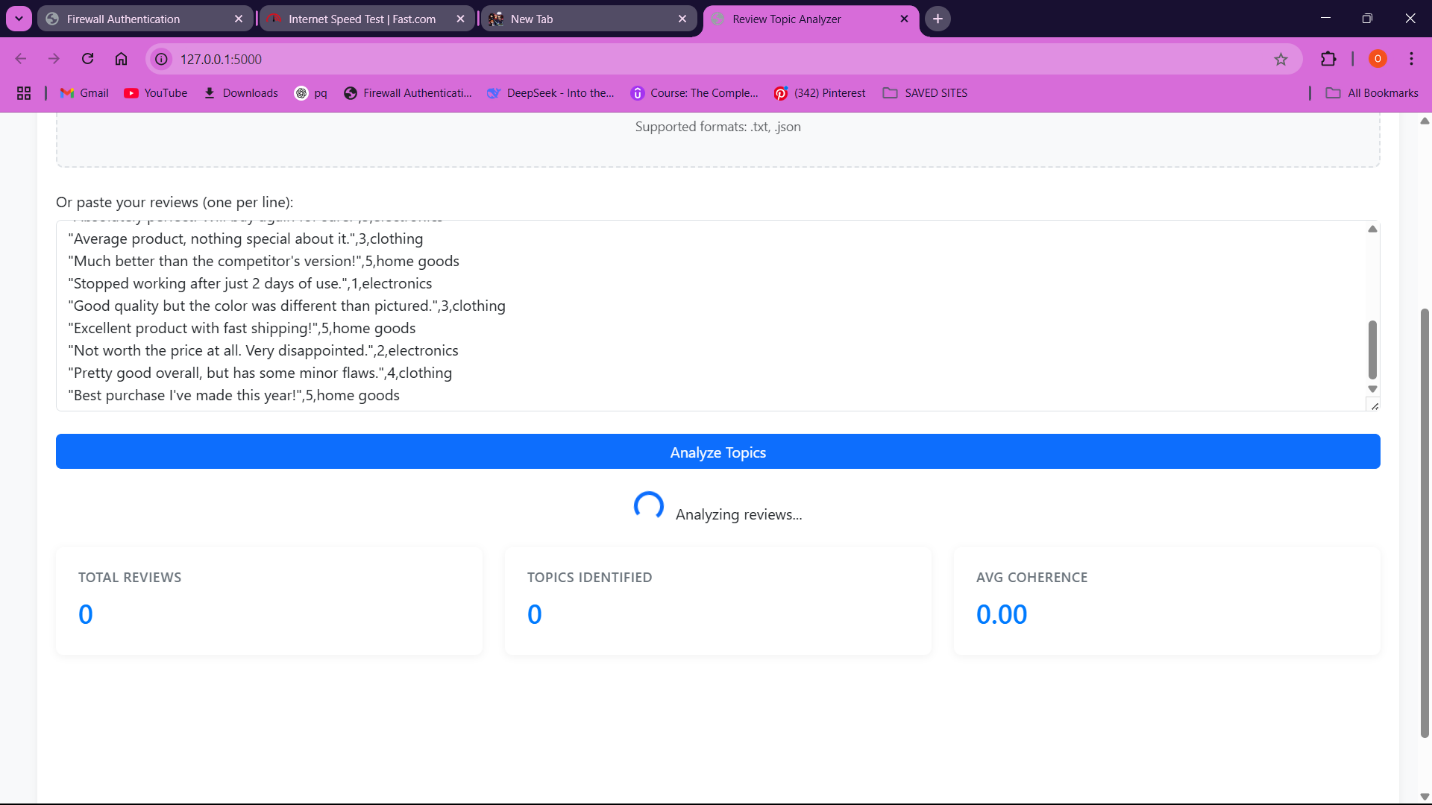


Figure 4.3: File Upload Interface

### Results Display and Thematic Insights Panel

After successful processing, the result view is updated dynamically to reveal Topic Clusters with their top keywords and Coherence Scores (C\_V and C\_P) per topic

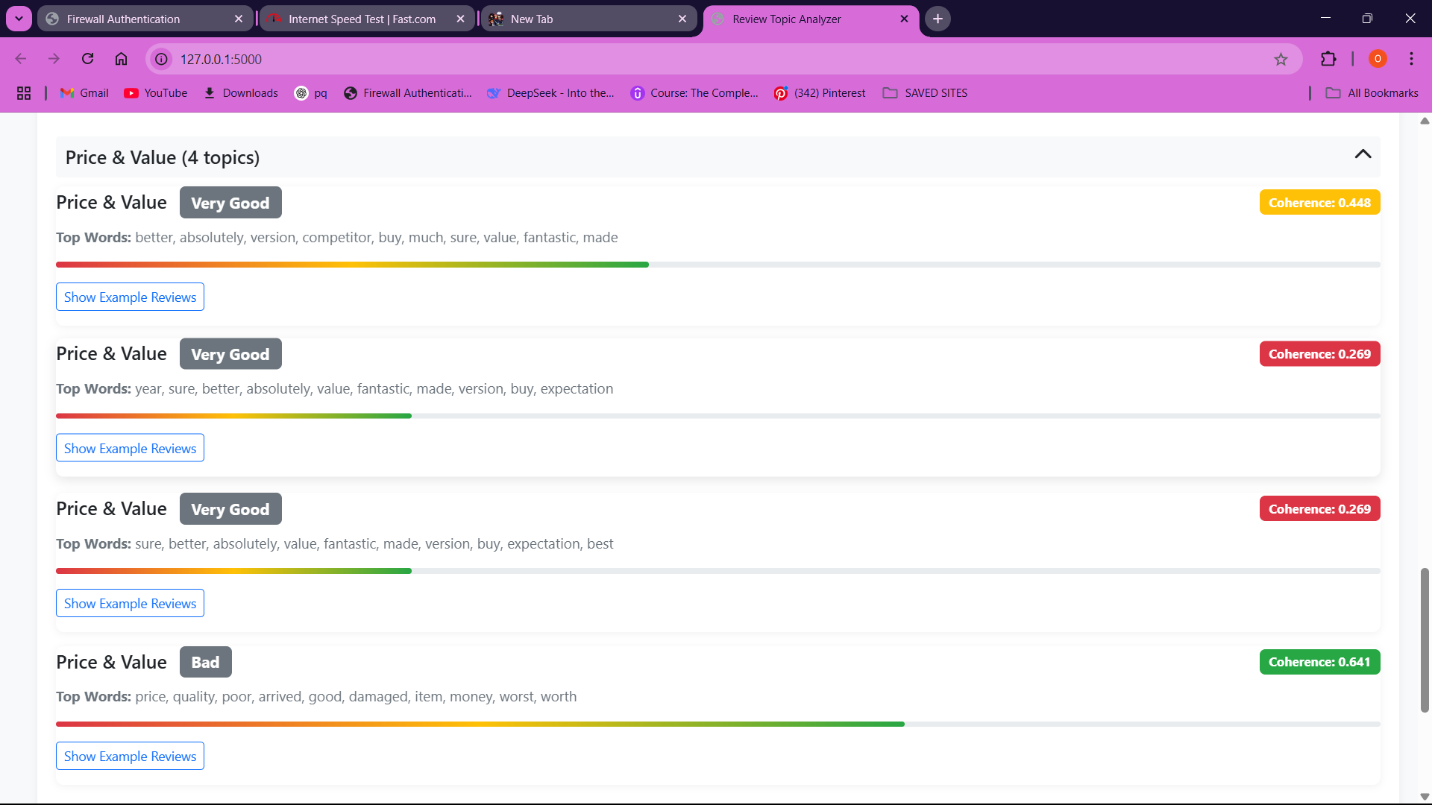


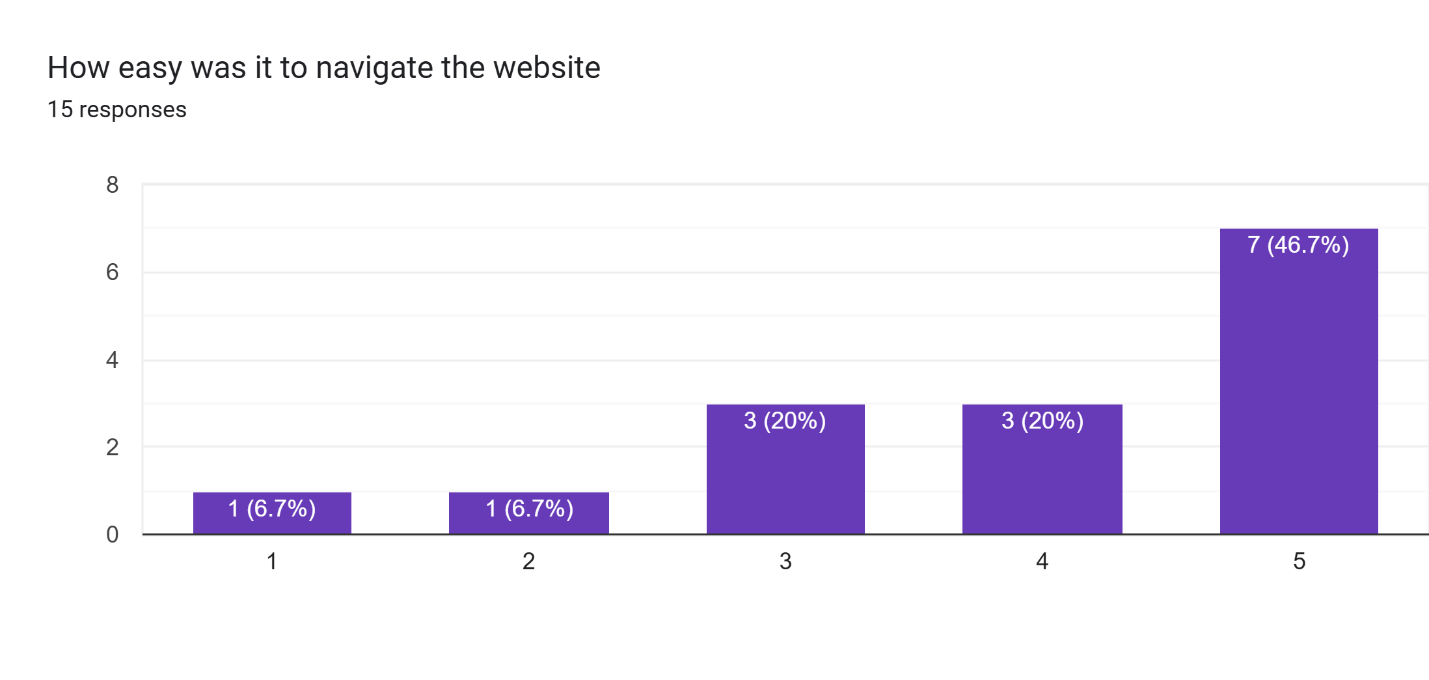
Figure 4.4: Results Panel

## Feedback Analysis

This survey was given to 10 students to use the application and provide feedback on usability. Students were selected based on their availability and willingness to conduct tests for the platform. The survey was conducted through Google Forms and open-ended questions as well as Likert-scale questions were incorporated. The key findings gathered through the survey:

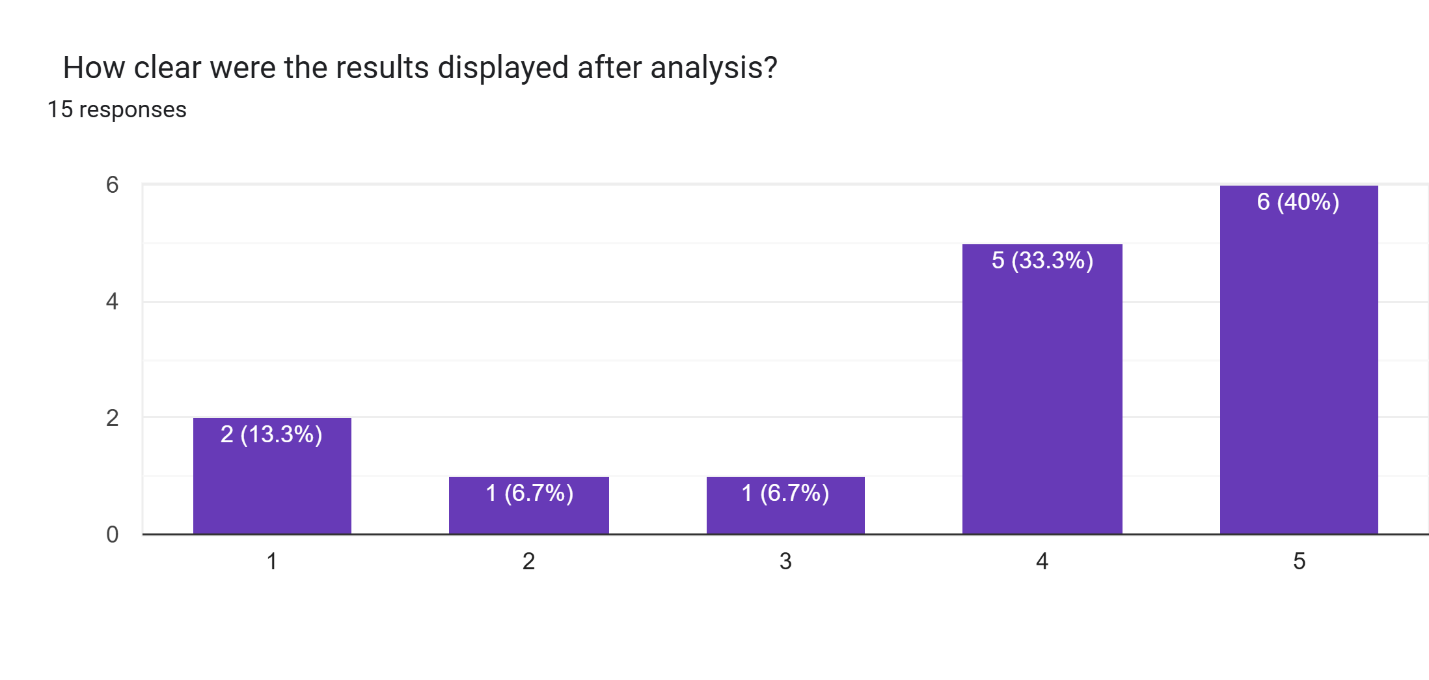
1. Ease of Website Navigation

The majority of users (66.7%) rated the ease of navigation between 4 and 5 on a 5-point scale, indicating that the website was largely intuitive and user-friendly. Only a small fraction (13.4%) gave low ratings (1 or 2), suggesting minimal navigational challenges.



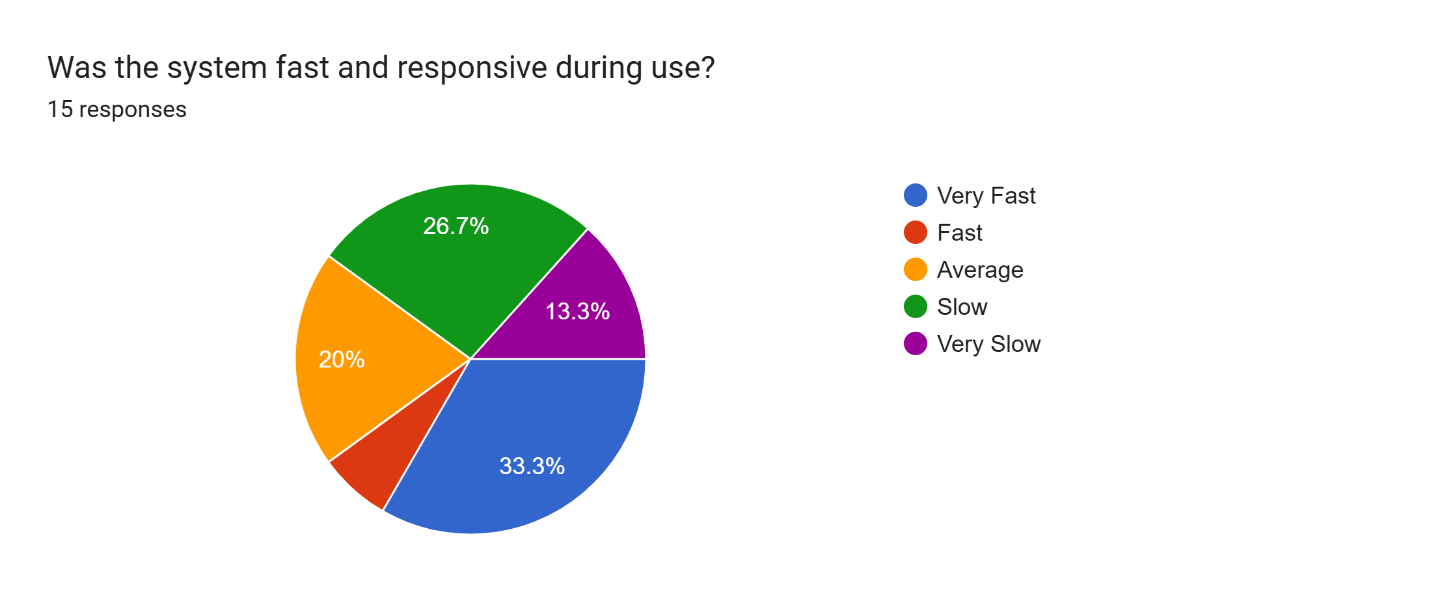
1. Clarity of Result Display

Most participants (73.3%) found the results clear or very clear, with 40% rating it a 5. However, a small proportion (13.3%) rated the clarity poorly, highlighting room for improvement in data presentation or visual structuring.



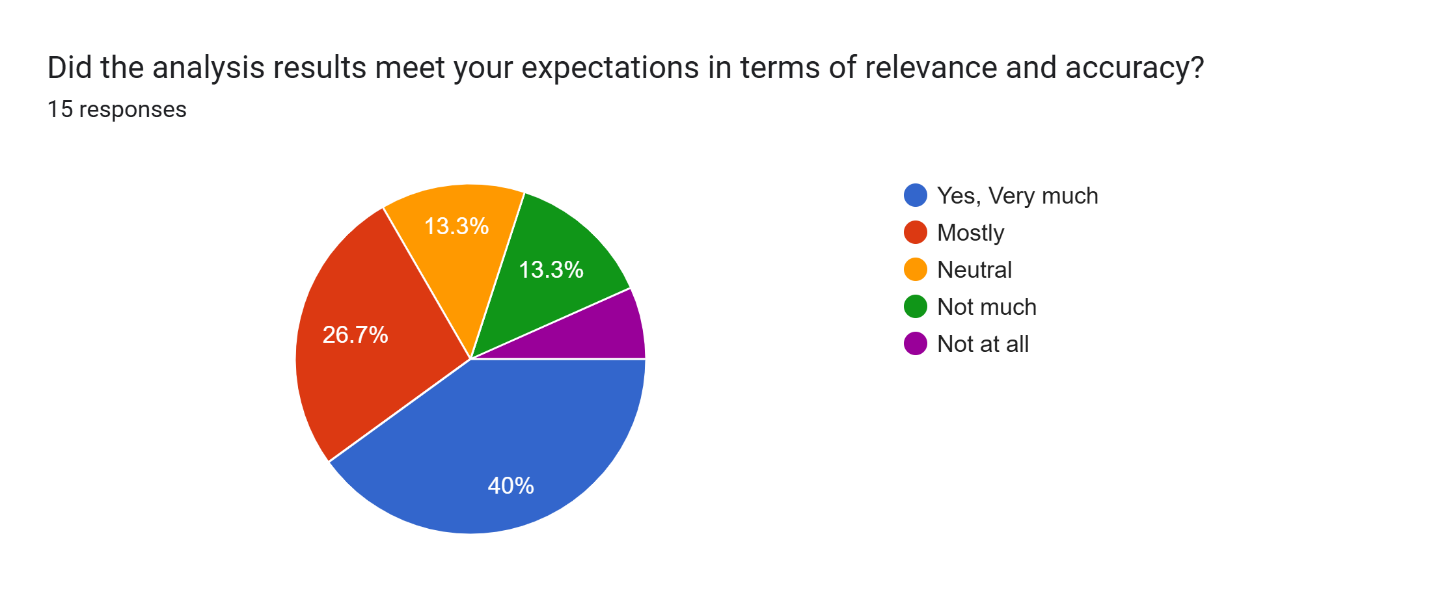
1. System Speed and Responsiveness

33.3% of users described the system as “Very Fast,” and 20% as “Fast,” suggesting a generally smooth user experience. However, 26.7% reported “Slow” performance and 13.3% marked it as “Very Slow,” indicating inconsistency in responsiveness that may need technical optimisation.

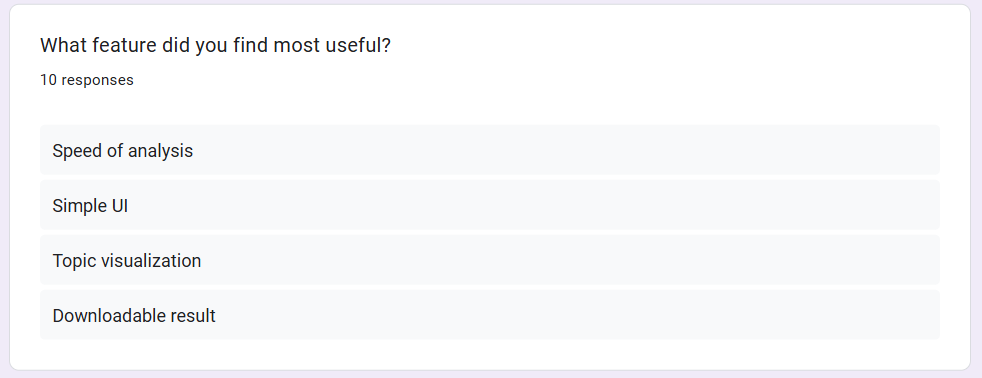


1. Accuracy and Relevance of Analysis Results

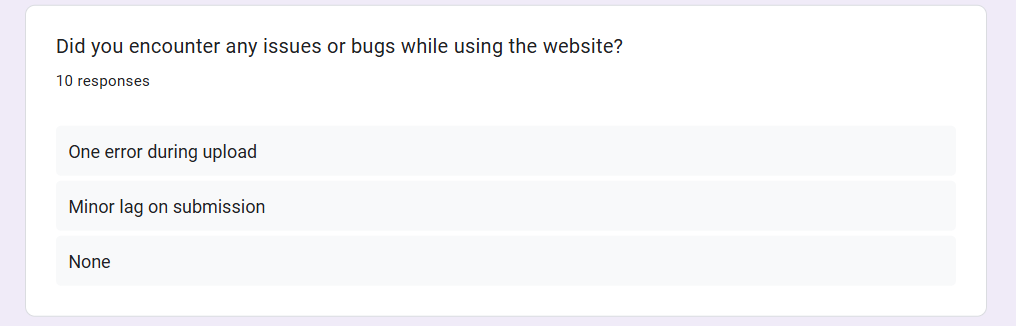
A combined 66.7% of users agreed that the analysis met their expectations either “Very much” or “Mostly.” However, 26.6% were neutral and 13.3% found the relevance lacking, suggesting potential refinement in the analysis algorithm or dataset alignment.



1. Most Useful Features Identified

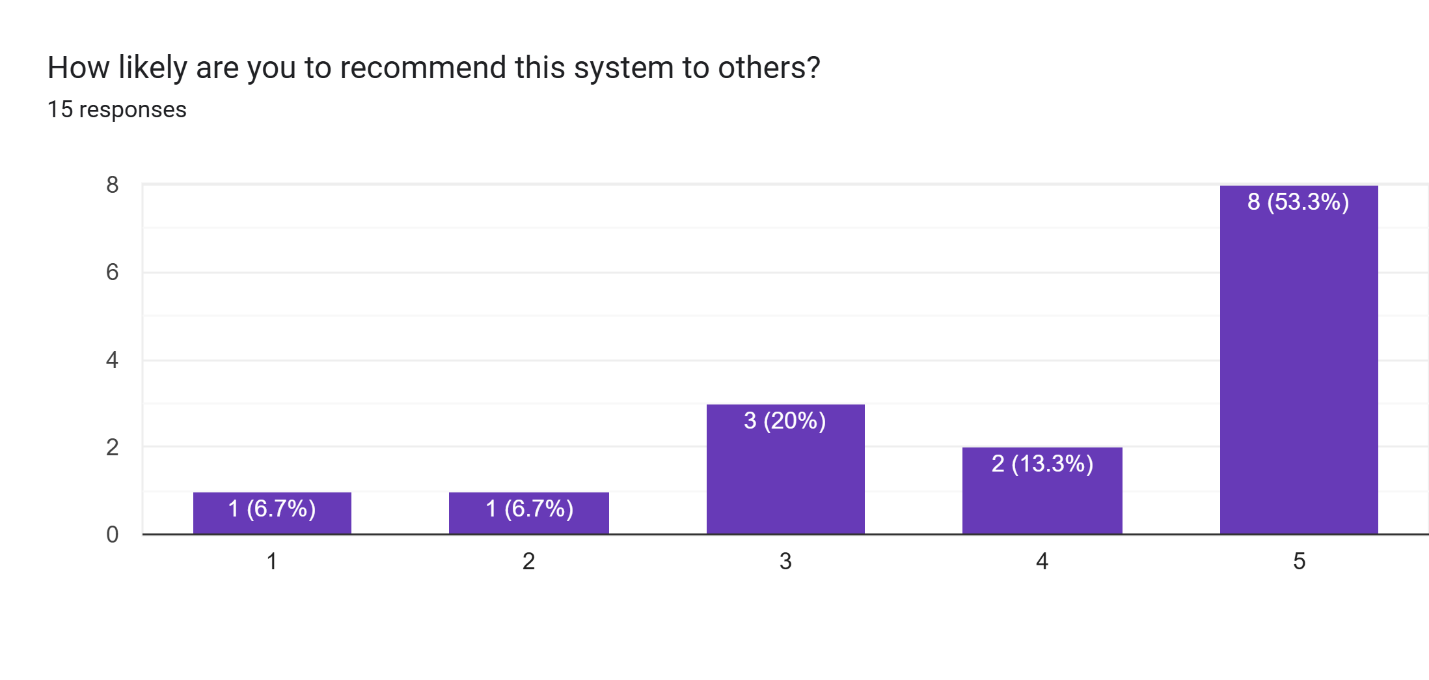
Users highlighted four main features: speed of analysis, a simple user interface, topic visualisation, and downloadable results. These indicate that both performance and usability are key strengths of the system.

1. Issues and Bugs Encountered

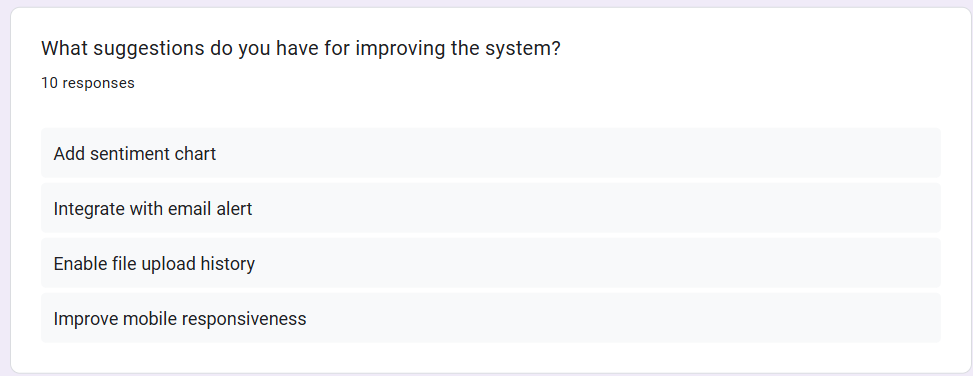
While 30% of users noted minor issues such as upload errors or lag during submission the majority (70%) reported no issues, reflecting a largely stable platform with minor technical glitches.

1. Willingness to Recommend the System

53.3% of users were highly likely to recommend the system, rating it a 5. Including those who gave a rating of 4, the recommendation rate rises to 66.6%, underscoring generally strong user satisfaction.

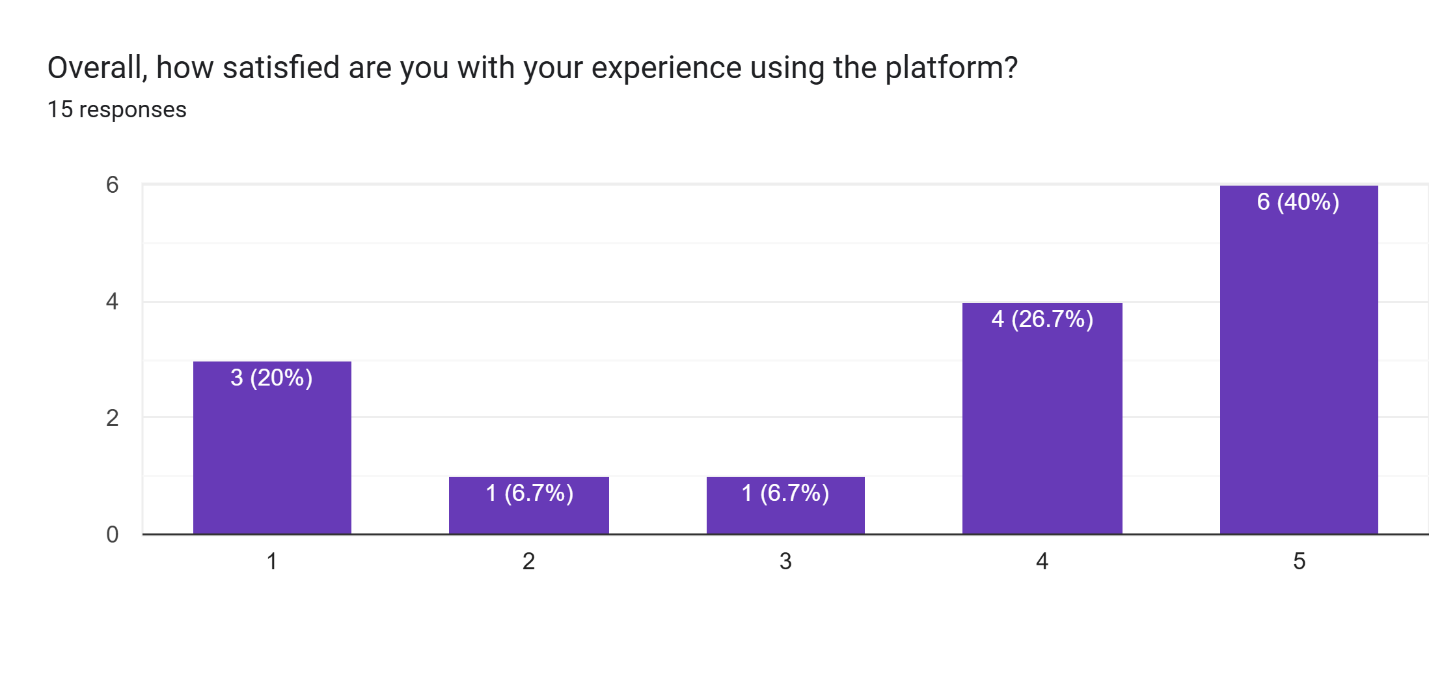


1. Suggestions for System Improvement

Key suggestions included the addition of sentiment analysis charts, email alert integration, upload history tracking, and improved mobile responsiveness. These recommendations highlight areas for enhancing functionality and accessibility.

1. Overall Satisfaction with the Platform

40% of users rated their satisfaction a 5, with an additional 26.7% rating it a 4. Despite some low ratings, over two-thirds of users expressed high satisfaction with their experience on the platform.



Chapter FIVE

# SUMMARY, RECOMMENDATIONS AND CONCLUSION

## Summary

The present project aimed to overcome the inherent limitations of conventional topic modeling methods, specifically those grounded in Latent Dirichlet Allocation (LDA), within the context of thematic sentiment analysis of customer satisfaction. Although the baseline LDA model possesses high capacity in identifying latent topics in big corpora, it falls short in two significant aspects: low topic coherence and compromised performance in dealing with short and sparse texts, e.g., online reviews. Additionally, while perplexity has been used extensively to quantify LDA performance in the past, it does not reflect human interpretability of the produced topics.

To mitigate these problems, this study proposed a new and enhanced topic modeling technique called LDA-GSAPO Latent Dirichlet Allocation with Gibbs Sampling, Perplexity Optimization, and Topic Coherence Integration. The integration of the C\_V and C\_P coherence metrics was key in ensuring that topics were not only statistically accurate but also semantically meaningful. This enabled a more interpretable and actionable form of sentiment analysis that is tailored to customer satisfaction feedback.

The model was subsequently deployed in an online setting that facilitated real-time examination of uploaded datasets through a basic web interface. Clients were able to post text data (e.g., CSV or JSON files) for analysis, whereby the expanded model generated coherent topics along with sentiment labels. The system was assessed based on both quantitative performance metrics (perplexity, coherence metrics) and qualitative user satisfaction metrics gathered through usability questionnaires.

Key findings include:

1. A considerable decrease in topic perplexity, resulting in greater predictive accuracy of the model.
2. Increasing topic coherence and interpretability of text-extracted themes.
3. Competent processing of brief and informal text messages.
4. An extremely usable and accessible web-based system with high user satisfaction levels.

## Recommendations

According to the research results, some suggestions can be proposed for future research and applications:

1. Broader Dataset Integration

Future research applications need to examine the inclusion of multi-domain datasets (e.g., finance, health, education) in order to ascertain the generalizability of the LDA-GSAPO model to customer reviews of retail or e-commerce and ensure.

1. Multi-language Support

Making the model capable of handling multilingual datasets would render it more relevant at the global level, especially for companies working in multilingual settings.

1. Incorporation of Real-Time Learning

The model can be augmented with online learning capabilities where it learns ongoing from new data, updating topics dynamically rather than being retrained from the beginning.

1. User Analytics and Feedback Loop

Utilizing user analytics to monitor how business stakeholders utilize the system can provide valuable feedback on refining interface design and feature applicability.

## Conclusion

In this research, a robust and semantically rich sentiment analysis model the LDA-GSAPO model was custom-designed for customer satisfaction analysis. Through the incorporation of coherence measures (C\_V and C\_P) into the LDA pipeline and leveraging the benefit of Gibbs Sampling and Perplexity Optimization, the model surmounts the conventional limitations of incoherent topic creation and statistical overfitting in regular LDA.

The enhanced model was effectively applied in a web-based environment, demonstrating that it is technically sound and simple to use. The key concepts provided by the system are more precise, concise, and easy to comprehend, which makes it a valuable asset for companies aiming to acquire significant information from customer feedback.

In summary, the LDA-GSAPO model bridges the gap between statistical performance and interpretability, enabling superior theme and sentiment analysis. It contributes significantly to the natural language processing and customer experience management fields and opens up new avenues for future research and business application.

References

Birjali, M., Kasri, M., & Beni-Hssane, A. (2021a). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, *226*, 107134. https://doi.org/10.1016/J.KNOSYS.2021.107134

Birjali, M., Kasri, M., & Beni-Hssane, A. (2021b). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, *226*. https://doi.org/10.1016/j.knosys.2021.107134

Campagnolo, J. M., Duarte, D., & Bianco, G. D. (2022). Topic Coherence Metrics: How Sensitive Are They? *Journal of Information and Data Management*, *13*(4). https://doi.org/10.5753/JIDM.2022.2181

Chauhan, U., & Shah, A. (2022). Topic Modeling Using Latent Dirichlet allocation: A Survey. *ACM Computing Surveys*, *54*(7). https://doi.org/10.1145/3462478/SUPPL\_FILE/CHAUHAN.ZIP

Cuevas, E., Barba-Toscano, O., Escobar, H., Zaldívar, D., & Rodríguez-Vázquez, A. (2024). An initialization approach for metaheuristic algorithms by using Gibbs sampling. *Mathematics and Computers in Simulation*, *225*, 586–606. https://doi.org/10.1016/J.MATCOM.2024.05.010

Dunn, W. L., & Shultis, J. K. (2023). Markov Chain Monte Carlo. *Exploring Monte Carlo Methods*, 189–254. https://doi.org/10.1016/B978-0-12-819739-4.00014-7

Edwards, C. (2021). The best of NLP. *Communications of the ACM*, *64*(4), 9–11. https://doi.org/10.1145/3449049;WEBSITE:WEBSITE:DL-SITE;TAXONOMY:TAXONOMY:ACM-PUBTYPE;PAGEGROUP:STRING:PUBLICATION

Fanni, S. C., Febi, M., Aghakhanyan, G., & Neri, E. (2023). *Natural Language Processing*. 87–99. https://doi.org/10.1007/978-3-031-25928-9\_5

Hamed, S., Ezzat, M., & Hefny, H. (2020). A Review of Sentiment Analysis Techniques. *International Journal of Computer Applications*, *176*(37), 20–24. https://doi.org/10.5120/ijca2020920480

Hanny, D., & Resch, B. (2024). Clustering-Based Joint Topic-Sentiment Modeling of Social Media Data: A Neural Networks Approach. *Information 2024, Vol. 15, Page 200*, *15*(4), 200. https://doi.org/10.3390/INFO15040200

Hoyle, A., Goel, P., Hian-Cheong, A., Peskov, D., Boyd-Graber, J., & Resnik, P. (2021). Is Automated Topic Model Evaluation Broken? The Incoherence of Coherence. *Advances in Neural Information Processing Systems*, *34*, 2018–2033.

Huang, Y., Wang, R., Huang, B., Wei, B., Zheng, S. L., & Chen, M. (2021). Sentiment Classification of Crowdsourcing Participants’ Reviews Text Based on LDA Topic Model. *IEEE Access*, *9*, 108131–108143. https://doi.org/10.1109/ACCESS.2021.3101565

Karras, C., Karras, A., Tsolis, D., Giotopoulos, K. C., & Sioutas, S. (2022a). Distributed Gibbs Sampling and LDA Modelling for Large Scale Big Data Management on PySpark. *7th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference, SEEDA-CECNSM 2022*. https://doi.org/10.1109/SEEDA-CECNSM57760.2022.9932990

Karras, C., Karras, A., Tsolis, D., Giotopoulos, K. C., & Sioutas, S. (2022b). Distributed Gibbs Sampling and LDA Modelling for Large Scale Big Data Management on PySpark. *7th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference, SEEDA-CECNSM 2022*. https://doi.org/10.1109/SEEDA-CECNSM57760.2022.9932990

Koltcov, S., Surkov, A., Filippov, V., & Ignatenko, V. (2024). Topic models with elements of neural networks: investigation of stability, coherence, and determining the optimal number of topics. *PeerJ Computer Science*, *10*. https://doi.org/10.7717/peerj-cs.1758

Kumar, V. V., Raghunath, K. M. K., Muthukumaran, V., Joseph, R. B., Beschi, I. S., & Uday, A. K. (2022). Aspect based sentiment analysis and smart classification in uncertain feedback pool. *International Journal of System Assurance Engineering and Management*, *13*(1), 252–262. https://doi.org/10.1007/S13198-021-01379-2/FIGURES/5

Ling, J., Li, X., Zhang, Q., & Shen, Y. (2025). Improved Latent Dirichlet Allocation (LDA) Method for Rock Tunnel Design by Knowledge Classification. *Springer Series in Geomechanics and Geoengineering*, 68–76. https://doi.org/10.1007/978-3-031-76528-5\_7/TABLES/2

Lo, K., Jin, Y., Tan, W., Liu, M., Du, L., & Buntine, W. (2021). Transformer over Pre-trained Transformer for Neural Text Segmentation with Enhanced Topic Coherence. *Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021*, 3334–3340. https://doi.org/10.18653/v1/2021.findings-emnlp.283

Madzík, P., & Falát, L. (2022). State-of-the-art on analytic hierarchy process in the last 40 years: Literature review based on Latent Dirichlet Allocation topic modelling. In *PLoS ONE* (Vol. 17, Issue 5 May). https://doi.org/10.1371/journal.pone.0268777

Mardones-Segovia, C., Wheeler, J. M., Choi, H.-J., Wang, S., Cohen, A. S., Mardones-Segovia, C., Wheeler, J. M., Choi, H.-J., Wang, S., & Cohen, A. S. (n.d.). Model Selection for Latent Dirichlet Allocation In Assessment Data. In *Psychological Test and Assessment Modeling* (Vol. 65, Issue 1).

Mashaabi, M., Alotaibi, A., Qudaih, H., Alnashwan, R., & Al-Khalifa, H. (2022). Natural Language Processing in Customer Service: A Systematic Review. *ArXiv*. https://arxiv.org/pdf/2212.09523

Mewada, A., Dewang, R. K., Goldar, P., & Maurya, S. K. (2023). SentiBERT: A Novel Approach for Fake Review Detection Incorporating Sentiment Features with Contextual Features. *ACM International Conference Proceeding Series*, 230–235. https://doi.org/10.1145/3607947.3607991;SUBPAGE:STRING:ABSTRACT;TOPIC:TOPIC:CONFERENCE-COLLECTIONS>IC3;CSUBTYPE:STRING:CONFERENCE

Mishra, M. (2024). A holistic review of customer experience research: topic modelling using BERTopic. *Marketing Intelligence and Planning*, *43*(4), 802–820. https://doi.org/10.1108/MIP-09-2023-0457/FULL/XML

Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, *11*(1), 1–19. https://doi.org/10.1007/S13278-021-00776-6/METRICS

Pais, S., Cordeiro, J., & Jamil, M. L. (2022). NLP-based platform as a service: a brief review. *Journal of Big Data*, *9*(1), 1–26. https://doi.org/10.1186/S40537-022-00603-5/FIGURES/3

Pandya, S., & Mehta, P. (2020). A Review On Sentiment Analysis Methodologies, Practices And Applications. *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, *9*, 2. www.ijstr.org

Patwardhan, N., Marrone, S., & Sansone, C. (2023). Transformers in the Real World: A Survey on NLP Applications. *Information 2023, Vol. 14, Page 242*, *14*(4), 242. https://doi.org/10.3390/INFO14040242

Raghunathan, N., & Saravanakumar, K. (2023). Challenges and Issues in Sentiment Analysis: A Comprehensive Survey. *IEEE Access*, *11*. https://doi.org/10.1109/ACCESS.2023.3293041

Rahimi, H., Mimno, D., Hoover, J. L., Naacke, H., Constantin, C., & Amann, B. (2023). Contextualized Topic Coherence Metrics. *EACL 2024 - 18th Conference of the European Chapter of the Association for Computational Linguistics, Findings of EACL 2024*, 1760–1773. https://arxiv.org/abs/2305.14587v1

Rodríguez-Ibánez, M., Casánez-Ventura, A., Castejón-Mateos, F., & Cuenca-Jiménez, P. M. (2023). A review on sentiment analysis from social media platforms. *Expert Systems with Applications*, *223*, 119862. https://doi.org/10.1016/J.ESWA.2023.119862

Sharma, D., Sabharwal, M., Goyal, V., & Vij, M. (2020). Sentiment Analysis Techniques for Social Media Data: A Review. *Advances in Intelligent Systems and Computing*, *1045*, 75–90. https://doi.org/10.1007/978-981-15-0029-9\_7

Stein, F., Schein, J., Bayrle-Kelso, N., & Bartel, M. (2025). *Perceived Innovativeness: An Slda-Based Approach*. https://doi.org/10.2139/SSRN.5145166

Tekin, Y. (2024a). Initialization in Gibbs Sampling Implementation of LDA. *32nd IEEE Conference on Signal Processing and Communications Applications, SIU 2024 - Proceedings*. https://doi.org/10.1109/SIU61531.2024.10600919

Tekin, Y. (2024b). Initialization in Gibbs Sampling Implementation of LDA. *32nd IEEE Conference on Signal Processing and Communications Applications, SIU 2024 - Proceedings*. https://doi.org/10.1109/SIU61531.2024.10600919

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review 2022 55:7*, *55*(7), 5731–5780. https://doi.org/10.1007/S10462-022-10144-1

Xu, Q. A., Chang, V., & Jayne, C. (2022). A systematic review of social media-based sentiment analysis: Emerging trends and challenges. *Decision Analytics Journal*, *3*, 100073. https://doi.org/10.1016/J.DAJOUR.2022.100073

Yu, D., & Xiang, B. (2023). Discovering topics and trends in the field of Artificial Intelligence: Using LDA topic modeling. *Expert Systems with Applications*, *225*, 120114. https://doi.org/10.1016/J.ESWA.2023.120114