Hybrid vs. Unsupervised: A Comparative Study of Topic Modeling Approaches for Review Classification

Pranay Kumar Reddy Pinninti
p.pinninti@ufl.edu

Tanuj Venkata Satya Sridhar Karuturi
karuturi.t@ufl.edu

Sreekar Reddy Nathi
sreekarreddnathi.j@ufl.edu

Jyothi Santoshi Karuturi
karuturi.j@ufl.edu

Abhinav Mandala
abhinav.mandala@ufl.edu

Abstract

In today's online commerce environment, consumer decision-making heavily depends on user reviews, which provide critical insights into product experiences and satisfaction. However, the vast amount of unstructured textual data presents significant challenges, necessitating effective categorization methods. Topic modeling, particularly through techniques like Latent Dirichlet Allocation (LDA), has emerged as a fundamental approach for extracting latent themes from such data. Although crucial, LDA often struggles with coherence and relevance in domain-specific contexts. This paper proposes a novel hybrid approach that combines unsupervised and supervised learning techniques to extract latent topics and subsequently classify Amazon reviews into these topics. By analyzing over 20,000 reviews from the electronics category, this study evaluates the hybrid model against traditional unsupervised LDA using coherence metrics like C_V and U_Mass, which are widely accepted for assessing the performance of topic modeling techniques. Our findings advance the methodology for deriving meaningful insights from unstructured data, improving the online shopping experience and guiding future research in review classification and Categorization.[1]

1 Introduction

In the digital era, online shopping has become a prevalent choice for consumers worldwide, resulting in an abundance of user-generated content in the form of product reviews. These reviews are pivotal as they influence purchasing decisions and enhance consumer trust. However, the sheer volume of reviews can overwhelm potential buyers and impede the decision-making process. Effective review classification not only categorizes content into coherent topics, facilitating easier navigation, but also enhances the shopping experience by highlighting the most relevant reviews based on consumer interests. Natural Language Processing (NLP) emerges as a critical tool in this context, where topic modeling (Xia et al., 2019), a fundamental NLP technique, is employed to extract meaningful patterns and themes from the extensive amounts of unstructured text presented by product reviews. By organizing these reviews into distinct topics, topic modeling allows for a structured analysis to understand consumer sentiments and preferences better, thus simplifying data and aiding in more informed decision-making.

Traditional topic modeling methods, such as Latent Dirichlet Allocation (LDA), typically operate within an unsupervised learning framework and face several limitations(Andrzejewski et al., 2009). These methods often rely on a bag-of-words approach, which can inadvertently include irrelevant or ambiguous terms in the topics generated. Furthermore, the inherent randomness of these algorithms can result in variability in topic quality between different iterations, complicating the task of achieving consistent and interpretable outcomes from the modeling process.

Our study aims to explore the efficacy of hybrid topic modeling techniques, combining both unsupervised and supervised learning methodologies, in the classification of product reviews into distinct topics. Our research integrates the flexibility of unsupervised learning with the precision of supervised learning to understand and organize user reviews more effectively than traditional methods. At the heart of our approach is the use of Correlation Explanation (CorEx)(Reing et al., 2016), a semi-supervised learning model that leverages both labeled and unlabeled data to enhance topic modeling. Unlike purely unsupervised methods, CorEx allows for the incorporation of domain-specific knowledge through anchor words that guide the topic modeling process, potentially uncovering

[1] https://github.com/PRANAY8055/topicmodeling-amazonreviews.git

nuanced topics that are often overlooked. This method is complemented by traditional supervised techniques, which refine these topics by learning from labeled instances, thereby improving the overall coherence and relevance of the topics generated.

We employed two topic coherence metrics (Rosner et al., 2014), C_V and U_mass, to quantitatively evaluate our hybrid model against traditional unsupervised models. C_V measures semantic similarity within topics, indicating their interpretability, while U_mass assesses the model's use of contextual relationships. These metrics provide a holistic view of the model's performance. Our results show that the hybrid model significantly outperforms traditional unsupervised models in terms of topic coherence and accuracy. The hybrid approach produces more relevant and distinct topics, suggesting a promising direction for future research in review classification and text analysis.

2 Related Work

The importance of topic modeling (George and Birla, 2018) in understanding and classifying large volumes of text has been well documented across various studies (Brookes and McEnery, 2019), providing valuable insights into different methodologies and techniques. Such investigations are fundamental for navigating the complexities of document classification and thematic analysis, and they establish a critical foundation for comparative studies.

Over the years, Latent Dirichlet Allocation (LDA) has been prominently featured across a diverse array of applications, attesting to its versatility and effectiveness in text analysis. For instance, (Acheadeth et al., 2022) highlight the application of LDA in e-commerce for sorting online product reviews into relevant and irrelevant categories based on underlying thematic structures. This use of LDA not only sharpens the accuracy of review classification but also facilitates crucial pre-processing steps for more intricate text analysis tasks, such as sentiment analysis. The ability of LDA to extract meaningful insights from a broad spectrum of text offers a solid basis for exploring advanced hybrid topic modeling techniques.

Despite its widespread application, traditional topic modeling methods like LDA and Probabilistic Latent Semantic Analysis (PLSA) often face significant challenges, especially when processing short texts (Qiang et al., 2017) from niche retail categories or unique consumer feedback contexts. These methods struggle to integrate necessary contextual cues for a nuanced understanding, heavily relying on sparse word co-occurrence data (Van Rijsbergen, 1977), which severely limits their effectiveness. The limitations of LDA and PLSA in managing concise textual content have been extensively examined, underscoring their inadequacies in producing consistently meaningful and interpretable outcomes. Moreover, the complexity of configuring LDA with precise model assumptions and hyperparameter settings further complicates its application, particularly when attempting to incorporate human inputs directly into the model.

In response to these challenges, the Correlation Explanation (CorEx) model presents a robust alternative. Developed by (Gallagher et al., 2017), CorEx does not rely on an underlying generative model. Instead, it utilizes an information-theoretic framework to learn topics, significantly enhancing the flexibility to include domain-specific knowledge through the use of anchor words. This method markedly improves topic separability and relevance, addressing critical gaps left by traditional techniques and providing a more adaptable solution for complex topic modeling challenges.

Our study seeks to build upon these foundational insights by comparing the performance and applicability of both hybrid and unsupervised topic modeling approaches, particularly in the context of review classification. By integrating the strengths of unsupervised learning with the precision of supervised methodologies, our research aims to advance the current understanding and implementation of topic modeling, thereby offering more refined tools for extracting and interpreting user-generated content on e-commerce platforms. This synthesis of related works not only reinforces the necessity for innovation in topic modeling but also sets the stage for future advancements that could further enhance the interpretability and utility of extracted topics in real-world applications.

3 Proposed Approach

3.1 Datasets

Selecting an appropriate dataset[2] was crucial to ensure the relevance and effectiveness of any analysis, as it forms the cornerstone for training and evaluating the developed model. Given the inherent variability in the topics discussed within different

[2] https://www.kaggle.com/datasets/magdawjcicka/amazon-reviews-2018-electronics/data

product categories, we recognized the importance of focusing on a specific category for our study. Consequently, we chose to target the electronics category due to its prominence and the likelihood of encountering diverse and nuanced themes within this domain. We opted for an Amazon reviews dataset available on Kaggle. This dataset encompassed a substantial collection of over 20,000 reviews, meeting our criteria for size, relevance, and accessibility. The abundance of reviews on Amazon provides us with a rich and diverse dataset, enabling us to explore various aspects of review classification and topic modeling within the context of the electronics category.

3.2 Methodology

The paper introduces a hybrid approach to topic modeling for review classification, intending to assess its efficacy compared to the traditional unsupervised method, Latent Dirichlet Allocation (LDA). The methodology commences with the collection of an Amazon review dataset from the electronics category, serving as the foundation for subsequent analysis. Following data acquisition, a preprocessing stage is initiated to refine the dataset for further evaluation.

Latent Dirichlet Allocation (LDA) infers topics from a collection of texts by assuming that documents are mixtures of topics and that topics are distributions over words. The implementation of LDA, after preprocessing the text, involves transforming it into a document-term matrix using the bag-of-words model and configuring the model with parameters like the number of topics and iterations using the Gensim library. Key hyperparameters such as Alpha and Eta are adjusted to optimize the model's performance, addressing the sparsity of document-topic and topic-word distributions, respectively.

The hybrid method comprises three distinct stages. First, unsupervised term frequency-inverse document (TF-IDF) (Yantao et al., 2007) is employed to identify high-frequency words within the reviews. This initial step provides valuable insights into the most prevalent terms. The second stage involves leveraging labeled data through the incorporation of anchor words within the Correlation Explanation (CorEx) framework. Unlike traditional models like LDA, which assume a generative model for document creation, CorEx focuses on learning topics that are maximally informative about the documents themselves. By identifying significant correlations within the dataset, CorEx facilitates the extraction of meaningful topics without making prior assumptions about topic generation. Moreover, the flexibility of CorEx allows for the seamless integration of domain knowledge through the inclusion of anchor words. These anchor words serve as guiding points, enhancing the interpretability and relevance of extracted topics. In the final stage, the model is trained using labeled data to classify each review under specified topics. When new data is provided, the model predicts the topic classification for each review. Given potential overlaps in topics, the model adopts a nuanced approach to ensure accurate categorization across relevant themes.

The tuned LDA model serves as a baseline for our research, providing a reference point for quantifying improvements brought about by our hybrid approach.The effectiveness of both the LDA and hybrid models is further validated through the use of coherence metrics such as C_V and U_Mass. Figure 1 represents the flow diagram of the methodology.

3.3 Experiment

This section elucidates the practical implementation of the three phases outlined in the methodology section, encompassing dataset collection, data preprocessing, and hybrid topic modeling.

3.3.1 Dataset Collection

We've obtained a robust Amazon review dataset from Kaggle, specifically focusing on the electronics category, comprising an extensive collection of 20,000 reviews. This dataset offers a holistic perspective on consumer sentiments and experiences, providing invaluable insights into their interactions with electronic products.

3.3.2 Data Preprocessing

In the data preprocessing phase, several steps were undertaken to clean and prepare the collected text data for further analysis.

Lowercasing - All characters were transformed to lowercase using the ‘lower()‘ method to ensure uniformity.

Cleaning Text - Unnecessary details such as punctuation were removed from the text.

Tokenization - The text was tokenized by splitting the text into individual words based on whitespace.

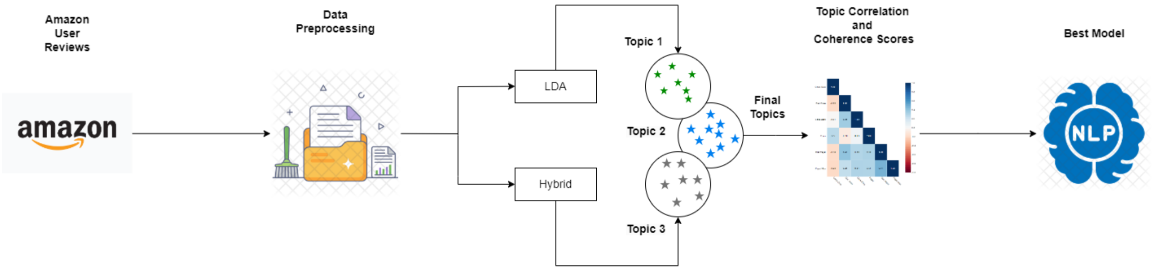


Figure 1: Flow Diagram

Stopwords Removal - Stopwords, which add little meaning to sentences, were removed to enhance the relevance of the text.

Lemmatization - The lemmatization process was applied to group together the various inflected forms of words, facilitating further analysis.
The result of the preprocessing phase was a cleaned and standardized text dataset ready for further analysis.

3.3.3 LDA Topic Modeling

LDA(Eletter et al., 2021) is a probabilistic model that infers topics from a collection of texts by assuming that documents are mixtures of topics and that topics are distributed over words. Our study employs LDA to establish a benchmark against which we can evaluate our advanced hybrid topic modeling approach.

The implementation of Latent Dirichlet Allocation (LDA) begins with data pre-processing. Following pre-processing, the text is transformed into a document-term matrix using the bag-of-words model. This matrix format is essential for the subsequent topic modeling with LDA. In configuring our model, we utilize the Gensim library to set up a basic Latent Dirichlet Allocation (LDA) model tailored to our dataset's specific characteristics. The model's key parameters include the Number of Topics, which determines the distinct topics to be extracted from the documents, and the Number of Iterations, which sets the convergence criteria to ensure the model sufficiently captures the word distributions. Additionally, the hyperparameters Alpha and Eta are critical in fine-tuning the model's performance; Alpha regulates the sparsity of the document-topic distributions, with higher values allowing documents to encompass more topics, while Eta controls the topic-word distributions, with higher values permitting topics to include more words. These parameters are essential for optimizing the LDA model to achieve precise and relevant topic identification.

To optimize the LDA model for our specific needs, we adjust the alpha and eta hyperparameters. This tuning process is critical as it enhances the model's adaptability and accuracy in topic classification by aligning the model assumptions more closely with the inherent structure of our data.

By setting a baseline with this tuned LDA model, we create a foundation for assessing the advancements enabled by our proposed hybrid topic modeling approach. This baseline allows us to quantify improvements in topic discovery and classification, highlighting the benefits of integrating supervised learning techniques into the topic modeling process. To validate the model's topic coherence and relevance, we employ widely recognized coherence metrics such as C_V and U_Mass. These metrics measure the semantic similarity and consistency of the topics generated by the LDA model, providing a quantitative basis for the initial evaluation of the model's effectiveness.

3.3.4 Hybrid Topic Modeling

The implementation of hybrid topic modeling encompasses three key stages, that include unsupervised learning, defining anchors, and supervised learning. Figure 2 illustrates the sequential flow of the approach.

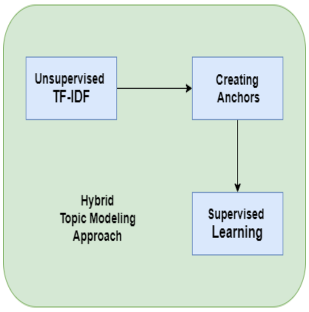


Figure 2: Hybrid Topic Modeling Approach

Unsupervised Learning

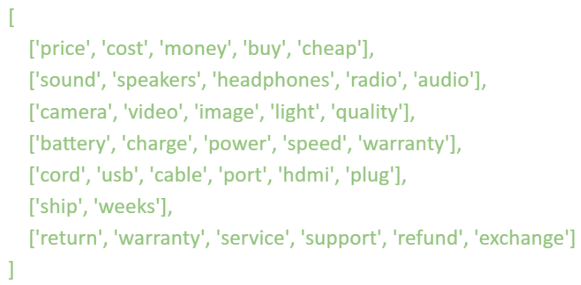
The pre-processed text undergoes tokenization using the TF-IDF vectorizer. This vectorization process transforms the text into a numerical representation, assigning higher scores not only to words with greater frequency within the reviews but also considering the rarity of the words, thus capturing their importance in the context of the entire corpus. Also, the TF-IDF vectorizer converts the text into a matrix representation, where each row corresponds to a document (review) and each column represents a unique term in the corpus. Additionally, a list of top 300 words and their corresponding TF-IDF scores are generated, providing insights into the most significant terms within the corpus, enabling a deeper understanding of the underlying themes present in the text data.

Defining Anchors

Anchors play a crucial role in the topic modeling process by linking sets of words to specific topics, thus providing structure and focus to the analysis. They serve to emphasize particular themes or subjects within those topics, guiding the identification of relevant terms associated with each theme. In Figure 3, words like "sound," "speakers," "headphones," "audio," and "radio" are grouped together and associated with the sound-related topic. These anchor words are carefully selected based on their relevance to the domain being studied, ensuring that they accurately represent the key concepts within each topic. By incorporating anchor words into the topic modeling framework, the CorEx model is provided with valuable guidance on how to allocate terms to different topics. This process helps to ensure that the resulting topics are coherent and meaningful, as the model can prioritize words that are most indicative of each topic's theme.

Supervised Learning

In the supervised learning phase, each review was meticulously assigned topics relevant to its content based on predefined anchors, marking a crucial step in our exploration of latent topics within the pre-processed text data. Our model was configured with seven latent topics, serving as foundational elements for subsequent analysis. These topics structure the analysis and organize the semantic content of the corpus. Utilizing the TF-IDF document-word matrix as input, the Corex model underwent fitting of the pre-processed text data. To ensure a robust evaluation of the CorEx model's performance, we implemented a training-test split of 80-20 percent. This setup utilized 80% of the data for training to optimize the model's learning processes and 20% for testing to accurately assess its predictive performance on unseen data. Through this process, the model adeptly learned the intricate relationships between words and documents, enabling it to discern patterns and unveil latent topics that encapsulate the underlying themes within the corpus.



```
[
  ['price', 'cost', 'money', 'buy', 'cheap'],
  ['sound', 'speakers', 'headphones', 'radio', 'audio'],
  ['camera', 'video', 'image', 'light', 'quality'],
  ['battery', 'charge', 'power', 'speed', 'warranty'],
  ['cord', 'usb', 'cable', 'port', 'hdmi', 'plug'],
  ['ship', 'weeks'],
  ['return', 'warranty', 'service', 'support', 'refund', 'exchange']
]
```

Figure 3: Defining Anchors

At the core of the Corex framework is use of specific anchor words, marking advanced topic modeling. These carefully chosen anchor words act like road signs, guiding the topic modeling process toward a clearer and easier-to-understand outcome. By using these anchor words wisely, our model gained valuable insights into the detailed meaning of the text, making the topics we extracted more understandable and relevant than ever before. A key part of training our Corex model was considering the strength of these anchor words. This helped us understand how much each anchor word influenced the direction of the topic modeling process. By adjusting these strengths carefully, researchers could control how much the specific topics reflected the context of the text, creating a topic model that made more sense and had deeper meaning.

4 Results and Analysis

In this section, we present the results of our hybrid topic modeling approach using domain knowledge. Our analysis focuses on the coherence scores as shown in the Figure 4 and includes a comparison with traditional LDA models, examining the topics and words generated by our model. The effectiveness of our methods was assessed using two evaluation metrics, C_V coherence and U_mass coherence. The C_V coherence score evaluates the semantic similarity between the top N words within a topic, while the U_mass coherence score is based on the log-likelihood ratio of the corpus, assessing the logical consistency of the topics and their words.

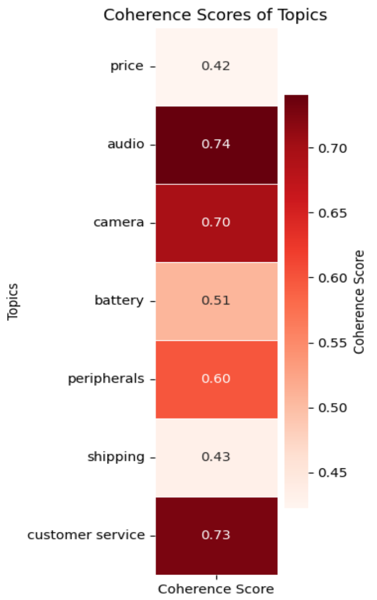


Figure 4: Topic wise Coherence Score for the topics generated by Hybrid model

4.1 Comparison with LDA

Our Hybrid model achieved an average coherence score of 0.59, indicating a high degree of semantic similarity among the top words within each topic and suggesting that the topics are meaningful and interpretable. For example, in a topic related to 'customer service', words like 'support', 'helpful', and 'responsive' appeared together frequently, reflecting a coherent theme recognized by consumers.

In contrast, a baseline LDA model applied to the same dataset under similar conditions yielded a lower average coherence score of 0.48, whereas the tuned LDA model achieved a coherence score of 0.51. This comparison underlines the effectiveness of using anchored words in our hybrid model, which guides the topic modeling process to focus on more specific and relevant themes, thereby enhancing coherence.These results are summarized in Table 1."

Another factor to consider is that our model generate topics which are build around the anchor words given by the user based on their specifications or needs and still acheive equal or higher coherence scores than baseline LDA models. Other than this the quality of the topics generated by LDA is very low and the topic words are not relevant to a particular topic, which can be observed from the top 10 topic words from each topic as shown in Table 2.

Model	**C_V score**	**U_mass score**
Basic LDA Model	0.48	-2.29
Tuned LDA Model	0.52	-2.37
Hybrid Model	0.59	0.73

Table 1: Comparing Coherence Scores: LDA vs. Hybrid Topic Modeling Methods

Topic	**Results of LDA (Unsupervised Learning)**	**Results of CorEx (Semi-Supervised Learning)**
Camera	camera, drive, card, video, picture, use, monitor, adapter, usb, ok	camera, quality, light, video, image, lens, poor, picture, focus, canon
Battery	work, battery, great, fit, charge, well, bag, charger, camera, perfect	charge, battery, power, warranty, speed, life, charger, batteries, supply, mah
Audio	sound, speaker, dont, player, music, radio, like, better, volume, money	sound, headphones, speakers, radio, audio, bass, ear, speaker, pair, music

Table 2: LDA vs Hybrid Topic Modeling Results

By examining similar topics which are generated using LDA and our model further illustrates its capability to capture key aspects of the data. For example, one prominent topic that emerged was centered around 'audio', with anchor words like 'sound', 'speakers' and 'audio' leading to the aggregation of related terms such as 'ear', 'bass', 'pair' and 'music', but when you observe a similar topic generated by LDA there are words like

’like’, ’better’, ’dont’, ... which are not related to ’audio’. This not only demonstrates the model’s responsiveness to the anchored guidance but also its utility in extracting commercially relevant themes that are valuable for product development and customer satisfaction analyses.

4.2 New Topic Discovery

We can leverage the hybrid model to find undiscovered topics by providing relevant anchor words, we selected the anchor words based on the required topics and most frequent words returned by TF-IDF. Table 3 displays anchor and generated topic words for the "Customer Support" topic.

Topic	Anchor words	Topic words generated
Customer Support	service, return, exchange, support	return, service, support, warranty, refund, exchange, customer, tech, email, repair

Table 3: Discovering hidden topics using anchor words

Canon does not provide any warranty coverage!
Nice idea, but it doesn’t work and the customer service is weak
Extremely poor fit. Bought them for my workout. Just wouldn’t stay. I returned it.
I emailed support about this and they were very rude and rather than answering my questions about the sorting, they suggested I return it if I’m not happy with it.
helpful tech support suggest wait hours replace batteries. radio working well

Table 4: Reviews categorized under ’Customer Support’ topic

In Table 4, a selection of reviews categorized under the ’Customer Support’ topic is presented for reference.

In conclusion, our results vividly demonstrate the superior performance of our hybrid topic modeling approach using Anchored CorEx over traditional LDA models. Notably, the hybrid model achieved significantly higher coherence scores, exemplifying its capacity to generate more semantically cohesive topics that consist of words that logically cohere together. This advantage is largely attributed to the strategic use of anchor words, which guide the discovery process towards more relevant and interpretable topics. Moreover, our model has successfully uncovered underlying topics that were not identified by LDA, revealing new dimensions of the data that hold practical significance for domain-specific applications. These findings underscore the efficacy of incorporating domain knowledge (Anchor Words) into topic modeling, promising enhanced topic quality that can pivotally influence product development and customer satisfaction strategies.

5 Discussion

In comparative study between hybrid and unsupervised topic modeling methods, specifically focusing on review classification, highlights several critical insights. The hybrid approach, incorporating supervised elements such as Correlation Explanation (CorEx), significantly improves topic coherence and relevance over traditional unsupervised methods like Latent Dirichlet Allocation (LDA). This enhanced coherence is evident in the higher coherence scores achieved by the hybrid model, indicating more meaningful and interpretable topics. These improvements are crucial for effectively parsing consumer feedback in e-commerce, where precise and actionable insights into product reviews can directly influence business strategies and customer satisfaction.

Moreover, this study also reveals that while hybrid modeling offers substantial benefits, it introduces complexities such as the need for initial labeled data. The selection of anchor words, a key element in the hybrid approach, requires domain expertise and careful consideration to ensure they are effectively guiding the topic modeling process. Despite these challenges, the benefits of hybrid topic modeling, particularly in terms of producing relevant and coherent topics, justify its consideration for advanced analytical tasks in diverse fields like e-commerce where understanding consumer feedback is crucial.

6 Conclusion

In this study, we demonstrated that our hybrid topic modeling approach using the Anchored CorEx model significantly outperforms traditional unsupervised LDA methods in categorizing themes within Amazon electronics product reviews. The integration of both supervised and unsupervised techniques, along with the strategic use of anchor words, improves the interpretability and relevance of topics to specific domains. This approach not only aligns the topics more closely with domain-specific needs but also facilitates more efficient organization of information, critical for businesses seeking actionable insights from consumer feedback.

The hybrid model's ability to generate semantically cohesive topics has been quantitatively validated through higher coherence scores compared to those achieved by traditional LDA. This indicates a deeper, more precise understanding of the text, enabling the extraction of nuanced themes that conventional methods often overlook.

In conclusion, our study paves the way for advanced topic modeling techniques that significantly improve the interpretability and utility of extracted topics, offering a substantial upgrade over traditional methods and providing a robust framework for future enhancements. This method promises to significantly influence product development and customer satisfaction strategies, offering a more customer-centric analysis that could drive business decisions.

7 Limitations and Future Work

A primary limitation of our current research is its focus solely on the electronics category. While this provides detailed insights into a significant sector of e-commerce, it restricts the applicability of our findings to other product categories, each characterized by distinct topics and consumer concerns that may not be adequately addressed by a model trained only on electronics-related reviews.

Building on the insights from our research, numerous opportunities for future development emerge to broaden the scope and enhance the effectiveness of our hybrid topic modeling approach. Extending our model to cover more product categories beyond electronics will enable us to assess its adaptability and refine it across a wider range of consumer goods, thereby capturing a more diverse set of consumer feedback. Additionally, employing machine learning algorithms to automate the selection of anchor words could greatly improve the model's scalability and accuracy.

Additionally, future iterations could explore the integration of sentiment analysis within the hybrid model to not only categorize topics but also gauge the emotional tone within them. This sentiment-enhanced topic modeling could provide even richer insights, enabling companies to tailor their strategies more effectively and connect with their customers on a deeper emotional level.

References - (All Team Members)

Lay Acheadeth, Nunung Nurul Qomariyah, and Misa M. Xirinda. 2022. Utilizing topic modelling in customer product review for classifying baby product. In *2022 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*, pages 52–57.

David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. Incorporating domain knowledge into topic modeling via dirichlet forest priors. *Proceedings of the ... International Conference on Machine Learning. International Conference on Machine Learning*, 382:25–32.

Gavin Brookes and Tony McEnery. 2019. The utility of topic modelling for discourse studies: A critical evaluation. *Discourse Studies*, 21(1):3–21.

Shorouq Fathi Eletter, Kholoud Ibrahim AlQeisi, and Ghaleb Awad Elrefae. 2021. The use of topic modeling in mining customers' reviews. In *2021 22nd International Arab Conference on Information Technology (ACIT)*, pages 1–4.

Ryan J. Gallagher, Kyle Reing, David Kale, and Greg Ver Steeg. 2017. Anchored correlation explanation: Topic modeling with minimal domain knowledge. *Transactions of the Association for Computational Linguistics*, 5:529–542.

Laya Elsa George and Lokendra Birla. 2018. A study of topic modeling methods. In *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 109–113.

Jipeng Qiang, Ping Chen, Tong Wang, and Xindong Wu. 2017. *Topic Modeling over Short Texts by Incorporating Word Embeddings*, page 363–374. Springer International Publishing.

Kyle Reing, David C Kale, Greg Ver Steeg, and Aram Galstyan. 2016. Toward interpretable topic discovery via anchored correlation explanation. *arXiv preprint arXiv:1606.07043*.

Frank Rosner, Alexander Hinneburg, Michael Röder, Martin Nettling, and Andreas Both. 2014. Evaluating topic coherence measures.

Cornelis Joost Van Rijsbergen. 1977. A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of documentation*, 33(2):106–119.

Linzhong Xia, Dean Luo, Chunxiao Zhang, and Zhou Wu. 2019. A survey of topic models in text classification. In *2019 2nd International Conference on*

Artificial Intelligence and Big Data (ICAIBD), pages 244–250.

Zhou Yantao, Tang Jianbo, and Wang Jiaqin. 2007. An improved tfidf featurfe selection algorithm based on information entropy. In *2007 Chinese Control Conference*, pages 312–315.