Automating Customer Insights: Optimizing Product Review Analysis on Amazon

Shalini Tigga   
*B.Tech CSE*  
*Lovely Professional University*Phagwara, India  
[shalinitigga26@gmail.com](mailto:shalinitigga26@gmail.com)

*Abstract*— Amazon, a leading e-commerce platform with over 310 million active users, generates vast data, including 6.2 million product reviews—around 1-2% of users contribute. These reviews are crucial for insights into customer satisfaction, product quality, and preferences. However, manually analyzing this data is labor-intensive and costly. By optimizing and automating review analysis, Amazon can quickly derive valuable insights, enhance customer satisfaction, and improve decision-making.

Keywords— *Topic Modelling, Latent Dirichlet Allocation, Non-negative Matrix Factorization, Latent Semantic Analysis, BERT-based topic modeling*

# Introduction

The proliferation of computers and internet-accessible information has led to the digitization and archiving of tremendous volumes of data over the internet in e-commerce sites, news portals, and social networks. Reviews of products on Amazon, for example, provide a colossal data set with immense potential to unearth consumer sentiment, product quality, and preference. Searching and analysing such data has now become a great challenge. Our research aims at building a computational tool that will help effectively organize, search, and analyse Amazon review data, using topic modelling and text mining techniques as methods to gain an advanced understanding and extraction of meaningful insight from this wealth of information source.

# Background

## Topic Modelling

Topic modeling is one of the techniques used in Natural Language Processing (NLP) and machine learning, aimed at uncovering the hidden thematic structure within a collection of texts. The intention of topic modeling is detection of the embedded semantic structures within textual information, thereby allowing for optimal and insightful data grouping and assimilation by the user. The procedures of topic modeling have great significance within natural language processing as they can discover seminal or relevant information from arrayed contents. In certain applications, users want to recognize general themes in large data mining processes, so there is a need to develop automated knowledge extraction to support the cognitive dimensions in documentation. [1].

1. Topic modelling [10]

Fig. 1 illustrates the key steps of topic modeling, including the bag of words (BoW), model training, and model output. We first assume that there are N documents, V words, and K topics in a corpus.

## Latent Dirichlet Allocation

In the realm of natural language processing, Latent Dirichlet allocation (LDA) is a Bayesian network and generative statistical model that automatically learns topics from corpora of text. While the LDA is often thought of as a Bayesian topic model where observations (words) appear in documents and each word is assigned to a topic of the document and limited topics will be chosen from the available pool of topics, it stands in stark contrast to the previous system called probabilistic latent semantic analysis (pLSA), which substitutes the older pLSA's uniform Dirichlet prior for LDA. [2].

1. Latent Dirichlet Allocation [11]

## Non-negative Matrix Factorization

Non-negative Matrix Factorization (NMF) is a dimension reduction machine learning approach that decomposes a non-negative data matrix into two lower-dimensional, non-negative matrices. Given a data matrix V with mxn dimensions, NMF seeks to approximate it by means of two matrices W of size mxk and H of size kxn, where k is usually much smaller than m or n. Such a decomposition is useful to uncover the latent features in the data by constraining all elements of the matrices to be non-negative, which usually makes the results more interpretable. For example, according to text mining, NMF could uncover some topics within documents by representing each document as a non-negative linear combination of topic vectors. Thus, the interpretability and effectiveness of NMF in extracting meaningful features from high-dimensional data constitute probably the main elements making NMF quite popular for text and image analysis applications. [3].

## Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a technique in natural language processing that detects relationships between terms and documents in a large corpus of text. It stands upon Singular Value Decomposition (SVD), which breaks down a given term-document matrix into three matrices: U, Σ, and VT. Thus, it succeeds in representing a matrix with a reduced number of dimensions that keep a record of the hidden semantic organization of the data by putting together similar terms and documents in their space by means of a vector. In this way, inference on these hidden patterns would serve to address synonymy (different words that have roughly the same meaning) and polysemy (one word has different meanings), thus serving the purposes of document clustering, topic modeling, and information retrieval. This dimensionality reduction works towards removing noise and enhancing the model's generalizability leading it to perform a better and more accurate job in similarity assessment and topic identification. [4].

## BERT-based Topic Modelling

BERT-based topic modeling exploits BERT (Bidirectional Encoder Representations from Transformers), a deep learning model leveraged by Google for grasping the contextual framework of words in documents. Whereas, conventional topic modeling approaches for instance, LDA rely on the co-occurrence patterns of words, treating them as independent words, BERT captures semantic meaning based on the context in which it appears around words. This peculiarity allows BERT to be rare but very effective in identifying subtle topics embedded in text data, even when the term associations are not explicitly defined. [5].

BERT-based topic modeling takes the general approach of converting a document into contextual embeddings using BERT: a dense vector representation designed as a numerical form to capture the meaning of each document. Then the embeddings are clustered using any of the approaches such as K-Mean, Hierarchical Clustering, and UMAP (Uniform Manifold Approximation and Projection) for dimensionality reduction to bring similar documents together as a function of an underlying topic. A cluster is a distinct topic that may then be described with the going terms or phrases within this cluster. This approach generally produces more coherent and specific topics compared to traditional models, especially when context is key, such as in the understanding of social media posts and customer reviews analysis.

# Literature Review

David et al.'s [6] paper gives the presentation of an approach, the LDA, which models discrete datasets hierarchically using a Bayesian framework and is particularly targeted toward the analysis of text corpora. LDA treats documents as mixtures of topics whereby each topic is modeled as a mixture of topic probabilities. The authors developed efficient inference techniques, namely variational methods and an EM algorithm for parameter estimation, and showed that the LDA modeling framework can effectively be used for document modeling, text classification, and collaborative filtering. Comparisons of LDA with other existing models, such as mixture of unigrams and probabilistic Latent Semantic Indexing (LSI), are also carried out.

Lee et al. [7] explored parts-based versus holistic representation in perception, specifically with regard to object recognition. They provided a nonlinear non-negative matrix factorization algorithm to effectively learn parts of faces and semantic features from text, in stark contrast to traditional holistic representations of a linear principal component analysis. The authors stressed how the non-negativity constraint of NMF allows the use of additive combinations, permitting the emergent parts-based representations during neural network implementations, where firing rates of neurons and synaptic strengths are inherently non-negative.

The challenges of searching, managing, and exploring the huge volume of daily-generated digital text are discussed by Zhou et al. [8], introducing text mining and LDA, a probabilistic topic model. Experiments are conducted on Wikipedia articles to create a document topic model for enhancing article search and recommendation. The second experiment is conducted on users' tweets to create a user topic model, through which users' interests on Twitter are analyzed. The experiment process, ranging from data collection to preprocessing to model training, is documented thoroughly by the authors, in which their findings are said to be a potential computational tool for social and business research applications.

Elisabeth et al. (2022) improved the research of conspiracy theories by having multimodal topic modeling to analyze audio-visual content in German-language Telegram channels and diminish the shortcomings of conventional textual analysis. They examine a corpus of about 40,000 messages from 571 channels known for the propagation of conspiracy theories in October 2023 using the BERTopic approach in concert with CLIP. The study highlights both the possibilities and drawbacks of graphical analysis, presenting primary themes across various modalities and text-image issues that are qualitative case studies on narrative strategies in the communication of conspiracy theories.

# Methodology

## Dataset Description

The dataset comprises Amazon ratings for select products, randomly sampled to include approximately 1,600 customer reviews. It contains 27 rows and 1,597 columns. Below are descriptions of each row in the dataset:

### id: It is a unique identifier for each review.

### asin: Amazon Standard Identification Number which is a unique, 10-character alphanumeric identifier assigned by Amazon for product identification within their catalog.

### brand: It is the product brand, with possible values being Amazon or Moshi.

### categories: It is the product category, including values such as Amazon devices, routers, smart home, etc.

### colors: It is the color of the product (contains null values).

### dateAdded: It is the date when the product was added to Amazon.

### dateUpdated: It is the date when the product information was last updated.

### dimension: It is the product dimensions (contains null values).

### ean: European Article Number (EAN), which is a barcode for product identification and merchandise tracking.

### keys: It is a special assigned key for the product.

### manufacturer: It is the manufacturer of the product (contains null values).

### manufacturerNumber: It is the manufacturer-specific product number (contains null values).

### name: It is the name of the product.

### prices: It contains the product price information, presented in a dictionary format.

### review.date: It is the date on which the product was reviewed.

### review.doRecommendation: It is boolean value indicating whether the reviewer recommends the product (contains true, false, and null values).

### reviews.numHelpful: It is the number of helpful upvotes received by the review.

### reviews.rating: It is the product rating given by the reviewer.

### review.SourceURLs: It is the source URLs of the reviews.

### review.text: It is the text content of the product review.

### review.title: It is the title of the review.

### reviews.userCity: It is the city of residence of the user (only contains null values).

### reviews.userProvince: It is the province of residence of the user.

### Reviews.username: It is the username of the reviewer.

### sizes: It is the product size details.

### upc: Universal Product Code (UPC), which is a barcode used for product identification and tracking in stores.

### weight: It contains product weight information (contains null values).

## Data Preprocessing

### The following preprocessing steps were taken to prepare the data for easy text standardization and noise reduction:

### Conversion to Lowercase: The whole contents of the text were converted to lowercases to minimize case-based variability and deliver uniformity.

### Removal of Short Words: Removal of very short words with 1 or 2 characters was performed, as they are usually less significant or non-informative.

### Whitespace Removal: Unnecessary whitespaces were removed so as to do away with text irregularities

### Non-Alphabet Character Removal: Non Alphabetic characters were removed so as to retain only meaningful text.

### Stop Word Removal: Common stop words were removed to reduce noise and focus on significant terms in the text data.

## Technique Specifications

### CountVectorizer(): It converts text to a matrix of token counts with parameters:

#### max\_df=0.9: It ignores terms that appear in more than 90% of documents (removes overly common words).

#### min\_df=10: It ignores terms that appear in fewer than 10 documents (removes rare words).

#### max\_features=1000: It limits vocabulary to the 1,000 most frequent words.

### LatentDirichletAllocation():It is a topic modeling algorithm with parameters:

#### n\_components=5: It sets the number of topics to extract.

#### random\_state=42: It ensures reproducible results by fixing the random seed.

### WordCloud(): It visual represents of topics:

#### width=800, height=400: It sets the word cloud dimensions.

#### max\_words=50: It limits the number of words displayed per topic.

### TfidfVectorizer(): It converts text to a TF-IDF matrix, where:

#### max\_df=0.9, min\_df=10, max\_features=1000: It is similar to CountVectorizer parameters.

### NMF(): It extracts topics with non-negative matrix factorization:

#### n\_components=5: It specifies 5 topics to extract.

#### random\_state=42: It fixes the random seed for consistency.

### TruncatedSVD(): It is a dimensionality reduction technique:

#### n\_components=5: It reduces data to 5 dimensions, corresponding to the number of topics.

#### random\_state=42: It ensures consistent results by setting a fixed seed.

### SentenceTransformer(): It is pretrained BERT model for sentence embeddings:

#### 'all-MiniLM-L6-v2': It is model variant that balances accuracy and efficiency.

### KMeans(): It is a clustering algorithm for topic discovery:

#### n\_clusters=num\_topics: It specifies the number of clusters (topics).

#### random\_state=42: It fixes random seed for reproducibility.

# Discussion

1. This is a WordCloud representing all the topics of LDA
2. This is a WordCloud representing all the topics of NMF
3. This is a WordCloud representing all the topics of LSA
4. This is a WordCloud representing all the topics of BERT

From these word clouds, we can deduce the following themes that the reviews focus on:

### Device Performance and User Satisfaction: Across LDA and NMF topics, keywords such as "battery," "storage," and "speed" exaggerate the user focus on device performance. Users frequently invoke technical features, including comments on the quality of the screen and battery duration, implying those to be paramount attributes influencing purchase decisions.

### Comparative Product Analysis: All models highlight a strong emphasis on product comparisons,especially between Amazon devices and their competitors, like Apple. Keywords like "Apple," "Fire," and "Kindle" antecede a valuation that consumers make by comparing features, pricing,functionality, and user experience across brands.

### Entertainment and Media Usage: Entertainment is a major theme in LSA topics, with keywords like "movies," "music," and "content" highlighting media consumption as a primary use case. The BERT-based analysis, however, adds depth by revealing user narratives in which individuals recount experiences like reading a series of books or consuming various kinds of media to provide contextual insight into device use.

### Audio and Accessory Preferences: On the subject of audio accessories, their topics shed light on features like "sound quality," "comfort," and "noise cancellation," indicating what are the user priorities in their high-quality audio experience. The NMF and LSA topics stress distinctive preferences that relate to the earbuds and headphones as fit, upliftment, and durability, which clearly emphasizes what the consumers genuinely care about in their audio accessories.

### Practicality and Portability: Practical attributes include the recurrent themes of "portability" and "durability" across models. Again to a certain degree though one could be wrong, users like portable devices and accessories, with LDA and BERT topics particularly drawing upon accessories to confer a sense of portability, like protective cases and slings for your Alexa devices.

# Future Scope

Some developments that could enhance this paper's application include, Dynamic topic models for temporal analysis will allow adaptive topic tracking and such models will detect changes in discourse over time across social media platforms, periodicals, and newspapers. Such models will also give rise to event detection, in real-time tracking of the emerging trends and events and their impacts on the discourse. Also, the area within interpretability and usability will be very crucial. With explainable AI, it can explain why a particular topic is assigned; this is even more important in areas such as healthcare and law. User-guided topic models equipped will support users chipped with interactive refinements on topics based on their insights, streamlining personalized content filtering, e-learning, and knowledge discovery.

The Hybrid and Multimodal Topic Models integrated with advanced learning strategies such as those based on BERT, GPT, or T5 will bring enhanced depth in semantic analysis while defending the topic structures. Multimodal topic modeling will enable simultaneous handling of text, images, audio, or any set of data formats useful for social media analysis, medical research, and content recommendation systems. The focus on Scalability and Real-Time Applications will obviously lead to real-time topic modeling systems developed for analysis provided by prediction problems, social media tracking, and instant customer support. Scalable topic models operating in distributed systems utilizing parallel processing will be the infrastructure necessary for working with large-scale datasets common to the IoT, marketing, and public opinion policy.

Ultimately, to some extent, the evolution of cross-domain multifarious models will enhance inclusivity in NLP and give cross-linguistic topic modeling-a discourse across different languages. The expectation can be an improved global sentiment analysis that enhances market research. Moreover, it would be domain-based topic modeling that could easily illicit domain-specific knowledge assimilatively into healthcare, finance, or even law, yielding much more clarity in interpretation. This would conclude with a good dent in sentiment analysis enabling considerations of contextual insights.

# Conclusion

From this study, it shows that customer satisfaction depends on various interrelated factors, such as technical performance, value, entertainment capabilities, audio quality, and user-friendliness. Identification and analysis of these factors empower Amazon's development and marketing teams to fine-tune its products so they are more appealing to users. Based on the model of behavior, BERT finds a way to go beyond the simplistic aspect of sentiment analysis to actually go deep into complex customer narratives. These narratives aren't simple opinions. They provide context and reasoning that explain why particular features resonate with users, adding a layer of qualitative insight to complement the traditional quantitative metrics.

In conclusion, it is apparent that without the complementary input of natural language processing in analyzing reviews, the process may become tedious and, additionally, less effective. Millions of reviews abound, which makes the thought of a manual analysis practically unrealistic; thus, automation makes it possible to present a whole lot of information and conduct real-time actionable insight. This way, Amazon tours general observations into concrete, evidence-based decisions, allowing them to continuously tune its products in accordance with recent customer feedback. This, therefore, allows Amazon to significantly drive customer satisfaction, improve product enhancement, and boost marketing strategies which are well-aligned to real user needs '' and increase its competitive edge in e-commerce.

# References

1. GeeksforGeeks, "What is Topic Modeling," *GeeksforGeeks*, [Online]. [Accessed: 20-Oct-2024].
2. J. Boyd-Graber, Y. Hu, and D. Mimno, *"Applications of Topic Models,"* Department of Computer Science, University of Maryland, [Online]. [Accessed: 20-Oct-2024].
3. A. S. Patil, "Step-by-Step Guide to Master NLP Topic Modelling Using NMF," *Analytics Vidhya*, Jun. 2021. [Online]. [Accessed: 23-Oct-2024].
4. IBM, "Topic Modeling," *IBM*, [Online]. [Accessed: 30-Oct-2024].
5. L. George and P. Sumathy, "An Integrated Clustering and BERT Framework for Improved Topic Modeling," *International Journal of Information Technology*, vol. 15, pp. 2187–2195, 2023.
6. D. M. Blei, A. Y. Ng, and M. I. Jordan, *"Latent Dirichlet Allocation,"* Computer Science Division, University of California, Berkeley, CA, USA. [Online]. [Accessed: 30-Oct-2024].
7. D. D. Lee and H. S. Seung, *"Algorithms for Non-negative Matrix Factorization,"* BelJ Laboratories, Lucent Technologies, Murray Hill, NJ, USA, and Dept. of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. [Online].[Accessed: 30-Oct-2024].
8. Z. Tong and H. Zhang, *"A Text Mining Research Based on LDA Topic Modelling,"* Jodrey School o f Computer Science, Acadia University, Wolfville, NS, Canada. [Online].[Accessed: 30-Oct-2024].
9. E. Steffen, *"More than Memes: A Multimodal Topic Modeling Approach to Conspiracy Theories on Telegram,"* HTW Berlin, Germany. [Online]. [Accessed: 30-Oct-2024].
10. L. Liu, L. Tang, W. Dong, *et al.*, *"An Overview of Topic Modeling and Its Current Applications in Bioinformatics,"* *SpringerPlus*, vol. 5, p. 1608, 2016. doi: 10.1186/s40064-016-3252-8.
11. MarkovML, *"LDA Topic Modelling," MarkovML Blog*, [Online]. [Accessed: 30-Oct-2024].

[**https://colab.research.google.com/drive/1HR4yiGugjCVliH9qv6b2Ek9yl4pIElbb?usp=sharing**](https://colab.research.google.com/drive/1HR4yiGugjCVliH9qv6b2Ek9yl4pIElbb?usp=sharing)