POLITECNICO
MILANO 1863

Politecnico di Milano

Master's Thesis

Scuola di Ingegneria Industriale e dell'Informazione

Decoding Customer Feedback: Employing Topic Modelling (BERTopic) to Understand Product Return Drivers from Online Reviews

Author:
Romain Marc P Léchaudé
Person Code: 10914499
Identification Number: 219347

Supervisor:
Professor Andrea Mor

15 Credits
Academic Year: 2023-24
M.Sc. in Management Engineering - Ingegneria Gestionale

Decoding Customer Feedback: Employing Topic Modelling (BERTopic) to Understand Product Return Drivers from Online Reviews

Romain Léchaudé

June 21, 2024

Abstract

In the era of rapidly growing online sales, customer reviews have become a crucial factor in purchasing decisions. These reviews not only influence potential buyers but also provide valuable feedback to sellers about product performance and customer satisfaction. This thesis explores the application of topic modelling techniques to analyze customer reviews from the online retail sector, specifically focusing on product critiques and return reasons. The objective is to compare the efficacy of two prominent topic modelling algorithms, Latent Dirichlet Allocation (LDA) and BERTopic, in identifying underlying themes in customer feedback.

Through a detailed literature review and thorough dataset pre-processing, the algorithms are employed to uncover latent topics in customer reviews. The performance of these models is evaluated using both quantitative metrics, such as coherence, silhouette and perplexity scores, and qualitative assessments of topic interpretability and relevance. The insights derived from this study aim to potentially assist companies in developing effective methods for understanding the primary drivers of product returns, thereby enhancing product design and customer satisfaction strategies.

The findings reveal that the data cleaning steps and decisions in both LDA and BERTopic processes are crucial, significantly influencing quality metrics and interpretability of the topics. The research also demonstrates that a focused sub-category analysis leads to better clustering for interpretability and quality metrics. Product critiques are more comprehensible, and previously hidden issues emerge when the algorithm is applied stepwise to each sub-category of products rather than to the entire range, which may encompass diverse critique types. However, this approach necessitates effective sub-categorization of the dataset into distinct product types and a potential further dataset cleaning based on the most frequent terms used in the targeted sub-category group of reviews.

This research contributes to the field of topic modelling by providing a comparative analysis of LDA and BERTopic, offering practical implications for improving customer engagement and reducing return rates through better-informed decision-making processes.

Keywords: Topic Modelling; BERTopic; Latent Dirichlet Allocation; Machine Learning; Text Mining; Product Critiques; Online Customer Reviews

Abstract

Abstract in lingua italiana

Nell'era delle vendite online in rapida crescita, le recensioni dei clienti sono diventate un fattore cruciale nelle decisioni di acquisto. Queste recensioni non solo influenzano i potenziali acquirenti, ma forniscono anche un prezioso feedback ai venditori riguardo alle prestazioni del prodotto e alla soddisfazione del cliente. Questa tesi esplora l'applicazione delle tecniche di topic modelling per analizzare le recensioni dei clienti nel settore del commercio al dettaglio online, concentrandosi specificamente sulle critiche dei prodotti e sui motivi di reso. L'obiettivo è confrontare l'efficacia di due importanti algoritmi di topic modelling, Latent Dirichlet Allocation (LDA) e BERTopic, nell'identificare i temi sottostanti nel feedback dei clienti.

Attraverso una dettagliata revisione della letteratura e una accurata pre-elaborazione del dataset, gli algoritmi vengono impiegati per scoprire i temi latenti nelle recensioni dei clienti. Le prestazioni di questi modelli sono valutate utilizzando sia metriche quantitative, come coerenza, silhouette e punteggi di perplessità, sia valutazioni qualitative dell'interpretabilità e della rilevanza dei temi. Le intuizioni derivanti da questo studio mirano potenzialmente ad aiutare le aziende a sviluppare metodi efficaci per comprendere i principali fattori che guidano i resi dei prodotti, migliorando così la progettazione dei prodotti e le strategie di soddisfazione dei clienti.

I risultati rivelano che i passaggi di pulizia dei dati e le decisioni nei processi di LDA e BERTopic sono cruciali, influenzando significativamente le metriche di qualità e l'interpretabilità dei temi. La ricerca dimostra anche che un'analisi focalizzata su una sottocategoria porta a una migliore suddivisione per interpretabilità e metriche di qualità. Le critiche ai prodotti risultano più comprensibili e emergono problemi precedentemente nascosti quando l'algoritmo viene applicato progressivamente a ciascuna sottocategoria di prodotti piuttosto che all'intera gamma, che può comprendere diversi tipi di critiche. Tuttavia, questo approccio richiede un'efficace suddivisione del dataset in distinti tipi di prodotto e una possibile ulteriore pulizia del dataset basata sui termini più frequenti utilizzati nel gruppo di recensioni della sottocategoria mirata.

Questa ricerca contribuisce al campo del topic modelling fornendo un'analisi comparativa di LDA e BERTopic, offrendo implicazioni pratiche per migliorare il coinvolgimento dei clienti e ridurre i tassi di reso attraverso processi decisionali meglio informati.

Parole chiave: Modellazione Tematica; BERTopic; Allocazione Dirichleti Latente; Apprendimento Automatico; Analisi del Testo; Recensioni di Prodotti; Recensioni dei Clienti Online

Contents

List of Figures

List of Tables

1 Introduction

1.1 The field of online reviews

With the growth of worldwide online sales and delivering for all types of products, as can be observed in Figure 1 (Statista, 2024), the importance of online reviews has taken an immense step forward. Indeed, when faced with products to buy online, customers tend to look and take decision in the light of past reviews related to the products they are interested in (Zhou, 2024). Good reviews thus play a key role in online sales.

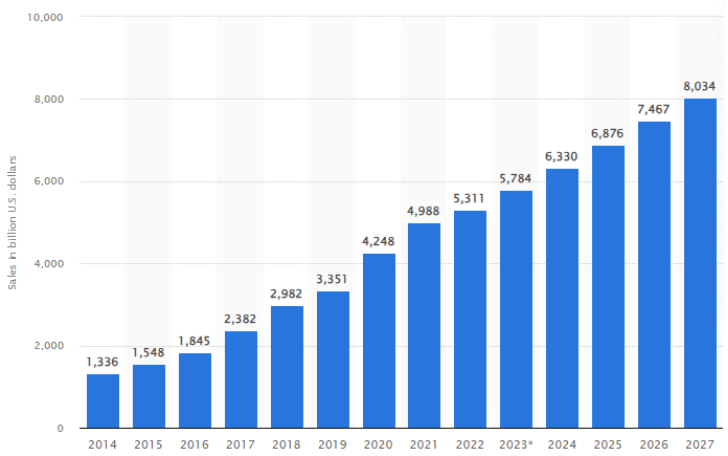

Figure 1: Worldwide online sales increase across time.

Even though, on average, only 5 to 10 percent of customers tend to write reviews (Zhou, 2024), around 88 percent of surveyed customers admit being influenced by past reviews in the buying decisions and almost 55 percent of surveyed customers read at least four reviews before making a purchase online (Zhou, 2024).

1.2 The objective of this paper

Online reviews can enable to tackle issues related to product critiques, and more precisely reasons for product returns. This paper has as objective to test several machine learning approaches to extract correctly different topics related to product critiques under the form of clusters. In this way, it would be possible to find a new way to identify issues related to products and identify needs of clients, and consequently improve continuously the product designs and marketing actions.

Several machine learning tools already exist to fit into the context of this research paper purpose. Such type of analysis is titled as Topic Modelling and can be performed through a wide range of algorithms: Latent Dirichlet Allocation (LDA), Biterm Topic Model (BTM), Self-Aggregation based Topic Model (SATM), BERTopic, etc. This research paper will focus on exploring the use of the newest method BERTopic, comparing it to LDA as a benchmark, and applying the analysis on Amazon Data reviews related to "All Beauty" products in the U.S., meaning many kinds of products related to the care of the body and of the hair: razor, brushes, perfumes, shampoo, tattoo machines, creams, etc.

This study delves into the utilization of Natural Language Processing (NLP) within the online retail sector, specifically regarding product returns. By applying NLP techniques to customer reviews, the aim is to discern the underlying reasons behind returns, thereby providing companies with valuable insights to address these issues more effectively.

1.3 Potential benefits of studies on online reviews for identifying product return drivers

The potential benefits of reducing returns extend beyond mere financial savings, encompassing reduced logistical costs and environmental advantages. As of 2022, the annual expenditure for online product returns in the USA alone amounted to 248 billion USD for an online sales market of 1.3 trillion USD in the U.S. (Chevalier, 2023). Moreover, that same year, 16.5 percent of online sales in the U.S. were returned.

Additionally, the impact of product returns extends beyond financial implications for businesses, contributing to environmental concerns as well. In 2022, emissions resulting from return shipping reached

an estimated 24 million CO_2 metric tons (Chevalier, 2023). The ultimate destination of returned items varies based on their condition and the policies of individual retailers. While a substantial portion of U.S. consumers, approximately 44 percent, believed that returned products would be resold, a significant proportion remained uncertain about their fate. Regrettably, the volume of returned items destined for landfills in the U.S. reached a staggering 4.3 billion tons in 2022 (Chevalier, 2023).

The investigation into the application of topic modelling to identify product returns drivers is still in its early stages (Mor, Orsenigo, Soto Gomez, & Vercellis, n.d.). This study utilizes free-text customer reviews sourced from online platforms as the principal dataset for knowledge extraction, facilitating an extensive analysis of return triggers across diverse product domains. Consequently, this method acknowledges the distinct attributes and complexities inherent in each product category. Furthermore, this investigation incorporates the contemporary BERTopic technique, contrasting its outcomes with those generated by the LDA algorithm to gauge the efficacy of the novel approach.

1.4 Explanation of the methodology followed

This paper adopts a structured approach to compare Latent Dirichlet Allocation (LDA) and BERTopic for analyzing online customer reviews. The methodology is divided into several key phases.

Clustering and Quality Metrics

The initial phase focuses on providing a comprehensive understanding of clustering techniques and the metrics used to assess their quality. Various metrics such as coherence scores and perplexity will be examined. The most appropriate metrics for this research will be selected based on their relevance and reliability, with justifications provided for each choice.

Latent Dirichlet Allocation (LDA)

The next phase involves an explanation of Latent Dirichlet Allocation (LDA). This includes describing its Bayesian probability model, the process of generating documents through a probabilistic model, and the use of the Document-Term Matrix for vectorizing text data. The distribution of topics within documents, represented by θ_d, and the assignment of terms to specific topics will also be detailed.

BERTopic

Following the LDA explanation, BERTopic will be introduced. The focus will be on how BERTopic leverages BERT embeddings for vector representation and its application in topic modelling. The advantages of using Sentence-BERT over the traditional Bag-of-words approach will be explained.

Data Preparation

With the theoretical groundwork laid, the practical analysis begins with an exploration and thorough cleaning of the dataset. This step is essential to ensure the data's quality and consistency, involving processes such as punctuation and stop word removal, and lemmatization.

Application of LDA and BERTopic

Subsequently, LDA and BERTopic will be applied to the cleaned dataset. The performance of each algorithm will be evaluated using the previously selected metrics. This phase includes detailed comparisons to identify strengths and weaknesses in the context of customer review analysis. Furthermore, based on the initial results, some methods will be investigated in order to optimize both the interpretation level and the values of the quality metrics.

Focused Sub-Category Analysis

Building on the initial findings, a more focused sub-category analysis will be conducted. This involves applying the algorithms to specific product sub-categories within the dataset, such as skin care and jewelry related products. The aim is to improve the interpretability and quality of the clusters by focusing on narrower, more homogeneous groups of products.

Conclusions and Recommendations

Finally, the insights gained from applying LDA and BERTopic will be synthesized into conclusions and recommendations. These will highlight the practical implications of the findings, offering actionable guidance for leveraging topic modelling to understand and address customer concerns effectively.

2 Clustering: Unsupervised Approach in Machine Learning

2.1 The different clustering methods and LDA's probabilistic approach

Machine learning can be divided in two types: Supervised and Unsupervised Learning. Supervised learning is the domain of machine learning where there is a dependent variable, a target. Unsupervised learning, on the other hand, has no target attribute. Seven data mining tasks exist (Vercellis, 2009, pp. 90-91):

- Characterization and Discrimination;
- Classification;
- Regression;
- Time series analysis;
- Associative rules;
- Clustering;
- Description and Visualization;

The four first tasks are part of the supervised learning domain while the three last ones are considered as unsupervised learning approaches.

This paper will focus on the application of the clustering method on text data, and will explore a part of the unsupervised learning field. The objective of clustering is to create homogeneous groups of points called "clusters" with the idea that observations within a cluster should be similar and observations from different clusters should be dissimilar. More precisely, in the case of topic modelling, which, as it is indicated in Figure 2 (Riva & Lerouge, 2022), is the unsupervised branch of text analysis, clusters are referred to as topics. David M. Blei has defined a topic in the following way (Blei, 2012):

"We formally define a topic to be a distribution over a fixed vocabulary."

Topic modelling refers to the use of algorithms in order to analyse text data, which is often known as unstructured data, to extract the different themes which are mentioned in the collection of text. An example of such a topic distribution in a document is observable in Figure 3 (Riva & Lerouge, 2022); in this case, the document is a review of the visit of a castle domain.

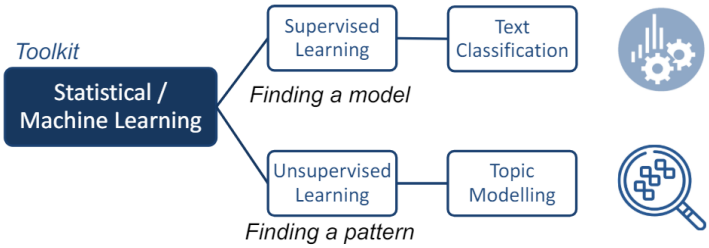

Figure 2: Text mining: supervised versus unsupervised learning.

There are different clustering methods that differ in the way the clusters are built (Vercellis, 2009, pp. 301-303):

- Partition methods (the data is divided into a pre-fixed number of clusters);
- Hierarchical methods (performs several partitions based on a tree structure);
- Density-based methods (looks at the number of observations lying within the neighborhood of each point);
- Grid methods (performs a preliminary partition based on a grid structure).

Bertopic employs a HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) algorithm that is a mixed of a Hierarchical method and Density-based one (DBSCAN being Density-based and the H standing for Hierachical). Classic Hierarchical methods are based on a tree structure (Dendogram) like the one that can be observed in Figure 4. It applies the distances among the

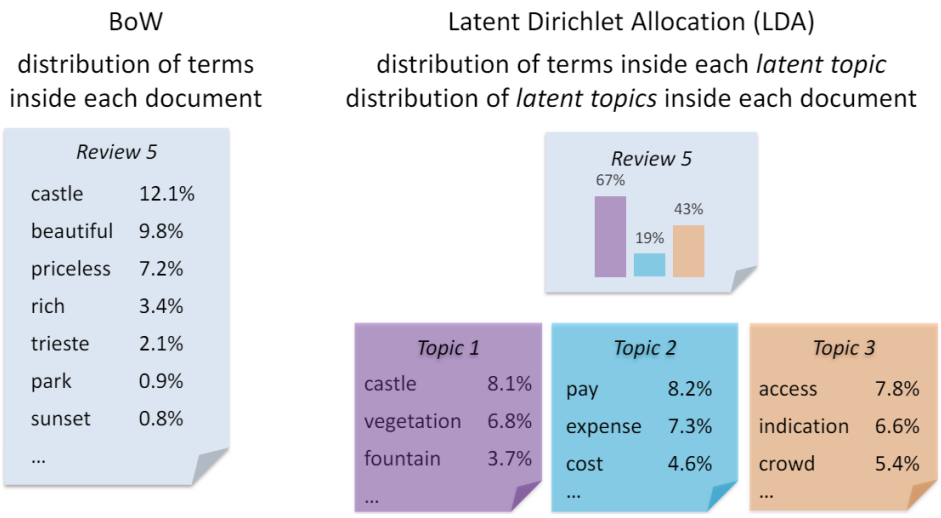

Figure 3: Three main themes are identified in the document (words are allocated to topics).

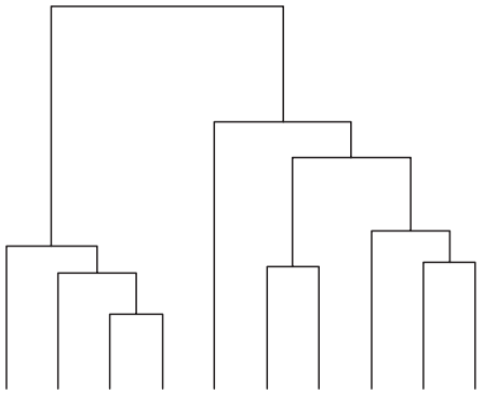

Figure 4: Example of a Dendogram, tree structure for a hierarchical clustering.

observations to derive clusters merging (Bottom-Up or Agglomerative Approach) or to split (Top-Down or Divisive Approach) them (Vercellis, 2009, pp. 314-319).

The classic DBSCAN method groups together observations with many nearby neighbours, and observations that lie alone in low-density regions are marked as outliers. An outlier denotes a data point that notably diverges from the rest of the dataset. Hawkins defines an outlier in this way:

"An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism."

In several fields, outliers are also referred to as abnormalities, discordants, deviants, or anomalies (Aggarwal & Aggarwal, 2017). A typical case of outliers identified through clustering can be observed in Figure 5. Moreover, in Figure 6, it can be seen how a box plot can be a powerful tool for detecting outliers in the data. Finally, as last example, in Figure 7, outliers can be identified through a scatter plot (Medium, 2024).

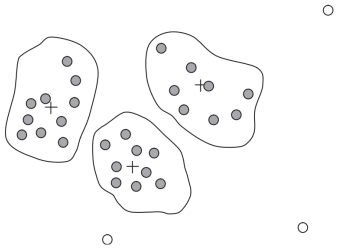

Figure 5: Example of outlier's identification through a clustering analysis.

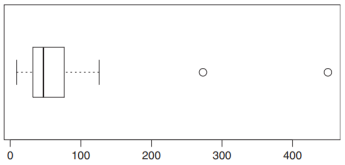

Figure 6: Example of outliers detected through a box plot.

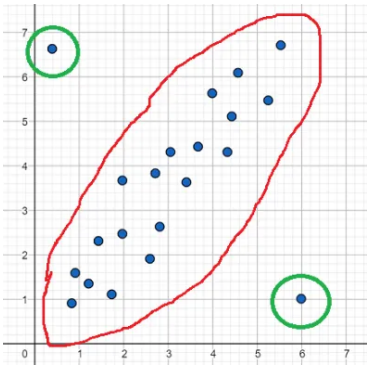

Figure 7: Example of outliers detected through a scatter plot.

The particularity of HDBSCAN, compared to classic DBSCAN, is that it enables to draw a distinction between local and global outliers. Local outliers are considered observations that might be different from other parts in their local neighbourhood though, not necessarily global outliers. An example of distinction between local and global outliers can be observed in Figure 8 (Medium, 2024).

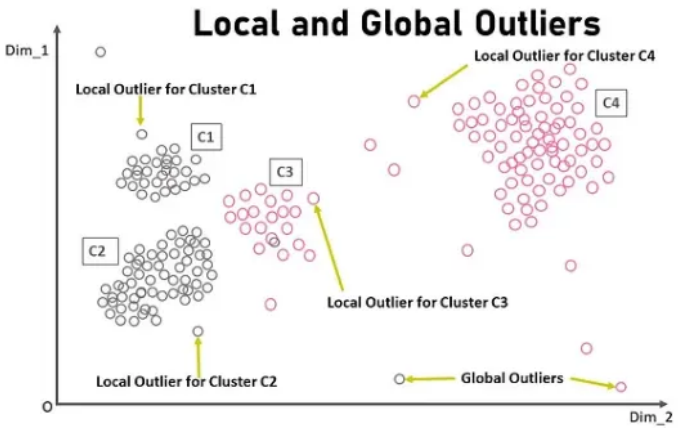

Figure 8: Local vs global outliers.

HDBSCAN uses a range of distances to separate clusters of varying densities from sparse noise. For each observation, a Neighbourhood Density (ND) measure is computed. The outlier score measures at what extend the ND of an observation is far from the NDs of its neighbour's. The outlier score distribution

can be observed in Figure 9. The higher the outlier score, the more likely the point is an outlier. Thus, this HDBSCAN method, identifying noise as outliers, ensures that unrelated documents are not forced into any cluster, thereby enhancing the clarity of topic representations (Grootendorst, 2022).

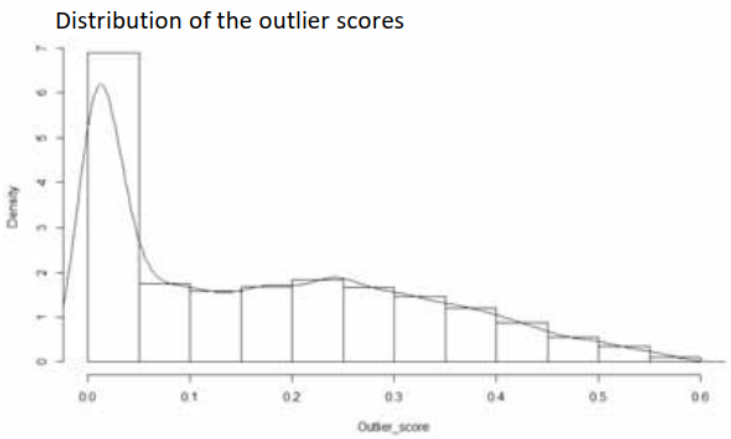

Figure 9: The outlier score distribution.

LDA, on the other hand, does not perform a real clustering as it does not measure distances but it employs a Bayseian Topic Modelling approach. Indeed, LDA does not allocate words to topics but works with probabilities: an observation (term) does not belong to a cluster but gets a score of probability of belonging to a cluster. Consequently, even though LDA can be considered as a "fuzzy clustering", its result is still a number of topics created out of the text data and can be therefore compared to BERTopic's result.

Representing text data is done using vector and matrix representation. It enables to transform qualitative and unstructured information into quantitative and structured information.

2.2 Assess Topic Model's quality

2.2.1 Cohesion, Separation and Silhouette metrics for Clusters' Assessments

How can it be identified that one clustering is better than another? How can the optimal number of clusters K be identified? Well, two approaches that will be used in a complimentary way exist:

- Relying on the interpretability of the results: is this clustering more interpretable than another one?
- Relying on statistical internal measures. Internal measures are used to measure the goodness of a clustering structure independently of external information. Five internal measures exist (Vercellis, 2009, pp. 319-321):

  1. The cohesion of each cluster: measures how closely related are objects in a cluster

  $$\text{coes}(\ C_k) = \sum_{x_i \in C_k} \sum_{x_k \in C_k} \text{dist}(x_i, x_k) \tag{1}$$

  2. The overall cohesion:

  $$\text{coes}(\phi) = \sum_{C_k \in \phi} \text{coes}(C_k) \tag{2}$$

  3. The separation of a pair of clusters: measures how distinct or well-separated a cluster is from other clusters

  $$\text{sep}(C_k, C_f) = \sum_{x_i \in C_k} \sum_{x_k \in C_f} \text{dist}(x_i, x_k) \tag{3}$$

  4. The overall separation:

  $$\text{sep}(\phi) = \sum_{C_k \in \phi} \sum_{C_f \in \phi} \text{sep}(C_k, C_f) \tag{4}$$

5. The silhouette coefficient: an observation should be very close to its siblings in the same cluster and should be less close, on average, to its cousins' observations from other clusters. It ranges in [-1;1], and the closer to 1 the better as we want v superior to u.

$$\text{silh}(x_i) = \frac{v_i - u_i}{\max(u_i, v_i)} \tag{5}$$

The lower the cohesion (i.e. the more the observations inside a cluster are near each other) and the higher the separation (i.e. the more observations from different clusters are far from each other) are, the better the clustering is.

2.2.2 Perplexity metrics for probabilistic models (LDA)

Besides these methods to identify the optimal number of clusters (K), other approaches have also been investigated in the field (Hasan, Rahman, Karim, Khan, & Islam, 2021), like the perplexity measure.

Perplexity acts as a gauge for how well a probabilistic model, like the LDA topic model, forecasts unseen or new data, reflecting the model's generalization capability. It is based on a measure of log-likelihood on a test set to evaluate the probabilistic topic model. The log-likelihood formula is (Pleplé, 2013):

$$L(w) = \log p(w|\phi, \alpha) = \sum_d \log p(w_d|\phi, \alpha)$$

The log-likelihood, denoted as $L(w)$, represents the probability of the words given the topics (ϕ) and the hyperparameters (α) for the topic distribution of documents. The summation over d indicates the aggregation across a set of unseen documents, where w_d signifies the words in each respective document. In this context, ϕ represents the topics, while α denotes the hyperparameter controlling the topic distribution of documents.

A low perplexity score indicates the model's confidence and precision in its predictions for unseen data, while a high score suggests uncertainty and inaccuracy (Blei, Ng, & Jordan, 2003). To compute perplexity, one must utilize a held-out test set, comprising documents not used during model training. The training set is utilized to train the topic model, while the test set comprises unseen documents, enabling evaluation of the model's predictive capability on novel data (Wang, Wang, Zhang, Wang, & Mao, 2019).

Perplexity is calculated using the following formula (Blei et al., 2003):

$$\text{perplexity}(D) = \exp\left(-\frac{\sum_{d=1}^{M} \log P(w_d)}{\sum_{d=1}^{M} N_d}\right)$$

where D represents the test set, M is the number of documents in the test set, d iterates over each document in the test set, $P(w_d)$ is the probability assigned by the model to the observed word sequence w_d in document d, and N_d is the number of words in document d.

While perplexity-based methods can provide meaningful insights into a model's generalization ability, they are not always stable and may lack logical rigor in determining the optimal number of topics. Some approaches use perplexity indicators to select the number of topics, but this approach may have limitations (Wang et al., 2019). Indeed, a study questions the log-likelihood approach by finding that (Chang, Gerrish, Wang, Boyd-Graber, & Blei, 2009):

"Surprisingly, topic models which perform better on held-out likelihood may infer less semantically meaningful topics."

This study points out that likelihood-based metrics, like the perplexity, for probabilistic models are often not correlated with the human interpretation capability of the results of the model (Chang et al., 2009). Therefore, this paper will go for two approaches, combining the statistical tests with human interpretability. It should be noted that as perplexities are used for probabilistic models, they can be used to assess LDA models, which are Bayesian based, but not BERTopic models, which are not probabilistic models.

The documentation does not explicitly provide the mathematical formula for the bound, only mentioning that it calculates the per-word likelihood bound using a subset of documents as the evaluation corpus (Rehurek, 2022). Thus, log-perplexity provides a likelihood bound, which is then used in the lower bound equation for perplexity. The values provided are negative ones. According to this interpretation, smaller bound values indicate deterioration, while larger values suggest a better model.

2.2.3 Coherence metrics for Topic Modelling

Another topic quality metrics is the coherence one. Topic coherence evaluation assesses individual topics by gauging the semantic similarity among the most relevant words within each topic. This evaluation aids in discerning between topics that exhibit semantic coherence and those that arise merely as statistical artifacts (Stevens, Kegelmeyer, Andrzejewski, & Buttler, 2012). This metrics is suited for topic modelling assessments. Several variations of coherence metrics exist, those adopt different approaches to estimate the words similarity inside topics.

The generalized topic coherence metrics is defined as follows:

$$\text{coherence}(V) = \sum_{(v_i, v_j) \in V} \text{score}(v_i, v_j, \epsilon)$$

where V represents a set of words describing the topic, and ϵ denotes a smoothing factor ensuring that the score function returns real numbers. The choice of ϵ can influence the coherence score; the original authors typically used $\epsilon = 1$ (Stevens et al., 2012). The value of the coherence metrics ranges between 0 ad 1, with a value near 1 indicating that the words in the topic analysed are more coherent with each others.

This measure is computed for each topic created by the clustering. However, in order to compare the LDA and BERTopic approaches, and to compare results of a same approach using different values of hyperparameters, this metrics is aggregated, through averaging, at the model level in order to evaluate the whole clustering quality level instead of the quality of individual clusters only.

A type of coherence measure is the UCI coherence one. It is based on pointwise mutual information (PMI) and cosine similarity, this measure uses word co-occurrences from a reference corpus. This type of coherence mertics has been shown to draw results that are more aligned with human judgement than other coherence measure types, like the UMass-coherence (Rosner, Hinneburg, Röder, Nettling, & Both, 2014). The UMass-coherence measure employs an asymmetrical confirmation measure between pairs of top words (using smoothed conditional probability). The summation of UMass-coherence takes into account the order among the top words of a topic. Word probabilities are estimated based on the document frequencies from the original documents used for learning the topic. It has been discovered by an experiment that both UCI and UMass coherence measures yield better performance when the parameter ϵ is chosen to be relatively small, rather than $\epsilon = 1$ as suggested in the original publications (Röder, Both, & Hinneburg, 2015).

The UCI-coherence measure is computed as (Stevens et al., 2012):

$$\text{score}(v_i, v_j, \epsilon) = \log\left(\frac{p(v_i, v_j) + \epsilon}{p(v_i)p(v_j)}\right)$$

with $p(v_i, v_j)$ being the the cosine similarity. This cosine similarity of the angles between vectors (words) can be represented as (Aletras & Stevenson, 2013):

$$\text{Simcos}(\tilde{v}_i, \tilde{v}_j) = \frac{\tilde{v}_i \cdot \tilde{v}_j}{\sqrt{\sum_k \tilde{v}_{ik}^2}\sqrt{\sum_k \tilde{v}_{jk}^2}}$$

This cosine similarity metrics is a tool enabling to determine the coherence of the topic based on a distributional similarity between top words in that topic. It computes a similarity score between pairs of words.

The PMI (pointwise mutual information for each word in the top terms of the topic) has the following form that can be observed in the UCI-coherence metrics description above (Aletras & Stevenson, 2013):

$$\text{PMI}(w_i, w_j) = \log_2 \frac{p(w_i, w_j)}{p(w_i)p(w_j)}$$

The PMI is used in order to give specific weights to vectors.

The UMass-coherence is expressed as (Stevens et al., 2012):

$$\text{score}(v_i, v_j, \epsilon) = \log\left(\frac{D(v_i, v_j) + \epsilon}{D(v_j)}\right)$$

where $D(v_i, v_j)$ represents the count of documents that include both words v_i and v_j, while $D(v_j)$ represents the count of documents that include the word v_j. Notably, the UMass metric calculates these counts based on the original corpus utilized for training the topic models, rather than relying on an external corpus. This metric aims to verify that the models capture data that is known to be in the corpus.

A variation based on using a normalized version of the PMI (NPMI) has been explored in the literature and has proved showing better correlation with human judgement than the classic PMI, whethere applied with UCI or with UMass coherence metrics (Aletras & Stevenson, 2013). It is expressed as (Aletras & Stevenson, 2013):

$$\text{NPMI}(v_i, v_j) = \frac{\text{PMI}(v_i, v_j)}{-\log(p(v_i, v_j))}$$

Additionally, a γ parameter can be applied to give greater importance to context features that have high PMI or NPMI values associated with a topic word (Aletras & Stevenson, 2013). The introduction of the γ parameter results in (Röder et al., 2015):

$$v_{ij} = \text{NPMI}(v_i, v_j)^{\gamma} = \left(\frac{\log \frac{P(v_i, v_j)+\epsilon}{P(v_i) \cdot P(v_j)}}{-\log(P(v_i, v_j) + \epsilon)} \right)^{\gamma}$$

The development of the NPMI approach has opened the door for allowing the field to identify a new better type of coherence measure: C_V (Röder et al., 2015). This measure (C_V) combines an indirect cosine measure with the NPMI and a boolean sliding window. A boolean sliding window calculates word counts by utilizing a sliding window that moves one word token at a time across the documents. Each step creates a new virtual document by copying the content within the window (Röder et al., 2015). In the context of this paper, it has been considered to use this C_V coherence metrics for its highest correlation with human interpretation of topics among all coherence metrics alternatives.

3 Functioning of LDA and BERTopic algorithms

3.1 How does Latent Dirichlet Allocation (LDA) operate?

3.1.1 The field of topic modelling

Among others, here are mentioned some classic topic modelling methods that can be used for deriving the main reasons for product returns: LDA, BTM, SATM (Mor et al., n.d.), LSA and NMF (Stevens et al., 2012).

- LDA: Latent Dirichlet Allocation
- BTM: Biterm Topic Model
- SATM: Self-Aggregation based Topic Model
- LSA: Latent Semantic Analysis
- NMF: Non-negative Matrix Factorization

The LDA method will be used as benchmark for comparison with BERTopic results.
To understand how the LDA algorithm works, let's establish some key terminology in Text Mining:

- C represents the Corpus of documents: the dataset consists of $|C|$ reviews, with each review denoted as d.
- V denotes the Vocabulary of terms: the dataset contains $|V|$ terms (or words), with each term referred to as t.
- W stands for the Document-Term Matrix: this matrix captures the weights of terms within documents.
- Tokenization refers to breaking down text structures into individual units (terms).

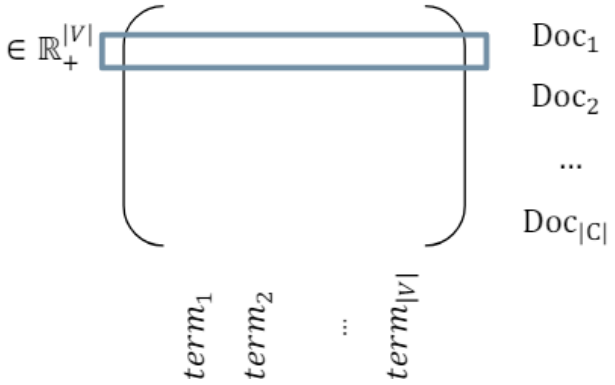

Figure 10: Document-Term Matrix (W): tool to vectorize text for LDA approach.

The Document-Term Matrix, also referred to as W, is the tool used to vectorize text data (unstructured data) into a matrix representation (structured data) in the case of a LDA approach. The Document-Term Matrix has $|C|$ rows, each one referring to a document d, and $|V|$ columns, each one referring to a term t. The structure of a document-term matrix can be observed in Figure 10 (Riva & Lerouge, 2022).

3.1.2 LDA: Bayesian Statistics

LDA, or Latent Dirichlet Allocation, is an unsupervised Bayesian statistical model (Blei et al., 2003) where documents are considered to be generated by a probabilistic model (Stevens et al., 2012). LDA assumes that documents are mixtures of topics, and topics are mixtures of words. LDA functions as follows: each document d within a given corpus D is regarded as a stochastic blend of K latent topics, where each topic represents, in the context of this paper, a critique pertaining to the product or service. In other words, it is assumed that some number K of topics exist for the whole corpus. In Figure 11 (Blei, 2012), it is possible to observe the topic distribution of a given document, and the word allocation

to the different topics present in the document. In this way, the document-term matrix is broken down in the two following matrices that can be observed in Figure 12 (Riva & Lerouge, 2022).

In LDA, the distribution of θ_i, which represents the probability of each topic appearing in each document, is assumed to be a sample from a Dirichlet distribution, thus similar to the ones that can be observed in Figure 13 (One-Off Coder, 2019). Moreover, as each document has a θ_i (a distribution of topics within the document), there is a distribution Φ_i that represents the probability of words being used for each topic and this distribution also follows a Dirichlet. These two sets of distributions correspond to the two matrices that can be observed in 12 (Riva & Lerouge, 2022). A distribution hypothesis is that terms that are from the same topics should appear in the same documents (Blei, 2012).

LDA functions, as an iterative process across all documents, as follows.

For each document D_i in the corpus:

1. Choose $\Theta_i \sim \text{Dir}(\alpha)$, a topic distribution for D_i.
2. For each word $w_j \in D_i$:
   (a) Select a topic $z_j \sim \Theta_i$.
   (b) Select the word $w_j \sim \Phi_{z_j}$.

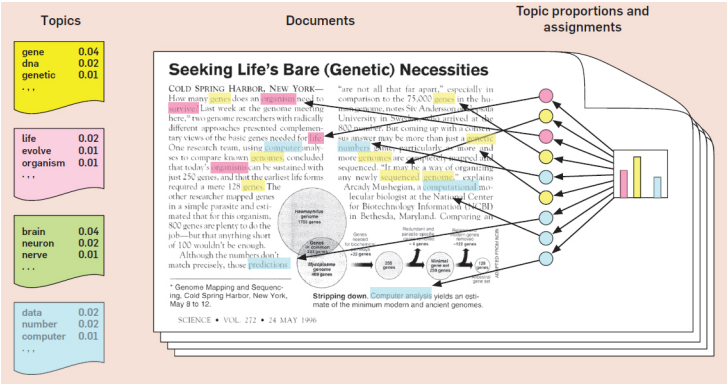

Figure 11: LDA's topic-word allocation mechanism.

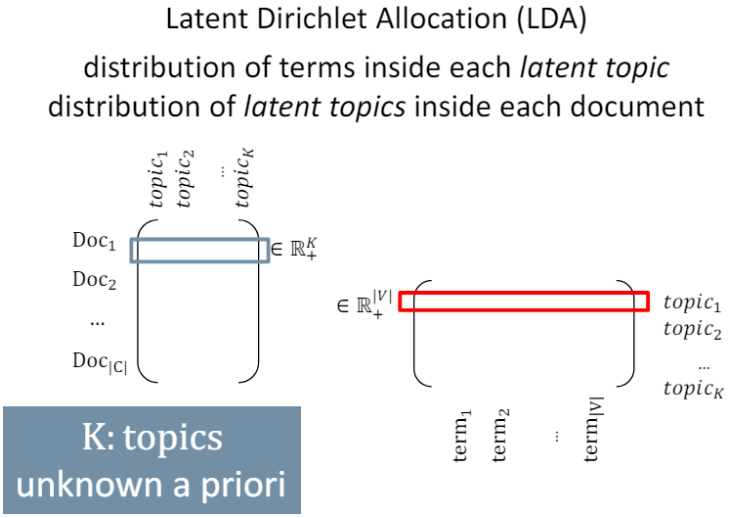

Figure 12: LDA breaks down the document-term matrix in two new ones.

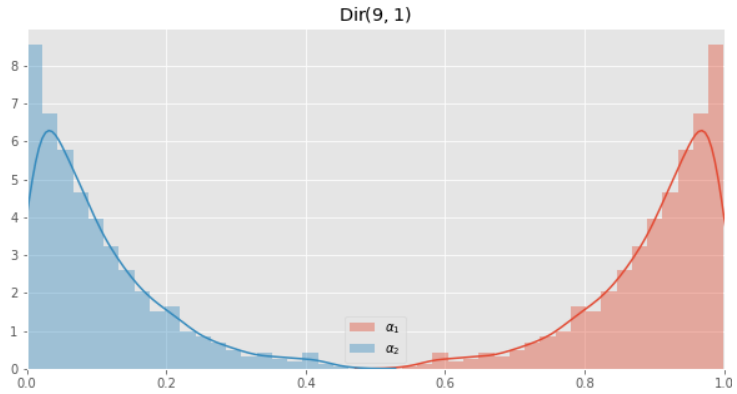

Figure 13: Example of two Dirichlet distributions.

3.1.3 Hyperparameter of LDA: number of topics K

One crucial hyperparameter in the LDA algorithm is K, representing the number of topics. It must be predetermined before initiating the model computation; it is different than parameters, which correspond to values derived from the model itself and are not set by the user. However, this requirement poses a limitation because misestimating the true number of topics can result in incomprehensible and misleading outcomes. Nonetheless, certain metrics, such as coherence and perplexity measures, along with the human oriented interpretability of the results, facilitate the evaluation of clustering quality. Consequently, these factors aid in determining the optimal K for segmentation through a trial-and-error phase.

3.1.4 Weaknesses of LDA approach

LDA faces certain limitations. Firstly, it can be hard to identify the optimal number of topics. The user can rely on some quality metrics but their full reliabilty has not been proved.

Secondly, LDA performs analysis based on individual terms without considering their order in the sentence. Indeed, each document is represented as a Bag-of-word. Consequently, this document-term matrix or Bag-of-Words approach can lead to a loss of contextual information, and more precisely it might oversee semantic relation between words used in the same document. BERTopic resolves this issue through its BERT embedding method (for vector representation), which considers the context of each term by taking into account the surrounding words. Therefore, LDA's methodology is akin to a Word2Vec approach, whereas BERTopic allows for a broader Sentence2Vec approach.

Lastly, while LDA accounts for a document being composed of a mix of different topics (in contrast to BERTopic, which typically associates one document with one topic), LDA assumes that within a document, each word is associated with a single specific topic. This may be a simplifying assumption since a word in the same document could be part of several topics.

3.2 How does BERTopic operate?

BERTopic exists as a python library developed by Maarten Grootendorst, a Dutch data scientist who also holds a background followed in psychology. BERTopic's first version (version 0.1.0) launched 24$^{\text{th}}$ of September 2020. Since the writing of this paper, BERTopic is in its 0.16.0 version.

BERTopic is an approach to topic modelling that harnesses the power of transformers and C-TF-IDF (Grootendorst, 2022). By combining these techniques, BERTopic is able to create clusters that represent topics in textual data. Unlike traditional topic modelling methods, BERTopic prioritizes the retention of important words within each topic, ensuring that the resulting topics are both interpretable and informative (Grootendorst, 2022).

One of the key features of BERTopic is its flexibility. It supports various modes of topic modelling, including guided, (semi-)supervised, hierarchical, and dynamic approaches. Moreover, BERTopic library provides several visualization tools.

BERTopic's algorithm is composed of five main phases that can be observed in Figure 14 (Grootendorst, 2022).

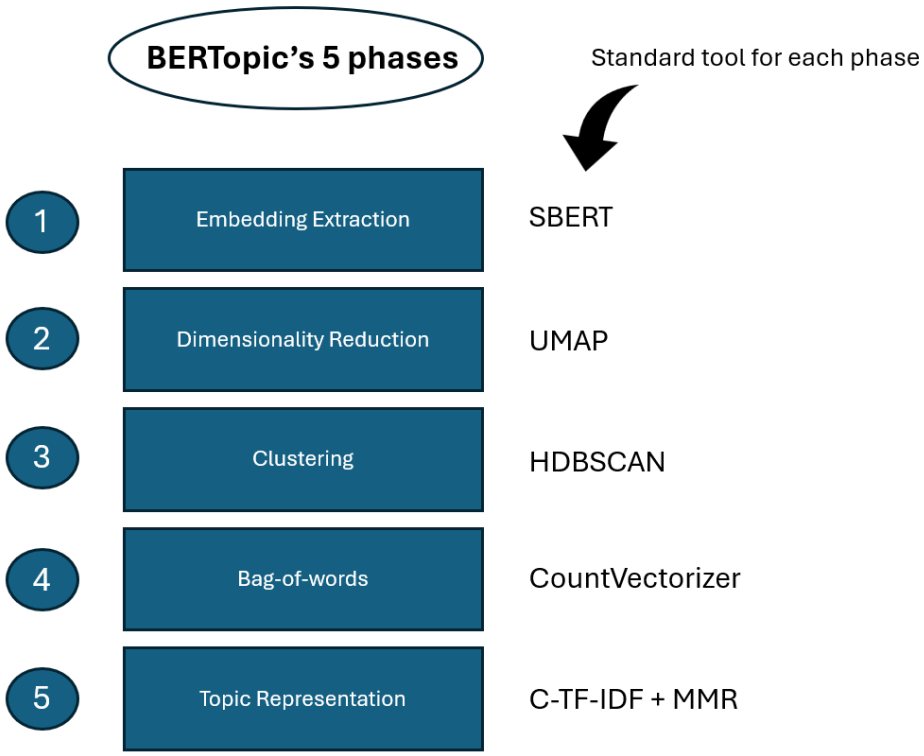

Figure 14: Sequence of steps of the BERTopic algorithm.

3.2.1 Embedding Extraction

The vector representation process of text data does not occur the same way as in LDA. Indeed, in BERTopic, no Document-Term Matrix is defined before launching the algorithm but a pre-trained embedding model, based on neural network, is used by BERTopic library itself. The reason BERTopic avoids Document-Term matrix for text vector representation is that one limitation of these models is that they use Bag-of-Words representations, which overlook the semantic relationships between words. Since these representations ignore the context in which words appear within a sentence, they may not effectively capture the true meaning of the documents (Grootendorst, 2022).

BERTopic relies on the BERT representation tool for its language representation in mathematical form. BERT stands for Bidirectional Encoder Representations from Transformers (Devlin, Chang, Lee, & Toutanova, 2018). It is a language representation model introduced by Devlin, Chang, Lee, and Toutanova in 2018 and which has given its name to BERTopic. Unlike traditional language models that read text sequentially in one direction (either left-to-right or right-to-left), BERT reads text bidirectionally. This means it can consider the context from both directions simultaneously, allowing it to develop a deeper understanding of the meaning and context of each word within a sentence (Devlin et al., 2018).

The architecture of BERT is based on Transformers, a type of neural network that utilizes self-attention mechanisms to process input sequences. This design enables BERT to effectively capture relationships between words in a sentence by looking at the entire input sequence at once (Devlin et al., 2018). During its pre-training phase, BERT engages in two tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). In MLM, BERT masks a portion of the words in the input and attempts to predict them, which helps it learn context from both directions. In NSP, BERT predicts whether a given sentence follows another, helping it understand sentence-level relationships (Devlin et al., 2018).

After the pre-training phase, BERT can be fine-tuned on task-specific datasets, which allows it to adapt efficiently to various NLP applications with relatively small amounts of task-specific data (Devlin et al., 2018).

Several BERT variations exist, all based on vector representation using semantic similarity. Nevertheless, the standard one is the SBERT (Sentence-BERT) framework. Using this SBERT methodology, many pre-trained sentence transformer models can be used directly and are still being produced on a regularly basis. Some embedding models, or also called "sentence transformers", might be better fit for different types of text data based on the language of the text data, on the subjects the text talks about, etc. A non-exhaustive list of embedding-models is: SFR-Embedding-Mistral; voyage-lite-02-instruct;

e5-mistral-7b-instruct; UAE-Large-V1; multilingual-e5-large-instruct (Grootendorst, 2022). The standard pre-trained sentence transformer for BERTopic, and the one that will be used in the report, is "all-MiniLM-L6-v2".

3.2.2 Dimensionality Reduction

In the realm of text analysis, the diversity of vocabulary stemming from the evolution of language is particularly notable in English. Originating as an old Germanic language from a region spanning Denmark, northern Germany, and parts of the Netherlands, English has been significantly influenced by French and Latin vocabulary over the centuries, especially following the Norman Conquest of England by William the Conqueror. This rich tapestry of vocabulary manifests in various forms, often leading to instances where different documents convey similar or identical meanings using distinct vocabulary choices. While this diversity enriches the corpus, it also introduces challenges that can complicate text analysis.

Firstly, diversity obscures semantic similarities between documents, complicating the identification of thematic connections. Secondly, it results in sparse and high-dimensional data representations, posing significant challenges for algorithmic processing. This phenomenon, commonly referred to as the "curse of dimensionality", highlights the inherent difficulties in navigating high-dimensional spaces (Akritidis & Bozanis, 2022). As data increases in dimensionality, the distance to the nearest observation approaches the distance to the farthest observation, rendering the concept of spatial locality ill-defined and diminishing the significance of distance measures (Grootendorst, 2022).

One significant consequence of the curse of dimensionality is the creation of lengthy vector representations dominated by zero elements. Algorithms operating on such data, such as text clustering algorithms, often struggle with computational complexity, requiring substantial memory resources and leading to increased execution times. This predicament underscores the need for effective dimensionality reduction techniques (Verleysen & François, 2005).

While Principal Component Analysis (PCA) has been a classic dimensionality reduction tool in data science, innovative techniques like UMAP (Uniform Manifold Approximation and Projection) have emerged as effective alternatives for nonlinear dimension reduction. UMAP allows for capturing both the local and global high-dimensional space in lower dimensions (Grootendorst, 2022), and is thus used as standard dimension reduction tool in BERTopic, directly applied on the embedding process's result. Furthermore, UMAP perfectly suits BERTopic as UMAP does not have computational restrictions on the number of dimensions of the embedding (McInnes, Healy, & Melville, 2018).

UMAP is a dimension reduction technique introduced by McInnes, Healy, and Melville in 2018. It is grounded in principles of manifold learning and topological data analysis, aiming to maintain the local and global structures of high-dimensional data when projecting it into a lower-dimensional space (McInnes et al., 2018). UMAP assumes that the data lies on a Riemannian manifold and uses manifold learning techniques to model local relationships within the data. By constructing a weighted k-nearest neighbor graph, it captures the local connectivity and represents the data's topological structure (McInnes et al., 2018).

One of the standout features of UMAP is its optimization process, which employs stochastic gradient descent to preserve the local neighborhood structure in the reduced dimensional space (McInnes et al., 2018). This method ensures that both the local and global features of the data are maintained, making the low-dimensional representation meaningful. UMAP is also highly scalable, handling large datasets efficiently with regard to computational speed and memory usage.

3.2.3 Clustering

The functioning of the standard clustering algorithm of BERTopic, HDBSCAN, has already been discussed above. Nonetheless, it is important to note that instead of working as observations being points, the algorithm will work with vectors (each vector representing a term).

In the context of clustering for numeric attributes, several approaches exist to compute distances between observations (Vercellis, 2009, pp. 297-298):

- Euclidean Distance: $\text{dist}(x_i, x_k) = \sqrt{\sum_{j=1}^{n}(x_{ij} - x_{kj})^2}$
- Manhattan Distance: $\text{dist}(x_i, x_k) = \sum_{j=1}^{n} |x_{ij} - x_{kj}|$
- Minkowski Distance: $\text{dist}(x_i, x_k) = \sqrt[q]{\sum_{j=1}^{n} |x_{ij} - x_{kj}|^q}$
- Mahalanobis Distance: $\text{dist}(x_i, x_k) = \sqrt{(x_i - x_k) \cdot V^{-1} \cdot (x_i - x_k)^T}$, with V^{-1} being the inversed of the covariance matrix

It can be decided which distance metrics the HDBSCAN part of BERTopic relies on by filling a parameter. Nevertheless, it was decided to keep the standard euclidean distance.

3.2.4 Bag-of-Words

Bag-of-Words (BoW), or also called Tokenizer, that has already been mentioned upon in which all the words (terms) of a document are put into a list and each given term has got its frequency appearance within the document. The Bag-of-Words can easily be displayed through a software and can result in an example like the one that can be observed in Figure 15 (Riva & Lerouge, 2022).

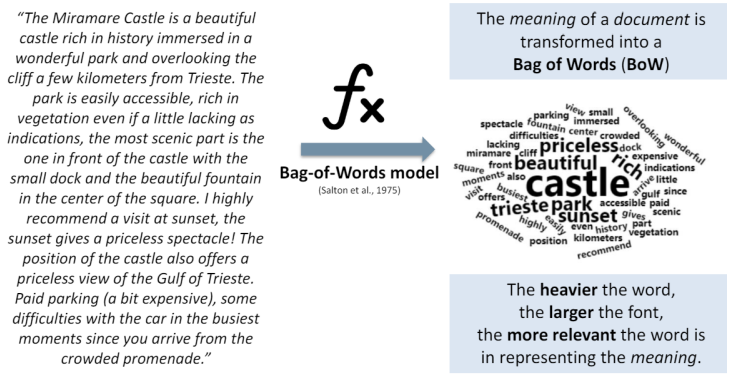

Figure 15: Bag-of-words transforms a text into a list of distinct terms with the frequency of appearance of each of those terms in the original text.

However, in BERTopic's case, this Bag-of-Words refers to the application of this listing of terms and frequencies on a topic vocabulary perspective; the Bag-of-Words model is not applied on all individual documents but on a concatenation of all the documents of one topic. Indeed, all the documents of one cluster are combined into one long document which will be achieved, for all different clusters. Then, the algorithm applies the Bag-of-Words model to count the frequency of the terms of each cluster.

BERTopic uses as standard Bag-of-Words model the "CountVectorizer" one. Changing the tool for this step does not modify the clusters themselves, as they have been identified in the previous step, but can modify how the clusters are analysed. For example, a strength of BERTopic compared to LDA, is that BERTopic's Bag-of-Words representation allows to count pairs, triples (or more) of terms to identify concepts and ideas that would remain hidden in a simple one-term frequency analysis. Indeed, counting only terms as individuals would not allow us to see that some words are often appearing together. An example of this can be seen in the analysis of the meta data that will be performed later (see the "art" case in 8.1 Use of Topic Modelling for identifying sub-categories of products). Moreover, this CountVectorizer model offers a lot of text cleaning and filtering tools that will already be dealt with during the initial cleaning phase of the data (stop words, words with low frequencies, etc.), and will therefore not be exploited any further at this step of the analysis.

3.2.5 Topic Representation

The topic representation phase, or also called weighting scheme, analyses the results of the clustering and Bag-of-Words phases in order to represent well each topic through its main components and characteristics, and through its singularity by contrast with other topics identified. The standard tool used by BERTopic for this step is called c-TF-IDF, which stands for cluster-Term Frequency-Inverse Document Frequency. The objective is to represent each topic by balancing two aspects, working from the clustered Bag-of-Words results of the previous step:

- Term Frequency-Inverse (TF): measures how frequently a term appears in a cluster; it is calculated by dividing the number of times a term appears in the cluster by the total number of terms in that cluster.
- Inverse Document Frequency (IDF): measures how unique or rare a term is across all clusters; it is calculated by dividing the total number of clusters in the corpus by the number of clusters containing the term, and then taking the logarithm of that ratio plus one (to force positivity).

These two computations, made for each term in each cluster, is balanced out in the way that can be observed in Figure 16 (Grootendorst, 2022) in order to obtain a representation that takes both into

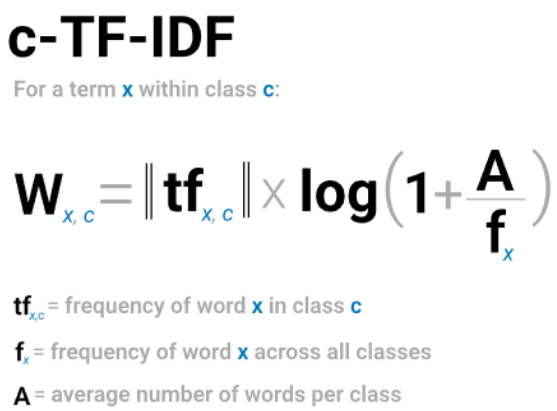

Figure 16: Balancing frequencies of terms with uniqueness of those terms compared to the other topics.

account the frequency of each term in the cluster (Bag-of-words perspective) and the presence or not of each term in the other identified clusters.

Consequently, each resulting topic representation should offer a good idea of the most important terms contained in each cluster as the Bag-of-Words' results are modified to also give less weight to terms that are also present in other topics and give more weight to the terms that are absent from the other topics.

3.2.6 Standard vs Situation-Specific Tools

As mentioned earlier, BERTopic allows high flexibility as each phase of the algorithm can be modified following the needs of the user and of the dataset's characteristics. Some examples can be observed in Figure 17 (Grootendorst, 2022).

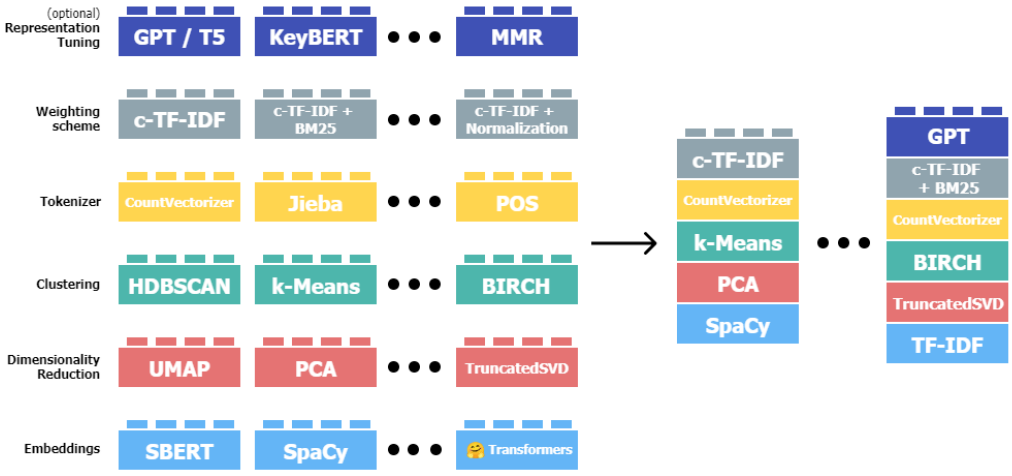

Figure 17: BERTopic offers a lot of flexibility in the tools the user can use.

3.2.7 Weaknesses of BERTopic approach

As previously explained, BERTopic is a new the state-of-the-art topic modelling framework. Its flexibility, embedding vector representation (transfomers), and its use of the c-TF-IDF for topic representation make of BERTopic a powerful tool.

Nonethless, BERtopic does contain some imperfections. One key limitation is its assumption that each document contains a single topic, which does not reflect the fact that documents can encompass multiple topics (Grootendorst, 2022). HDBSCAN, being a soft-clustering technique, helps to some extent by using its probability matrix as a proxy for the distribution of topics within a document (Grootendorst, 2022).

Another limitation is that while BERTopic provides a contextual representation of documents through language models based on the use of transformers, the topic representation is generated from Bags-of-Words. This means that the topic words highlight the importance of words in a topic but often result in redundancy and similarity among the top words, complicating the interpretation (Grootendorst, 2022).

4 The All-Beauty dataset cleaning steps

The dataset used for the analysis in this paper is composed of a subset of customer reviews from the Amazon Review Dataset (ARD dataset) (Ni, 2018) in the USA. This ARD dataset contains 233,1 million reviews for products divided across 29 categories. This paper will perform the analysis on the All-Beauty category.

The LDA model needs strong cleaning of the text data before being used, whereas the BERTopic is supposed to manage more efficiently uncleaned datasets. Nevertheless, in order to allow for a fair comparison between the two methods, it was decided to perform an identical cleaning procedure on the data in order to have a same input for both algorithms. Therefore, the filtering and cleaning step of the Bag-of-Words phase of BERTopic will be overlooked as the data will already have been through a cleaning process.

4.1 Exploratory phase of uncleaned data

The whole initial dataset covers a wide range of dates, from 2000 until late 2018, containing 371.345 reviews. It can be observed in Figure 18 the distribution of the reviews' frequency over time. It has been decided to narrow the analysis over three full years: from January 2015 until December 2017. The frequency of this period can be observed in Figure 19.

Consequently, the dataset contains 241.187 documents, all corresponding to one review made for a product of Amazon that belongs to the All-Beauty dataset and that was written between January 2015 and December 2017.

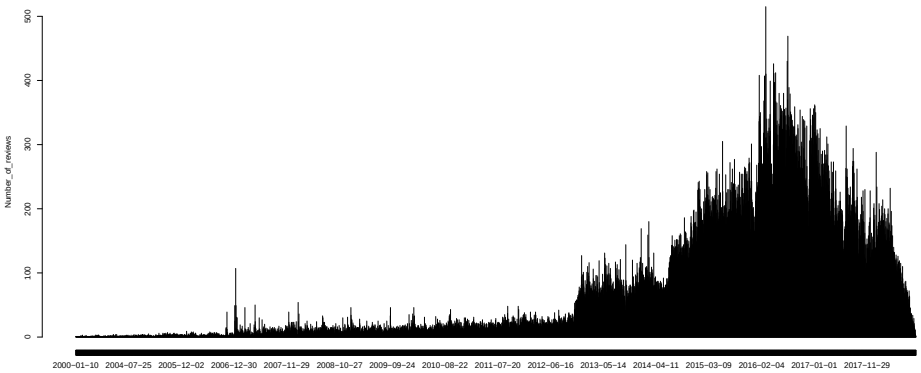

Figure 18: The dataset covers almost 20 years of All Beauty Amazon reviews, containing 371.345 reviews.

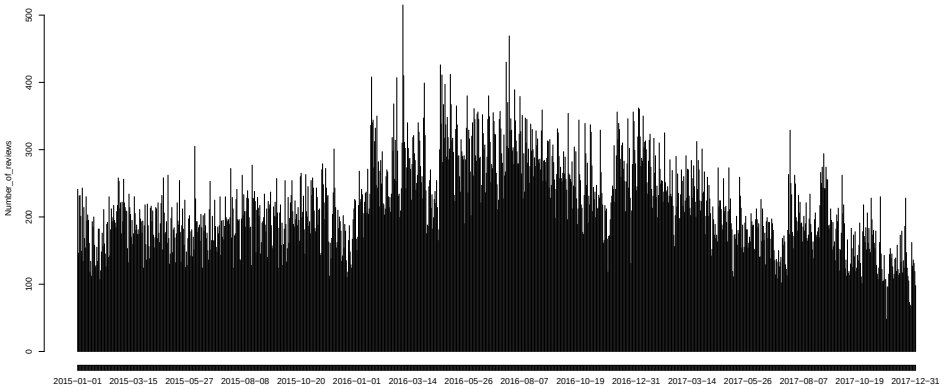

Figure 19: The analysis will focus on a three year period: 2015, 2016 and 2017, containing 241.187 reviews.

A first and fast analysis already indicates that most frequent words contained in the dataset are what are commonly considered as stop words. Stopwords can be defined as follows (Ghag & Shah, 2015):

"Stopwords are frequently occurring words that hardly carry any information and orientation."

Indeed, as can be seen in the word cloud in Figure 20, the most frequent symbols are: —, the, and, a, to, it, etc. Moreover, some words with capital letters are visible in the word cloud such as "Great" or "The". This shows that identical words might be considered as different because one version could contain a capital letter and the other one not (e.g. "the" and "The"). This can over-complexify the document-term matrix (for the LDA) by creating extra useless columns.

Figure 20: The reviews contain a lot of stop words that put forward no real meaning. Moreover, there are some words with capital letters.

4.2 Cleaning process of text data

Therefore, the following cleaning steps were applied:

1. Transforming all the texts from all documents to lower case. It prevents the same words to be considered as different by the algorithm because of the presence of one or several capital letters in one version and not in the other.

2. The punctuation was removed, which includes all kinds of special characters: — / . ? ! ; , ( ) : - etc.

3. The numbers were removed as the assumption was made that they do not influence the meaning of the critique.

4. Stop words were removed. A library on R was used that contains 175 words of the English language that are extremely common but are not adding further meaning to the analysis. The list of stop words can be found in Figure 21.

5. Spacing was adjusted to prevent having several "space" between words.

6. Stemmatization/Lemmatization/Normalization: grammatical variants of words were brought together to their stem form (ex: visit for visits, visiting, visited, visitor, etc.). This prevents us from considering as several words, variants of a same word. Indeed, the grammatical form of words should not impact the meaning of the review.

7. Uncommon words were also deleted. Indeed, an assumption was made that very infrequent words are not useful in the building process of topics as well as not representative of topics. Therefore, it was decided that words with a frequency lower than 60 occurrences in the whole dataset would be deleted.

8. Several manual changes were brought by deleting as well as regrouping words. Indeed, as infrequent words should be deleted because they do not help build clusters, very frequent words are present in too many documents to enable to correctly build the topics. In the case of BERTopic, the c-TF-IDF computation should balance out those terms from the topic representation. Nonetheless, in order to ease the topic identification algorithm and to keep the same input for LDA and BERTopic, it has been decided to already remove some of the most frequent words. Indeed, those terms might

operate such as noise that decreases the quality of the clustering process. Some examples of manual changes are: all colour-related terms were gathered under the term color, similar words to "good", like "great" for example, have been re-grouped under one term, different variations of perfume denominations have been regrouped under the term "perfume", some frequent typo errors (tthank, haveinformation, ive, highquality, etc.) have been addressed, etc.

9. Finally, reviews with stars of 5 and 4 have been deleted as the assumption was made that reviews with the top score only congratulate the product and do not make any critiques, which is what this paper is interested in. As can be observed in Figure 22, this removal means most of the reviews are deleted. Nevertheless, this filtering will help focus on critiques and product return drivers, which is the scope of this paper. The cleaned dataset contained 59.178 reviews.

The resulting word cloud can be seen in Figure 23.

```
> stpw.en
  [1] "i"          "me"         "my"         "myself"     "we"         "our"        "ours"       "ourselves"  "you"        "your"       "yours"      "yourself"
 [13] "yourselves" "he"         "him"        "his"        "himself"    "she"        "her"        "hers"       "herself"    "it"         "its"        "itself"
 [25] "they"       "them"       "their"      "theirs"     "themselves" "what"       "which"      "who"        "whom"       "this"       "that"       "these"
 [37] "those"      "am"         "is"         "are"        "was"        "were"       "be"         "been"       "being"      "have"       "has"        "had"
 [49] "having"     "do"         "does"       "did"        "doing"      "would"      "should"     "could"      "ought"      "i'm"        "you're"     "he's"
 [61] "she's"      "it's"       "we're"      "they're"    "i've"       "you've"     "we've"      "they've"    "i'd"        "you'd"      "he'd"       "she'd"
 [73] "we'd"       "they'd"     "i'll"       "you'll"     "he'll"      "she'll"     "we'll"      "they'll"    "isn't"      "aren't"     "wasn't"     "weren't"
 [85] "hasn't"     "haven't"    "hadn't"     "doesn't"    "don't"      "didn't"     "won't"      "wouldn't"   "shan't"     "shouldn't"  "can't"      "cannot"
 [97] "couldn't"   "mustn't"    "let's"      "that's"     "who's"      "what's"     "here's"     "there's"    "when's"     "where's"    "why's"      "how's"
[109] "a"          "an"         "the"        "and"        "but"        "if"         "or"         "because"    "as"         "until"      "while"      "of"
[121] "at"         "by"         "for"        "with"       "about"      "against"    "between"    "into"       "through"    "during"     "before"     "after"
[133] "above"      "below"      "to"         "from"       "up"         "down"       "in"         "out"        "on"         "off"        "over"       "under"
[145] "again"      "further"    "then"       "once"       "here"       "there"      "when"       "where"      "why"        "how"        "all"        "any"
[157] "both"       "each"       "few"        "more"       "most"       "other"      "some"       "such"       "no"         "nor"        "not"        "only"
[169] "own"        "same"       "so"         "than"       "too"        "very"       "will"
```

Figure 21: List of 175 deleted stop words.

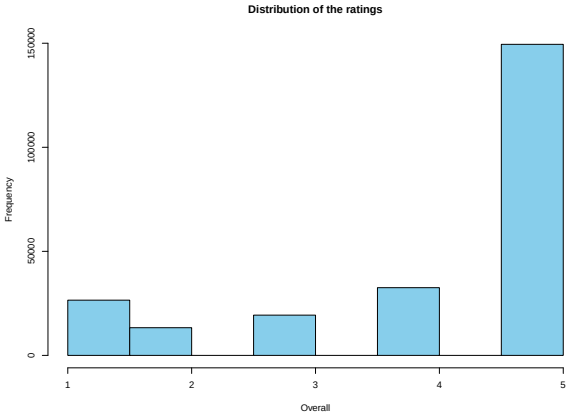

Figure 22: Most of the reviews contained in the dataset are five out of five ratings. It is considered that 4- and 5- stars reviews do not contain many critiques.

Figure 23: Word cloud of cleaned reviews.

5 Use of Topic Modelling on cleaned dataset

5.1 Use of LDA on cleaned dataset

A crucial hyperparameter for LDA algorithm is K, the number of topics. Its value must be set before launching the algorithm. To determine the optimal number of clusters, the perplexity and coherence measures can be used. The result of the perplexity analysis can be observed in the graph of Figure 24, and the result of the coherence analysis C_V can be observed in the graph of Figure 25.

The two graphs do not indicate the same optimal number of topics, 12 for the perplexity metrics and 5 for the coherence one. This difference in results can be due to the fact that perplexity and coherence do not look at the same elements for assessing the quality of the clustering. Indeed, as previously explained, the perplexity measures the model's generalization capability (its effectiveness on new data) while the coherence measure analyses the semantic similarity of the most relevant words within each topic. From an interpretable point of view, it has been observed that the number of topics of 12 shown by the perplexity metrics is better, even if still weak.

Other metrics based on distance computation cannot be used for assessing the LDA quality as this method does not rely on the computation of distances between observations to build the topics, in the contrary of BERTopic. Therefore, it has been considered that the optimal number of topics should be situated around 12 clusters.

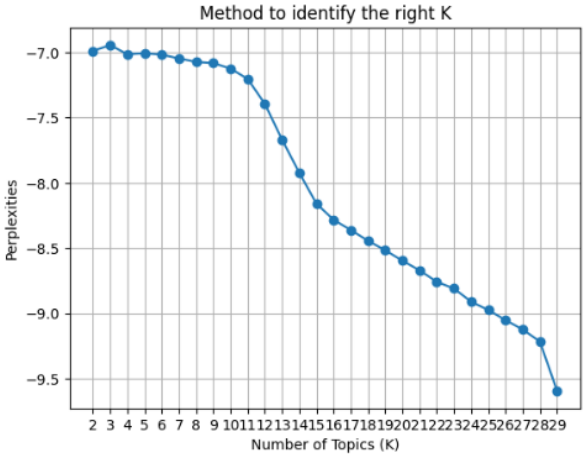

Figure 24: Optimal number of topics (optimal K) determined through perplexity measure.

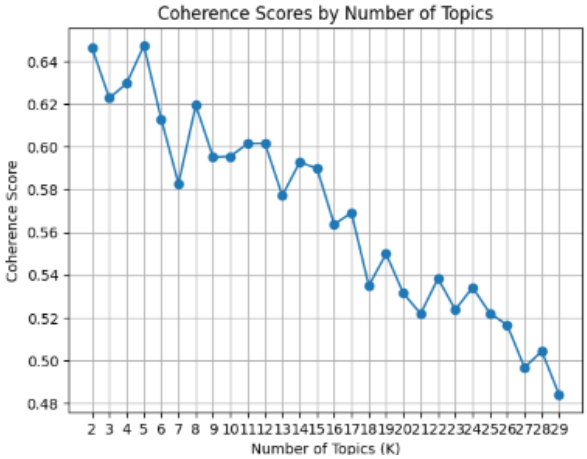

Figure 25: Optimal number of topics (optimal K) determined through coherence measure (C_V).

After this hyperparameter identification step, supported by the perplexity and coherence measures, where an optimum numbers of topics K equal to 12 was identified, it has been observed that the results are not easily interpretable. Indeed, many product names are present in the clusters and seem to overlap with the critiques. On one hand, the presence of those product names may seem to be useful as it enables to link the critiques to certain types of products; however, on the other hand, the product names seem to operate as noise that impacts the topic formation process as much as the critiques themselves. Thus,

a mix between critique clustering and product clustering seems to be at hand, and those two clustering dynamics may not always go hand-in-hand but might compete.

Some coherent topics of return drivers have been successfully extracted and can be observed in Table 1. Some remarks can be made. It can be observed that issues of size are regrouped with critiques towards too expensive prices. Moreover, a "nail" product cluster has been created that regroups several critiques already identified in other clusters. Furthermore, certain critiques like peel have been put into the "nail" cluster, but might affect other types of products as well.

Keywords
1) Different color than the one ordered
2) Size issue, expensive
3) Bottle leaks, is damaged
4) Smell, scent bad
5) Broken
6) Burn, irritate, sticky (skin related issues)
7) Nail products issues (size, difficult to remove, sticky, peel)

Table 1: Main critiques that have been identified with LDA's Topic Modelling approach.

Consequently, it might be appropriate to consider removing the product names from the dataset for the clustering process to only focus on the critiques, and less on other aspects. Before doing that, let us first observe the results of BERTopic with the same dataset without providing any new change to it.

5.2 Use of BERTopic on cleaned dataset

As previously explained, in BERTopic, there is no need to set the number of topics K. This is a great advantage of BERTopic over LDA. Nevertheless, there is another hyperparameter that must be set for BERTopic. This hyperparameter is the "minimum cluster size" which sets the minimum size of a cluster (the minimum number of documents needed to create a cluster). Consequently, this hyperparameter impacts the number of topics that will be generated. The higher the value for this hyperparameter, the lesser clusters there are (Grootendorst, 2022).

A first random approach was considered where an arbitrary low hyperparameter value for the minimum cluster size of 110 was set. This has lead to a model formation composed of 63 topics. Many of those clusters are related to each other and represent variances among same global critiques. For example, a general problem of size can be sub-divided into size issues for rings, toe, etc. A two-dimensional graph representing the different clusters (after dimensionality reduction) can be seen in Figure 26. A lot of clusters are in the same positions due to hierarchy between the topics. Using this representation, it is possible to re-gather all 63 topics into 13 main ones. From there, large critiques of types of products can be extracted. This value of 13 clusters is similar to the result obtained with LDA, in the precedent section, with its 12 clusters.

As a few real critiques could be identified out of the 63 created topics, it was assumed that the hyperparameter of the model was perhaps not optimal. Therefore, the value of the minimum cluster size was increased progressively to reach a number of topics close to the 13 identified, and for each iteration, quality metrics have been noticed and can be found in Table 2. Nevertheless, as addressed above earlier, it is noteworthy that the interpretation of those measures must be taken carefully as details and assumptions made by the formulas may lead to uncorrelated results with the human judgment. The number of clusters column omits the outlier cluster.

Minimum Cluster Size	Number of clusters	Number of Outliers	Coherence	Cohesion	Separation	Silhouette
1) 110	63	20197	0.530	24.057	11788.457	0.559
2) 170	35	18978	0.559	15.054	3463.181	0.619
3) 200	28	20163	0.562	14.220	2367.375	0.627
4) 230	26	19955	0.563	13.435	2009.168	0.622
5) 300	20	20372	0.607	12.203	1320.164	0.651
6) 350	17	22347	0.641	10.568	851.309	0.671
7) 400	15	24603	0.652	9.709	683.432	0.640
8) 450	12	20476	0.643	9.049	466.565	0.550

Table 2: Quality metrics of BERTopic's results.

The cohesion metrics, which one wants low, keeps decreasing when the value of the hyperparameter increases and the separation, which one wants high, is the highest for the lowest value of minimum cluster size. Therefore, it is observed that these two metrics offer opposite conclusions. Regarding the the coherence and the silhouette, that both indicate a better models when they show near one values,

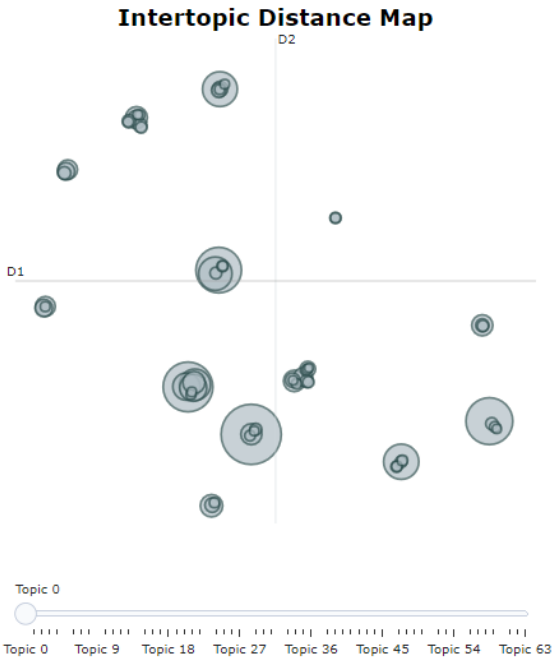

Figure 26: 2-dimensional representation of the cluster's distances.

say that the best clustering is the one with hyperparameter equal to 400 for the coherence measure, and 350 for the silhouette metrics.

Using the hyperparameter set to 350, the general critiques of products were extracted and summarized into fewer more general clusters. They can be observed in Table 3.

Keywords	Causes of returns
1) Size	Size fit issues: too large, too small, overpriced for the size, seems like child/toddler's size, no size of product described on website, not same size as the one shown in the website's picture.
2) Broken	Product or part of product is easily broken, fell apart, product or part of product arrived broken, broken too fast, bottle is leaking, razor blade feels like broken glass.
3) Battery	Battery does not turn on, battery stopped working after 1 or 2 uses, after replacing battery the product still does not work, stopped working after x months, necessary to replace batteries a lot/battery dies fast, cannot take out/replace the battery, battery leaks, battery runs out quickly, battery section does not close well.
4) Sticking	Sticks too much (Sticky feel on the product, made hair sticky, super sticky and hard to handle, very sticky and could not take off, a button sticks, hard to apply) or does not stick enough (does not stick at all, only sticks in the center, stickiness does not last in time).
5) Poor quality pieces	Cheap, messy, weak scent, broken easily, arrived with a hole (arrived broken).
6) Picture	The real product is different from the picture on the Amazon website (picture of old version, picture of other version of product, picture is deceiving, not as pictured, smaller than picture seems, larger than in picture, color does not match color in picture, looks better in picture, does not look like the picture, picture shows several versions of the product but just received one (brush)).
7) Leak	The product is leaking: bottle of perfume, arrived damage leaking, the first time it is used it leaks out on to the counter top, more than half of it leaked all over the box before it arrived, starts leaking after 6 months.
8) Smell	Bad smell (smells awful), smell (way too) strong, smell old, does not smell as it should, does not smile like the original, odd smell.
9) Delivering	Long time before delivering (late delivery, delivery was slow), not the right product was delivered, delivered broken or with missing element on the product.

Table 3: Main critiques that have been identified with BERTopic's Topic Modelling approach.

5.3 Methods to improve the current results

As the minimum cluster size increases, it has been observed that the product names take more and more role in the topics' formation and topics' representation, and it becomes more difficult to interpret the clusters. Indeed, this dataset contains the product names, and as such some topics are created around the product names and not about the critiques. For example, all jewellery products are regrouped in one cluster area but do contain several types of critiques: size issues and broken products. It is also the case for other broken products, like "mirror broken".

As the analysis has been performed on individual words and as there are many names of products as well as types of products that are present in the results and do impact the topics' formation, it makes the interpretation process quite difficult. Therefore, three additional approaches have been considered:

1. Using BERTopic abilities to take pairs (or triples) of terms instead of the classic LDA individual word by word approach. This should enable the algorithm to capture at the same time specific products and their associated critiques as well.

2. Removing all product names from the dataset so to keep the critiques only. This should generate more interpretable results but would prevent us from associating each critique to a sub-category of product. Indeed, the results would show critiques for the whole All-Beauty dataset, disregarding the different types of products that make it up.

3. Using BERTopic on different sub-categories of products of the All-Beauty dataset, enabling to perform an analysis on a range of related products, which should share common traits of critiques.

6 Use of BERTopic on groups of terms

As already explained, the Bag-of-Words step of BERTopic enables to take packages of several terms together when identifying the most frequent words for topic representation. This option enables to determine the most frequent critiques that a simple one-term Bag-of-Word approach like LDA would not indicate. Consequently, this tool from BERTopic enables us to reach Sentence2Vec methodology instead of a simple Word2Vec. Even more, in this case of presence of both critiques and names of products, this tool could potentially allow us to identify, after clustering, inside each cluster, the different product-critique associations.

It has been considered to allow for aggregation from 1 up until 3 terms to identify groups of terms that are often used together. The quality metrics of such analysis can be found in Table 4.

Minimum Cluster Size	Number of clusters	Number of Outliers	Coherence	Cohesion	Separation	Silhouette
1) 110	63	20197	0.468	24.057	11788.457	0.559
2) 170	35	19171	0.482	15.054	3463.181	0.619
3) 200	28	20447	0.495	14.220	2367.375	0.627
4) 230	26	20580	0.463	13.434	2009.168	0.622
5) 300	20	21813	0.489	12.196	1320.160	0.652
6) 350	17	23765	0.463	10.564	851.309	0.671
7) 400	15	24603	0.497	9.709	683.432	0.640
8) 450	12	20476	0.518	9.049	466.565	0.550

Table 4: Quality metrics of BERTopic's results with aggregated words.

It appears the quality metrics of the case with aggregation of terms are very similar to the one-term only scenario, except for the coherence measure which showcases lower values overall. While the coherence metrics has lower values, this aggregation approach is better from an interpretability point of view. Indeed, while the same global critiques as in point 5.2 are identified, many new interesting combinations of terms are valuable: size hand, fell apart, ring broken, chain broken, fit toe, poor job, idea poor, difficult appli, differ skin toe, leak everywher, bottl leak, start leak, smell bad, etc. Not only combinations of product and critiques are observed but also clarifications and specifications of types of critiques. Consequently, from an interpretable point of view, this aggregation of terms results in an easier interpretation of the topics and offers more detailed critiques in already identified global critiques (a hierarchical structure between critiques is highlighted).

Moreover, it was to be expected that the coherence measure would have been the only one potentially showcasing worse results than an individual term approach as this metrics measures the semantical similarity of the top words of each cluster; therefore, regrouping terms together naturally leads to less similar structures as groups of terms have more chances to be different between each other than individual terms. What could have been expected however, was an increase in the silhouette metrics; as it remains constant, it shows that the clustering may not change a lot but from an interpretability point of view, the model gains in strength.

7 Use of Topic Modelling on dataset without product names

7.1 Removal of product names from dataset

For this step, further dataset cleaning steps have been performed as frequent product names were identified and removed. The purpose is to attain clusters regrouping similar critiques of all different types of products of the dataset without drawing the distinction between the different range of product types. The list of removed terms can be observed in Figure 27 and the resulting word cloud can be found in Figure 28. This could provide us interpretable results for critiques' comprehension but might prevent us from linking the correct critiques to the right product. Indeed, a first necessary step would be to associate critiques to all possible products they may concern. This association might be easy for critiques like "skin burning" regarding cream type products, but it might be more difficult to associate critiques like "broken" or "broken first use" to the correct product types.

bag, brand_names, bracelet, brush, case, chain, coat, cream, deodor, eyelash, floss, gel, glue, lipstick, lotion, mascara, mirror, nail, necklace, perfume, phone_accessory, plastic, powder, product, pump, razor, ring, serum, shampoo, shower_gel, shower_tool, shoe, shower_products, sock, soap, spray, sticker, tattoo, teeth_products, thing, wig.

Figure 27: Product names removed from the dataset.

Figure 28: Word cloud without names of products.

7.2 Use of BERTopic on dataset without product names

The quality metrics of the BERTopic's results are shown in Table 5, the analysis consider the agglomeration of terms as well as it was observed that the interpretability increases with it but that the values of metrics remain constant, except for coherence metrics that has a general decrease.

Minimum Cluster Size	Number of clusters	Number of Outliers	Coherence	Cohesion	Separation	Silhouette
1) 110	86	25081	0.486	28.471	19063.502	0.670
2) 170	41	28769	0.477	18.185	4427.272	0.666
3) 180	38	26892	0.471	17.560	3884.945	0.636
4) 190	38	28321	0.464	20.616	3672.886	0.658
5) 200	6	4110	0.508	5.550	174.750	0.253
6) 230	5	4481	0.572	4.672	125.343	0.268
7) 300	4	5132	0.650	7.388	88.866	0.354

Table 5: Quality metrics of BERTopic's results with aggregated words with dataset without product names.

A first comparison between the results of Table 4 with Table 5, for low values of minimum cluster size, demonstrates that the number of clusters increases in the case the product names have been removed, for a same minimum cluster size. Hypothetically, it could be due to the fact that the product names were influencing the clusters formation, leading to topics around product names (with the critiques of the products in the cluster of the product name) instead of topics around critiques.

However, starting from the value of hyperparameter of 200 and above, the number of clusters with dataset without product names reduces drastically. The silhouette performance metrics also shows a decrease in performance of the resulting clusters. Additionally, even though the number of outliers decreases a lot, a strong unbalanced allocation of documents to one cluster appears.

Based on the silhouette metrics, the performance of BERTopic's algorithm on the dataset without product names could be considered as better for low values of the hyperparameter minimum cluster size. Even more, the optimal minimum cluster size for the dataset without product names based on the silhouette metrics decreases to reach the lowest value of the table, 110. The silhouette metrics at this hyperparmeter value is similar to the metrics obtained at the optimum of 350 with the dataset with the product names.

Moreover, regarding the coherence metrics, it reaches a higher value than in any situation encountered with the classic dataset with product names. Indeed, the coherence metrics reaches a value of 0.650 for a minimum cluster size of 230 while the highest value for the analysis with BERTopic for the dataset with product names and with aggregation of terms is of 0.497 for a hyperparameter of 400. Nevertheless, the interpretability linked to this metrics could be questioned as the number of topics associated with this case is very low (4 clusters).

From an interpretability point of view, with minimum cluster size of 110, the graph representing the distances between clusters after dimensionality reduction can be observed in Figure 29. It can be observed that between 15 and 17 isolated groups of bubbles exist.

Figure 29: 2-dimensional representation of the cluster's distances for BERTopic on dataset without product names.

The critique extraction process has been simplified by removing product names, leading to topics that are more centered on critiques rather than product associations. However, certain product parts or body parts remain in the dataset and influence the formation of topics. Despite this, these part-specific topics enhance our understanding of the critiques and thus improve interpretability. The extracted topics can be observed in Table 6.

Keywords
1) Bad smell
2) Size, fit, size toe, size finger, bracelet fit
3) Burn skin, burn face, burn scalp
4) Differ picture, notice differ
5) Hurt ear, hurt eye
6) Leak bottle, leak everywher
7) Batteri die
8) Arrived broken
9) Broken easili, fell apart easili
10) Need thinner
11) Piece crap, piece miss, half piece
12) Skin peel, start peel, hurt peel
13) Stone fell, stone cheap
14) Weak suction, size suction
15) Bristl fall
16) Irritate belli button
17) Zipper fell
18) Heavy

Table 6: Critiques that have been identified with BERTopic's Topic Modelling approach with dataset omitting product names.

7.3 Use of LDA on dataset without product names

To identify the optimal number of topics for the LDA algorithm, perplexities and coherence metrics have been used. The graphs representing the evolution of those metrics over the number of topics can be found in Figure 30 and in Figure 31. These two graphs indicate that the optimal number of clusters should be around 10. This could first be considered as strange results as the optimal number of clusters identified through BERTopic was of 86. Nevertheless, those 86 clusters of BERTopic may be re-grouped to more or less 16 clusters as was done in point 5.2. Furthermore, the coherence metrics for LDA for 10 clusters is of 0.635 while for BERTopic's best clustering it was of 0.670 (as can be seen in Table 4). Therefore, from a quality metrics point of view, BERTopic outperforms slightly LDA.

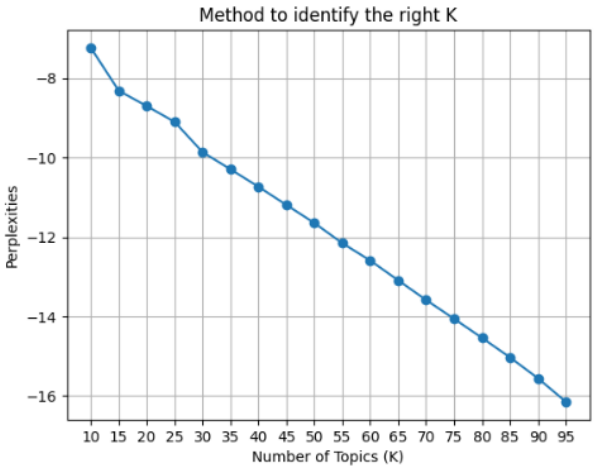

Figure 30: Perplexity measure for optimal K identification for dataset omitting product names.

Regarding the interpretability aspect, the identified products critiques have been extracted from the topics and are written in Table 7. The extraction of relevant critiques has been made easier through the removal of product names. Nevertheless, we observe that not all existing critiques have been identified. Therefore, an improvement is notable but does not seem to lead to the optimal method yet for identification of causes of product returns.

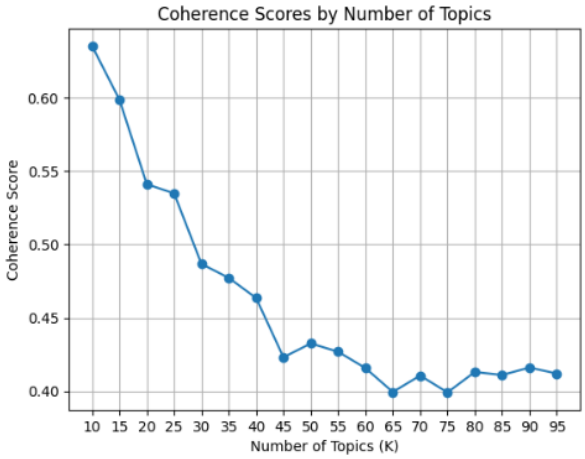

Figure 31: Coherence measure for optimal K identification for dataset omitting product names.

Keywords
1) Size
2) Expensive
3) Received (received aged, received another product than the one ordered)
4) Picture (picture, description or notice differ from real product)
5) Broken, fall (broken or falls apart in a few hours)
6) Instruction, scam, allergies (problem of instructions and allergy risk)
7) Skin, face, eye (peel, burn, irritate, sticky)
8) Uncomfortable
9) Smell (bad smell, old smell)
10) Leak (tube, battery)

Table 7: Critiques that have been identified with LDA's Topic Modelling approach with dataset omitting product names.

8 Use of Topic Modelling on sub-categories of products

8.1 Use of Topic Modelling for identifying the sub-categories of products

Even though the analysis has been performed on a dataset that contains specific products that focus on beauty products, our starting point is that products from a same category do share similar defaults, critiques. Nevertheless, the results of the analysis seem messy and difficult to interpret as it seems required to have the type of product with the default in order to draw coherent conclusions. Therefore, it has been decided to explore topic modelling on narrower categories, creating sub-categories of the All-Beauty dataset.

To carry it out, another dataset concerning meta data of the All-Beauty dataset has been used. This dataset contains a list of 32.892 products (all of the products contained in the All-Beauty dataset, with 371.345 reviews initially). Each product has its name associated and an identification number (id) under the name of "asin".

Firstly, a cleaning of the column of the names of the products has been performed on R. This cleaning step is quite similar to the one that was performed on the review (text column) of the All-Beauty dataset described upon, except for the stemming step that was overlooked as a stemming on names of products was not considered to bring any use in the clustering process. The pre-cleaned word cloud can be observed in Figure 32, and the cleaned version is in Figure 33.

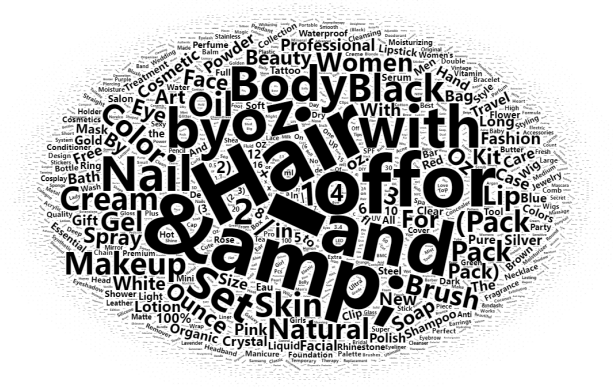

Figure 32: Uncleaned Word cloud of names of products.

Figure 33: Cleaned Word cloud of names of products.

Secondly, a topic modelling has been performed on Python with BERTopic and with LDA in order to categorise the products in different sub-categorises.

During the LDA analysis, it was observed that term "art" was very frequent and occurred present in different topics. Therefore, some investigation has been initiated and this revealed that art was a word that was almost always used with another word (pair or triple of terms): nail art; nail polish art; stickers art; body art; tattoo art; hair art; lipstick art. It was thus understood that the top frequency words that can be observed in the word cloud in Figure 34 (hair, nail, set, body; and "art" a bit later) are all

words that often appear together. As the LDA uses an analyses word per word, taking out the word from its sentence and nearby terms, such a case can easily be overlooked and may lead to an unnecessary complex document-term matrix and more difficult topics results regarding their interpretation. On the other hand, BERTopic enables to set groups of terms together (aggregation of terms for Sentence2Vec) when repeated often enough, which is a great advantage of this algorithm over the LDA.

	words	Freq
1	hair	3918
2	nail	2489
3	set	2400
4	body	2136
5	women	1727
6	perfume	1680
7	oil	1655
8	skin	1622
9	face	1603
10	cream	1575
11	brush	1488
12	makeup	1435
13	gel	1370
14	natural	1334
15	lip	1199
16	spray	1093
17	kit	1027
18	soap	998
19	eye	950
20	quot	935
21	beauty	920
22	art	902
23	case	868
24	wig	865
25	lotion	819
26	bag	768

Showing 1 to 27 of 1,004 entries, 2 total columns

Figure 34: Hair, nail, set and body all often appear with art.

Consequently, it was decided to work with the BERTopic approach for the product sub-category clustering. Some sense checks were performed and re-allocations of some products to other clusters than the ones initially allocated to by the algorithm. It is notably the case for the "Unclassified" cluster where many products could not be classified to any cluster and were thus considered as outliers. This type of clusters represented 23 percent of the products. Moreover, after cleaning of the text representing the description of the products, some rows ended up with only empty values. All those were allocated to a "???" cluster, and represents 6.93 percent of the overall dataset.

After the manual re-allocation of some products in other clusters (sub-categories of products), the different topics can be observed in Figure 35. The column CustName reflects names given manually, not by the algorithm.

	Topic	Count	Name	CustomName	Representation	Representative_Docs
0	-1	1959	-1_tattify_norelco_razor_wraps	Unclassified	[tattify, norelco, razor, wraps, philips, syst...	[body spray, body spray floz, body spray]
1	0	9257	0_skin_shower_cream_oil	Skin_Shower_Cream	[skin, shower, cream, oil, body, face, soap, m...	[vitamin korean collagen face cream sensitive ...
2	1	4029	1_makeup_powder_tattoo_eye	Makeup	[makeup, powder, tattoo, eye, brush, foundatio...	[powder cosmetic makeup brush, eyeliner eyebro...
3	2	2557	2_nail_sticker_menicure_uv	Nail_Menicure	[nail, sticker, menicure, uv, acrylic, tips, b...	[glitter shimmer uv false tips extension menic...
4	3	1726	3_nan___	???	[nan, , , , , , , ,]	[nan, nan, nan]
5	4	2325	4_ring_necklace_earring_bracelet	Jewelry	[ring, necklace, earring, bracelet, jewelry, s...	[fashion ring stainless steel leaf ring jewelr...
6	5	1799	5_case_bag_phoneaccessory_cosmetic	Phone_Accessories_and_others	[case, bag, phoneaccessory, cosmetic, travel, ...	[cat pen pencil case purse pouch cosmetic make...
7	6	1558	6_lipstick_gloss_matte_balm	Lipstick	[lipstick, gloss, matte, balm, wild, wet, stic...	[true lipstick, lipstick gloss lipstick matte ...
8	7	2002	7_hair_headband_clip_accessory	Hair_Accessories	[hair, headband, clip, accessory, clips, ponyt...	[style hair clip hair pin hair accessory, wedd...
9	8	1899	8_perfume_spray_fragrance_edt	Perfume	[perfume, spray, fragrance, edt, oil, mist, se...	[perfume perfume perfume fragrance spray, perf...
10	9	968	9_wig_cosplay_lace_curly	Cosplay	[wig, cosplay, lace, curly, straight, cap, par...	[straight hair wig heat resistant fiber synthe...
11	10	1146	10_hair_brush_styling_iron	Hair_brush	[hair, brush, styling, iron, ceramic, straight...	[hair straightener brush anti ceramic comb det...
12	11	422	11_teeth_mint_ning_kids	Teeth	[teeth, mint, ning, kids, clean, total, fresh,...	[teeth, teeth teeth ning teeth, teeth powder]
13	12	1245	12_razor_count_replacement_series	Razor	[razor, count, replacement, series, rechargeab...	[series razor count, series cc razor system co...

Figure 35: All 12 sub-categories of products.

8.2 Use of Topic Modelling on sub-categories of products

8.2.1 Use of BERTopic on sub-categories of products

The number of products in each sub-category cluster can be found in Figure 36. To analyse the results of the clustering on sub-categories of products, it has been opted to analyse a topic modelling analysis on two sub-categories: skin-shower-cream and jewelry. The approach utilizes the up to three aggregation and the dataset without product names is used.

```
The sub_category df of Skin_Shower_Cream has a length of 9995 reviews
The sub_category df of Makeup has a length of 8791 reviews
The sub_category df of Nail_Menicure has a length of 2980 reviews
The sub_category df of Jewelry has a length of 4891 reviews
The sub_category df of Perfume has a length of 3165 reviews
The sub_category df of Lipstick has a length of 2717 reviews
The sub_category df of Phone_Accessories has a length of 0 reviews
The sub_category df of Hair_Accessories has a length of 2414 reviews
The sub_category df of Cosplay has a length of 1049 reviews
The sub_category df of Breast_Bra has a length of 0 reviews
The sub_category df of Tattoo has a length of 0 reviews
The sub_category df of Hair_brush has a length of 2153 reviews
The sub_category df of Teeth has a length of 722 reviews
The sub_category df of Razor has a length of 1789 reviews
```

Figure 36: The number of reviews associated to each sub-category of products.

The results of the BERTopic analysis on the skin-shower-cream sub-category can be observed in Table 8. As the dataset contains less reviews, the minimum number of clusters hyperparameter has been brought to lower values. Starting from a minimum cluster size of 60, the clustering process no longer works as the algorithm cannot continue working because of the lacks of remaining reviews for the allocation.

Minimum Cluster Size	Number of clusters	Number of Outliers	Coherence	Cohesion	Separation	Silhouette
1) 10	198	3606	0.532	28.470	74921.885	0.699
2) 20	87	4597	0.508	17.916	16679.686	0.708
3) 30	52	4684	0.481	14.944	6510.594	0.668
4) 40	36	4201	0.450	12.737	3025.028	0.579
5) 50	26	4201	0.473	10.754	1786.883	0.553

Table 8: Quality metrics of BERTopic's results with aggregated words for skin-shower-cream products.

The critiques of the skin-shower-cream products are shown in Table 8. They were extracted from the results of the algorithm with minimum cluster size set to 20 (as the metrics indicate it is the best clustering). Nevertheless, some other critiques have been observed with minimum cluster size equal to 30 and 40, that were not in the hyperparameter equal to 20 results, and were therefore added in the table.

Keywords	Causes of returns
1) Scent	Bad smell, strong smell.
2) Arrived	Bottle arrived open, arrived damaged, long time of arrival.
3) Scalp	Burn scalp, itchi scalp.
4) Leak	Box leaks, leaks everywhere, broken leak, bottle leak.
5) Sticky	Sticky.
6) Size	bar size, size larger, size fit, hand size, size finger, handle long, handle short, size face.
7) Received	Order something but received another.
8) Notice	Notice differ from actual product.
9) Burn	skin burn, face burn, eye burn.
10) Fall apart	Broken apart, start fall apart.
11) Fake	Fake formula, fake seller.
12) Stain	Clean hard, difficult wash.
13) Peel	Skin starts peeling, face peel, peel follow.
14) Irritate	Seriously irritate skin.

Table 9: Critiques that have been identified with BERTopic's Topic Modelling approach for skin-shower-cream products.

A second sub-category, Jewelry, is analysed. The metrics resulting from the clustering processes for different minimum cluster sizes can be analysed in Table 10.

Minimum Cluster Size	Number of clusters	Number of Outliers	Coherence	Cohesion	Separation	Silhouette
1) 10	122	969	0.518	17.962	35517.478	0.754
2) 20	65	1085	0.475	14.847	10919.449	0.763
3) 30	39	1473	0.447	12.935	4256.211	0.746
3) 40	28	1657	0.434	12.272	2316.501	0.731
3) 50	19	2018	0.454	10.765	1223.386	0.736

Table 10: Quality metrics of BERTopic's results with aggregated words for jewelry products.

The critiques of the Jewelry products are shown in Table 10.

Keywords	Causes of returns
1) Broken	Broken soon, came broken.
2) Stone	Stone miss, stone fell, size stone.
3) Picture	Picture of product online does not fit the actual size of the product.
4) Nose	Size.
5) Finger	Turn finger, fit finger, turn toe.
6) Fell apart	Element fell apart.
7) Piece	Broken piece, piece crap, piece junk, piece fell.
8) Arrived	Arrived broken.

Table 11: Critiques that have been identified with BERTopic's Topic Modelling approach for jewelry products.

In the case of the jewelry sub-category analysis, the names of the body parts play an important role in the topics formation. To better improve the clustering around the critiques of the product, the body parts have been removed. The deleted terms can be seen in Figure 37 where also some other terms than body parts are present such as jewel and sterl. The quality metrics of this new clustering on the jewelry sub-category can be found in Table 12.

Ear, finger, hair, head, jewel, jewelri, nail, neck, nose, sterl, toe, wrist.

Figure 37: Body parts terms and some others that have been removed from the dataset for improving sub-category based topic modelling.

Minimum Cluster Size	Number of clusters	Number of Outliers	Coherence	Cohesion	Separation	Silhouette
1) 10	122	918	0.534	18.820	35379.992	0.760
2) 13	98	936	0.513	17.989	23396.688	0.763
3) 15	89	1007	0.512	17.353	19680.448	0.768
4) 20	73	1196	0.497	16.067	14295.870	0.763
5) 30	46	1558	0.483	15.087	6066.901	0.735
6) 40	31	1794	0.477	13.569	2901.550	0.710
7) 50	23	2301	0.469	10.930	1760.284	0.735

Table 12: Quality metrics of BERTopic's results with aggregated words for jewelry products after removal of body parts.

It can be observed that removing body parts from the dataset has led to an overall increase in coherence and silhouette metrics. Consequently, it can be concluded that topic modeling on product sub-categories may benefit from further sub-category-specific dataset cleaning.

The optimal minimum cluster size seems to be 15. In Figure 38 is the graphical representation of the distances between the topics (after dimensionality reduction). On this graph, it can be seen that the 89 clusters can be generalized to 16 topics. From the clustering based on the hyperparameter value of 15, 17 general topics were extracted and are written in Table 13.

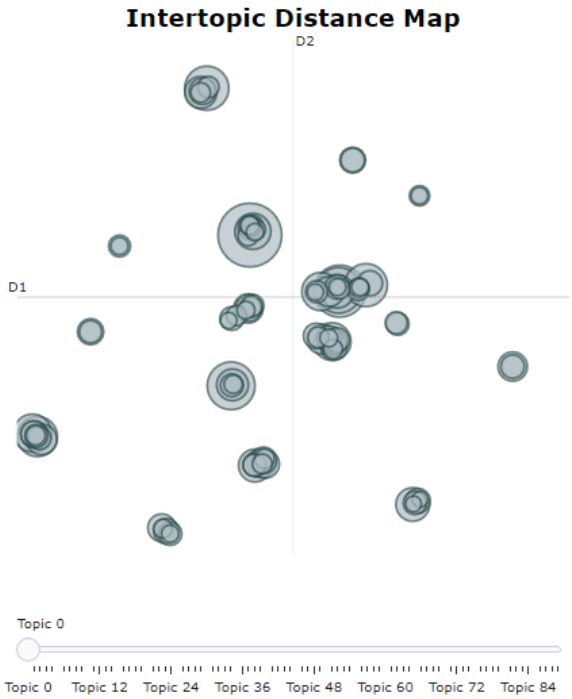

Figure 38: 2-Dimensional representation of the topics for sub-category Jewelry after removal of body parts vocabulary for minimum cluster size of 15.

Keywords	Causes of returns
1) Stone	Stone miss, stone fell, stone size.
2) Piece	Piece broken, piece fell.
3) Clasp	Clasp broken, clasp fell.
4) Size	Size kind, size wise, size hate, size round, size fit, large, size stone
5) Broken	Broken right away, bent broken, poor bent, hour broken, already received broken, Broken easily, soon broken, wore twice broken, arrived broken, arrived broken, fell apart, broken apart.
6) Ball	Ball does not screw.
7) Heavy	A bit heavy, quite heavy, heavy belli
8) Belli button	Belli button fell apart, irritate belli button.
9) Picture	Cheap picture, picture size.
10) Metal	Cheap metal, metal flake, size metal, oil smell, already turn copper.
11) Hurt	Hurt bad, hurt long, pain, bleed.
12) Bar	bar short, bar size.
13) Cuff	Cuff broken, smell rust.
14) Fade	Fade quick.
15) Fake	Fake feel.
16) Magnet	Magnet hold, broken magnet, cheap magnet, magnet not strong enough.
17) Tarnish	Broken tarnish.

Table 13: Critiques that have been identified with BERTopic's Topic Modelling approach for jewelry products after removal of body parts.

As can be observed in Table 13, even though the names of the products have been removed, the names of the components of the products have been kept in the dataset. It leads to the creation of some topics around those product components. It could be considered to remove them in order to try to further increase the quality metrics. However, keeping them at this sub-category level enables to link correctly the critiques with some specific elements of the products. For example, the fact that an element from the product fell is good, but here it is shown which types of elements can fall (stone, belli button, piece, clasp). Another example is the metal component of the product: it has issues of smell, poor quality, smelting, etc.

8.2.2 Use of LDA on sub-categories of products

To provide a comparison to the BERTopic's results on the sub-category analysis, the same analysis has been performed with LDA.

Regarding the skin-shower-cream sub-category, the optimal number of topics, K, has been identified both through the perplexity measure that is shown in Figure 39 and through the coherence measure that can be observed in Figure 40. Those two graphs reveal strange results as both graphs indicate that only two clusters are the optimal scenario. Therefore, it has been decided to create the same number of clusters as the number of generalized clusters identified through BERTopic's optimal minimum cluster size, which is of 20. As Table 8 indicate 14 identified critiques, it was decided to set the LDA hyperparameter to 14 as well. For this value, BERTopic showed a coherence measure that is between 0.699 and 0.708 (in Table 8), while LDA has a coherence metrics of 0.540 (in Figure 40). Consequently, following a pure metrics assessment method, BERTopic should outperform LDA.

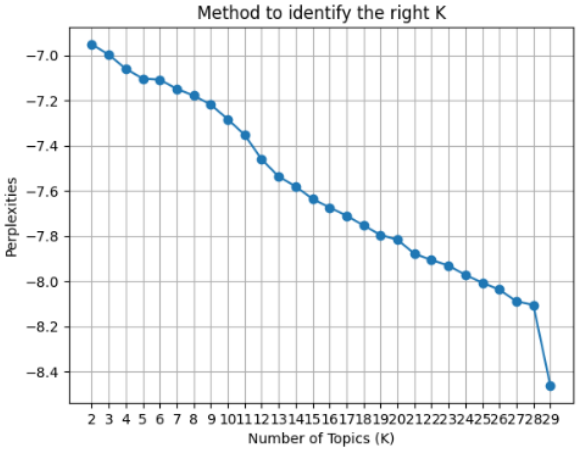

Figure 39: Perplexity measure for optimal K identification for Skin-Shower-Cream.

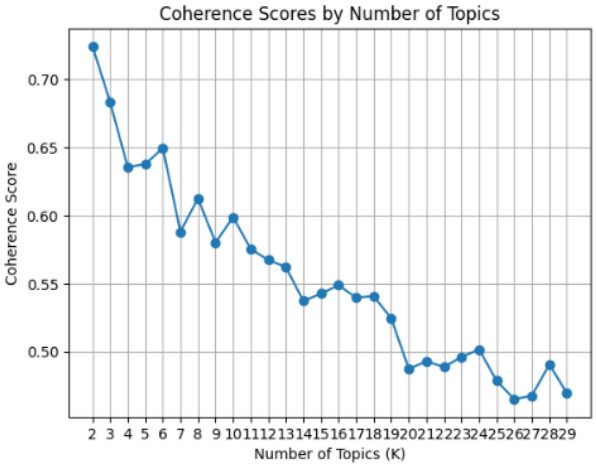

Figure 40: Coherence measure for optimal K identification for Skin-Shower-Cream.

The results of LDA's clustering interpretation results can be observed in Table 14. Those results are similar to the ones obtained with BERTopic, except for an "Overpriced, expensive" and an "Expiration day has already been reached" clusters that appear. Nevertheless, the interpretation of the results reveal itself to be quite difficult in comparison with the one for BERTopic's results. The ability of BERTopic of aggregation of terms make the results more easily interpretable, even though, as seen previously, it does not impact significantly the quality metrics.

Keywords
1) Overpriced, expensive
2) Burn (chemic, alcohol, itch)
3) Broken
4) Peel, irritate
5) Scent, smell
6) Size, fit
7) Fall
8) Arrived
9) Bottle (half, formula)
10) Scalp
11) Leak
12) Differ from notice/picture
13) Expiration date has already been reached
14) Notice differ

Table 14: Critiques that have been identified with LDA's Topic Modelling approach for skin-shower-cream products.

For the Jewelry sub-category, the dataset with body parts removal has been used. The optimal number of topics, K, has been identified both through perplexity measure in Figure 41 and through coherence measure in Figure 42. Like in the case of the skin-shower-cream sub-category analysis, the optimal K strangely seems to be of 2. Consequently, the same appraoch as with the previous sub-category has been used. Indeed, as BERTopic's jewelry results indicate that 17 topics could be identified, the number of topics set for LDA was of 17. For that K value, the coherence metrics for LDA is of 0.540 (Figure 42) while for BERtopic its coherence metrics had a value between 0.763 and 0.768 (Table 12). Consequently, from a metrics perspective, BERTopic should again outfperform LDA. The results of the clustering interpretation can be observed in Table 15.

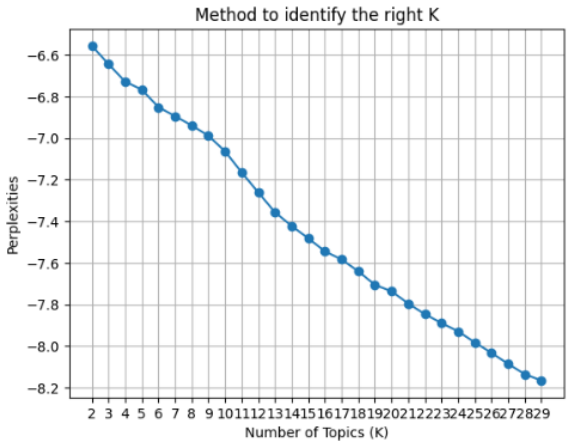

Figure 41: Perplexity measure for optimal K identification for Jewelry.

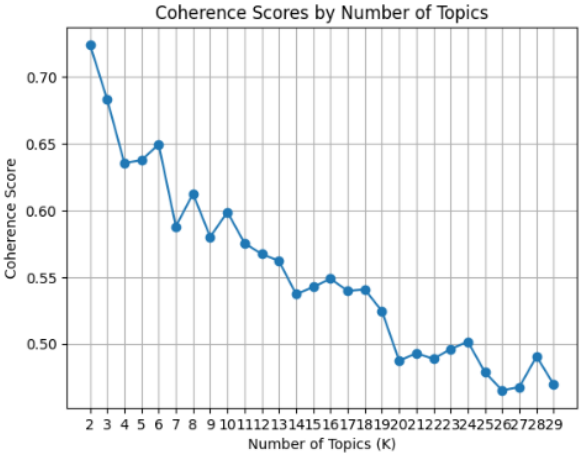

Figure 42: Coherence measure for optimal K identification for Jewelry.

Keywords
1) Fake
2) Broken
3) Hurt
4) Damaged
5) Uncomfort
6) Fall
7) Fit, size
8) Heavy
9) Cheap
10) Stone
11) Clasp
12) Arrived
13) Metal (issues related to metal part)
14) Order (ordered something different)
15) Infect (infect hand, turn bad)
16) Scratch
17) Stuck (jewelry gets stucked)

Table 15: Critiques that have been identified with LDA's Topic Modelling approach for jewelry products after removal of body parts.

The resulting clustering gives less information than the equivalent BERTopic result (Table 13). Indeed, only individual keywords for causes of returns can be extracted. It is more difficult to extract groups of terms that could offer a clear description of the problem. For example, in the case of the "Broken" cluster, other words like "within" and "hour" are present in the same topic. Through BERTopic's analysis, it has been concluded that those different elements of the cluster could be combined into "broken within hours". Another example is for the "fit" cluster, the word "stone" is present, and it has been observed is BERTopic's analysis that there is an issue of "size stone". The cluster "Arrived", refers to the a the critique of "arrived broken" that had been observed with BERTopic. A last example would be the "Cheap" cluster, just observing this word might lead us to think that it would be a positive aspect. However, BERTopic's results have shown that the "cheap" term refers to a low quality aspect. Therefore, this result for LDA might have been misleading. Nevertheless, a cluster "Uncomfort" has been created which does not exist in BERTopic's analysis, it had been incorporated in the "Hurt" cluster while in the case of LDA, a distinction is drawn. Furthermore, a cluster "Damaged" has been created in addition to the "Broken" one, this may be due to the fact that K has been set to 17 without beeing based on metrics derived from LDA performance's evaluation.

In can be concluded that, while the results of this LDA's approach give similar results to BERTopic's approach, the interpretability of the clusters is more difficult in the case of LDA. Indeed, in our case, the results of BERTopic have been used to correctly interpret the meaning between the different LDA clusters. In adddition, the number of topics used for the LDA approach are derived from the results observed with BERTopic, due to strange LDA results.

9 Summary table of critiques

Return cause	BERTopic	LDA	BERTopic aggregated	BERTopic without product names	LDA without product names	BERTOPIC skin-shower-cream	LDA skin-shower-cream	BERTopic jewelry	BERTopic jewelry modified datatset	LDA jewelry modified dataset
1 Size, fit issue (clothes, footwear, jewelry, nose)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
2 Product broken fast, easily, falls apart	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
3 Product arrived broken	✓	✓	✓	✓	✓		✓	✓	✓	✓
4 Falls apart										✓
5 Damaged product						✓	✓			✓
6 Battery charging	✓		✓	✓						
7 Poor quality pieces	✓		✓	✓						
8 Product different from picture	✓		✓	✓	✓		✓	✓	✓	
9 Inaccurate website description					✓	✓	✓			
10 Leaky product	✓	✓	✓	✓	✓	✓	✓			
11 Smell bad	✓	✓	✓	✓	✓	✓	✓			
12 Delivering issue	✓		✓							✓
13 Product arrived late						✓				
14 Received another product than one ordered					✓	✓				✓
15 Bottle arrived open						✓	✓			
16 Expiration date reached (aged)					✓		✓			
17 Color issue		✓								
18 Expensive, overpriced					✓		✓			
19 Burn, itchi (skin, eye, face, scalp)		✓		✓	✓	✓	✓			
20 Irritate (skin, eye, belly button)		✓		✓	✓	✓	✓		✓	
21 Sticking	✓	✓	✓		✓	✓				
22 Skin peels				✓	✓	✓	✓			
23 Product hurts (ear, eye, skin, bleed)				✓					✓	✓
24 Scratch										✓
25 Product quality issue								✓		✓
26 Missing piece								✓		
27 Broken piece								✓	✓	
28 Piece low quality										
29 Heavy product				✓					✓	✓
30 Instruction issue					✓					
31 Allergy issue					✓					
32 Uncomfortable					✓		✓			✓
33 Product gets stuck										✓
34 Counterfeit (fake seller, fake formula)						✓			✓	✓
35 Fade quick									✓	
36 Stain issue (difficult to wash)						✓				
37 Need thinner				✓						
38 Zipper fell				✓						
39 Weak suction, size suction				✓						
40 Bristl fall				✓						
41 Stone fell, stone cheap				✓				✓	✓	✓
42 Nail product issues		✓								
43 Clasp broken, clasp fell									✓	✓
44 Bar size									✓	
45 Ball does not screw									✓	
46 Metal (cheap, size, oil smell, turned copper)									✓	✓
47 Cuff broken, smell rust									✓	
48 Magnet (cheap, broken, not strong enough)									✓	
49 Tarnish broken									✓	

Figure 43: Summary Table of critiques.

10 Conclusion

This thesis explored the application of topic modelling techniques to analyze online customer reviews, with a particular focus on identifying drivers of product returns within the "All Beauty" category on Amazon. By employing Latent Dirichlet Allocation (LDA) and BERTopic, we aimed to extract meaningful insights from unstructured review data, thereby offering actionable recommendations for reducing return rates and enhancing customer satisfaction.

Through a comprehensive analysis of the dataset, several key findings emerged:

Data Preprocessing and Cleaning

Both LDA and BERTopic processes revealed that the initial steps of data cleaning and preprocessing are crucial. Decisions made during these stages significantly influence the quality metrics and interpretability of the resulting topics. Proper handling of text data, including tokenization, removal of stop words, and lemmatization, as well as vocabulary specific harmonization or deletion, was essential to ensure the accuracy and reliability of the topic models.

Comparative Performance of LDA and BERTopic

The comparative analysis demonstrated that BERTopic both showed a slight edge in quantitative performance metrics, such as coherence scores, and in terms of qualitative measures like interpretability and business relevance. The better interpretability is done among other things through the possibility of aggregation of words enabled by BERTopic, which does not increase metrics values but do increase interpretability of the results. BERTopic's ability to produce more coherent and human-interpretable topics suggests it is better suited for extracting actionable insights from customer reviews.

Sub-Category Focus

One of the significant findings of this study is the enhanced performance observed when the analysis was conducted on more focused sub-categories within the "All Beauty" product range. A targeted approach, where the topic modelling algorithms were applied separately to distinct sub-categories such as skin care and jewels, resulted in more precise and interpretable clusters. This method allowed for the identification of specific product defects and customer concerns that might have been overlooked in a broader analysis. Furthermore, a further sub-category specific data cleaning process proved to be a critical factor in obtaining those precise valuable and actionable insights.

Practical Implications

The findings of this research meet practical implications for online retailers and product manufacturers. By leveraging topic modelling techniques to analyze customer reviews, companies can gain a deeper understanding of the specific reasons behind product returns. This knowledge can be used to inform product design improvements, tailor marketing strategies, and ultimately enhance customer satisfaction. Additionally, reducing the rate of product returns not only saves costs but also contributes to environmental sustainability by minimizing waste and reducing carbon emissions associated with return logistics. However, the fact that only between 5 to 10 percent of customers write reviews (Zhou, 2024) has to be kept in mind as some biases might appear if the objective is to identify all or the most frequent issues encountered by clients.

In conclusion, this thesis demonstrated the efficacy of topic modelling techniques, particularly LDA and BERTopic, in extracting meaningful insights from online customer reviews. The importance of thorough data preprocessing, the benefits of sub-category focus, and the practical applications of the findings underscore the value of this approach. Future research could explore the integration of other advanced natural language processing techniques, or change the other hyperparameters of the models that have not been modified in this research, to further refine the analysis and expand its applicability across different product domains.

References

Aggarwal, C. C., & Aggarwal, C. C. (2017). *An introduction to outlier analysis.* Springer.

Akritidis, L., & Bozanis, P. (2022). Lifting the curse: Exploring dimensionality reduction on text clustering applications. In *2022 13th international conference on information, intelligence, systems & applications (iisa)* (pp. 1–8).

Aletras, N., & Stevenson, M. (2013). Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th international conference on computational semantics (iwcs 2013)–long papers* (pp. 13–22).

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

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, *22*.

Chevalier, S. (2023, December 18). *E-commerce returns in the united states - statistics & facts.* `https://www.statista.com/topics/10716/e-commerce-returns-in-the-united-states/#topicOverview`.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ghag, K. V., & Shah, K. (2015). Comparative analysis of effect of stopwords removal on sentiment classification. In *2015 international conference on computer, communication and control (ic4)* (pp. 1–6).

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Hasan, M., Rahman, A., Karim, M. R., Khan, M. S. I., & Islam, M. J. (2021). Normalized approach to find optimal number of topics in latent dirichlet allocation (lda). In *Proceedings of international conference on trends in computational and cognitive engineering: Proceedings of tcce 2020* (pp. 341–354).

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Medium. (2024, January 30). *All you need to know about outliers: Causes, types, and methods to detect them.* `https://medium.com/@pingsubhak/all-you-need-to-know-about-outliers-causes-types-and-methods-to-detect-them-0c331f9ec328`.

Mor, A., Orsenigo, C., Soto Gomez, M., & Vercellis, C. (n.d.). Shaping the causes of product returns: A topic modeling approach on online customer reviews. *Available at SSRN 4329729*.

Ni, J. (2018). *Amazon review data (2018).* `https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/`.

One-Off Coder. (2019, April 3). *Dirichlet multinomial distribution.* `https://datascience.oneoffcoder.com/dirichlet-multinomial-distribution.html`. (Last updated on Apr 03, 2024, 4:35:01PM)

Pleplé, Q. (2013, May 15). *Perplexity to evaluate topic models.* `https://qpleple.com/perplexity-to-evaluate-topic-models/`.

Rehurek, R. (2022, Dec). *Gensim: Topic modelling for humans.* `https://radimrehurek.com/gensim/models/ldamodel.html`.

Riva, P., & Lerouge, R. (2022, October 27). *Advanced performance measurement from data to kpis: Text.* PowerPoint slides. (Politecnico di Milano, Advanced Performance Measurement, Lecture 11)

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth acm international conference on web search and data mining* (pp. 399–408).

Rosner, F., Hinneburg, A., Röder, M., Nettling, M., & Both, A. (2014). Evaluating topic coherence measures. *arXiv preprint arXiv:1403.6397*.

Statista. (2024, February 6). *Retail e-commerce sales worldwide from 2014 to 2027.* `https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/`.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 952–961).

Vercellis, C. (2009). *Business intelligence: Data mining and optimization for decision making* (1st ed.). Hoboken, NJ: John Wiley & Sons.

Verleysen, M., & François, D. (2005). The curse of dimensionality in data mining and time series prediction. In *International work-conference on artificial neural networks* (pp. 758–770).

Wang, H., Wang, J., Zhang, Y., Wang, M., & Mao, C. (2019). Optimization of topic recognition model for news texts based on lda. *J. Digit. Inf. Manag.*, *17*(5), 257.

Zhou, L. (2024, April 13). *Online review statistics: The definitive list (2024 data).* https://www.luisazhou.com/blog/online-review-statistics/#:~:text=Only%205%25%20to%2010%25%20of%20customers%20actually%20write%20reviews&text=It%20turns%20out%20that%20just,go%20on%20to%20write%20them.