3D Printer Filament Review Topic Analysis

Nathan Hansen | naha1276@colorado.edu

ABSTRACT

In this project topic modeling is used to extract actionable insights from product reviews for 3D printer filament. Using this information the factors important to customers when purchasing 3D printer filament can be estimated, as well as more specific feedback on a case-by-case basis, such as per supplier, or filament type. Reviews were retrieved from the AMAZON REVIEWS 2023 [1] dataset after careful filtering was performed to identify relevant products, which required the use of supervised classification algorithms. Topic modeling was performed using the BERTopic model [2] to extract common discussion topics, from which actionable insights could be drawn. Topic comparisons were performed using a variety of metrics, including the frequencies at which topics were paired within reviews, and topic tones. These comparisons revealed several useful insights into customer preferences and common complaints, which could be expanded upon further in future analysis.

1 Introduction and Motivation

There is a tremendous wealth of information available to online retailers, from which meaningful insights can be extracted through a wide variety of methods. Such insights are highly desirable for companies attempting to improve their product offerings, marketing efforts, and their understanding of competing businesses.

Due to its popularity and wide product offerings Amazon was chosen as a suitable source of data for such analysis. The AMAZON REVIEWS 2023 dataset [1] contains both reviews and metadata for products listed on Amazon from 1996 to 2023, documenting over 48 million items and 570 million reviews. To make this dataset more approachable, a specific subset was chosen to focus on: 3D printer filament. This was deemed an interesting subset to analyze because there are many factors that can affect the perceived quality of the filament (color accuracy, dimensional uniformity, dryness, texture, etc.), and understanding which of those factors are most important to the consumer would be vital for filament sellers.

One challenge associated with this process is splitting products into distinct categories, or clusters for analysis. Most retailers provide limited granularity in their categorization structure, meaning existing metadata is often insufficient. Additionally, there is no guarantee that existing categories have been correctly assigned. Fortunately, statistical models offer an alternative approach in which categories can be inferred from other product information, such as titles and descriptions, with tunable granularity. Another challenge is the conversion of millions of reviews into a summarized format for easy interpretation. The large volumes of text must be converted into meaningful numeric representations through specialized algorithms, after which statistical models may be applied to group similar concepts. These summarized concepts would then be in a format appropriate for manual review and interpretation, significantly reducing the time-burden of human researchers.

1.1 Prior Work

1.1.1 Text Categorization Algorithms

Category extraction from text is a form of topic modeling, for which various methods have been developed. Notably, Latent Dirichlet Allocation (LDA) [4] and Non-Negative Matrix Factorization (NMF) [5], which use probabilistic approaches and the more recently developed BERTopic [2] which uses a pre-trained Large Language Model (LLM) are three common and effective methods. The Multi-LSTM/CNN approaches discussed by Krishnan et. al. [3] were also considered, but are based on the presence of both structured product attributes (size, color, material, capacity etc.) and unstructured (title and description). Since the Amazon dataset contains very little structured data for these methods to take advantage of it was determined that they would be a poor fit for this project.

Between LDA, NMF, and BERTopic it was found anecdotally that they can produce topics of similar utility, but the BERTopic implementation has been greatly streamlined and offers the opportunity to extract semantic meaning from text via the language model. This may aid in grouping products that are functionally identical, but described using different words. It was chosen as the topic-modeling and text category extraction algorithm.

1.1.2 BERTopic

The BERTopic algorithm consists of several main steps:

1. Embedding documents using a Sentence-BERT (also known as Sentence Transformer, or SBERT) [7] model, in which a pretrained LLM creates dense vector representations of the text documents
2. Dimensionality reduction using the Uniform Manifold Approximation and Projection (UMAP) [8] algorithm in which the embeddings are reduced to optimize the clustering process
3. Clustering of the reduced embeddings using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [9], which allows the number of generated topics to be adjusted by merging similar clusters

4. Topic extraction from the clusters using a class-based Term Frequency Inverse Document Frequency (cTF-IDF) [2] algorithm. During this step additional processes like stop-word removal and keyword extraction can be added to improve the human-readability of generated topic labels

One important assumption made by the BERTopic algorithm is that each document contains only one topic, which is often not true. If the model has difficulty extracting useful topics from reviews, then the reviews could be split into sentences, which are more likely to represent a single topic, but could lack some useful context.

1.2 Proposed Work

The high-level steps required for this project include: extracting target products from the dataset, extracting topics from the reviews associated with those products, interpreting the topics, and extracting insights.

Ideally, a fully unsupervised approach would be sufficient to extract the 3D printer filament products. This might be accomplished using BERTopic to create clusters of similar products using their titles. However, if the dataset proves too difficult to cleanly separate using this method, then some manual labeling and a supervised classifier would be required.

Once the dataset has been sufficiently filtered to just the 3D printing filament products the BERTopic model could again be used for topic modeling, but of the relevant product reviews instead. These topics would inform us of the general concepts that users are most frequently mentioning in their reviews, which are likely to represent the aspects of the products that are most important to them. Extracting insights requires careful inspection of the reviews associated with each topic, since the automatically generated topic descriptions are often vague.

While numerical ratings (out of five stars) are included with each review, there is no guarantee that they are representative of the tone of a given review. Using specialized sentence-transformer models would allow the sentiment of the reviews to be estimated and compared to the numeric ratings, which might provide a more reliable indication of how strongly customers feel about each topic.

2 Target Product Extraction

2.1 Initial Exploration

The AMAZON REVIEWS 2023 dataset is provided in multiple subsets, divided by main category. In this case, the “Industrial & Scientific” category was assumed to contain all products of interest (approx. 427,500 products and 5.2 million reviews) and was used for the analysis.

Initial exploration of the dataset revealed a convenient “3D Printing Filament” category, however further exploration showed that this contained more than just the relevant products. Many other types of printer components and consumables, such as printing platforms, photosensitive resins, and entire 3D printers were also included in this category. Given the frequent misclassifications of products into the “3D Printing Filament” category it seemed likely that there were also filament products missing from it, making it an unreliable method for filtering the dataset.

2.2 Dataset Reduction

Since existing categories were unreliable it was determined that an algorithmic approach to dataset segmentation would be required. Additionally, rather than operating on the full 427,500 products it was determined that certain criteria should be met for a product to be considered. Specifically, that the title contained the word “filament” and at least one known filament type (case-insensitive: PLA, ABS, PETG, PC, Nylon, PVA, HIPS, ASA, TPU, FPE, PET, PETT, PMMA, POM, PP, TPC, or TPE), which reduced the candidate products to approximately 7,000.

In preparation for the topic modeling step, in which we assume that products with large numbers of reviews will be the most interesting, the dataset was further reduced to only products with at least five reviews. This reduces the candidate product count to approx. 2800.

Some examples of resulting product titles, which highlight the difficulty in clear segmentation, or classification from titles:

- *1.75MM Filament PLA Refills, Jekon PLA Filament for 3D Pen/3D Printer 1.75mm 20 Colors One Pack, Each Color 33 feet, 660 feet in Total*
- *#1 Best Filament ABS Black 1.75 mm +/-0.02mm Top Accuracy, 3D Printer Spool Extruder Holder Stand, XYZ Printing, 1 kg Clear Print Flexible Platform, Plastic Smooth 2.2lb Refill Cartridge,Infographics*
- *Athorbot Desktop 3D Printer ABS PLA Nylon Filament Large Printing Size 11.8"x11.8"x11.8" Brother (11.8"x11.8"x11.8")*
- *TCPoly Thermally Conductive Ice9 Nylon 3D Printing Filament*

The wide mix of formats and term usage made fully filtering the data through simple word inclusion and exclusion rules unrealistic, at least without knowing the appropriate word sets ahead of time (if they exist).

2.3 Unsupervised Clustering

The reduced dataset was then passed through the BERTopic model to group the products into new categories that would make them easier to separate into “relevant” (filament) or “irrelevant” products. However, while the generated topics were frequently useful and consistent (e.g. all products were nozzles, printers, filament, etc.), they were also frequently a mix of relevant and irrelevant products. Since this was insufficient to achieve the product extraction goal it was determined that a more involved, supervised technique would be required.

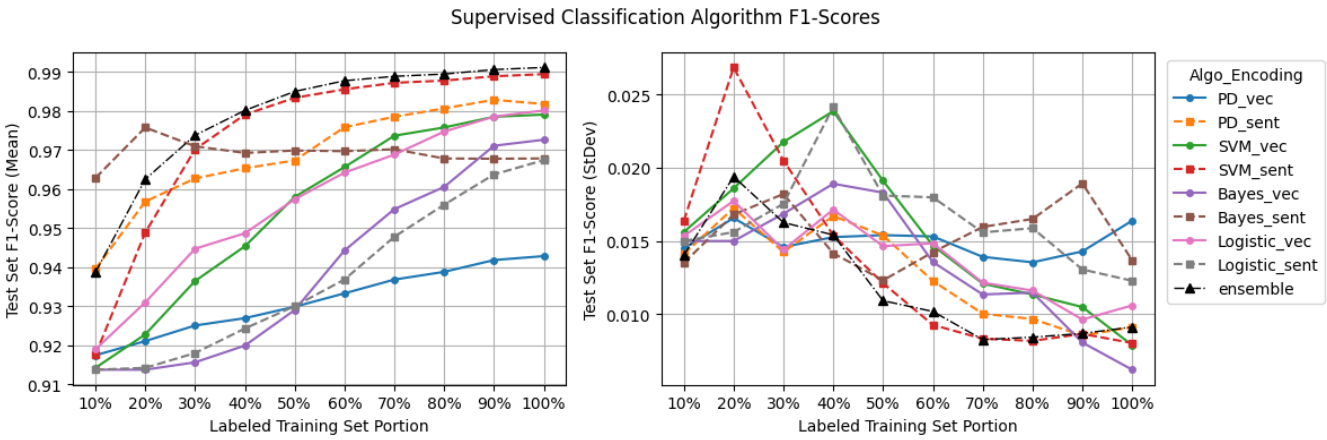

Figure 1: F1-Scores of each classifier and text-encoding pair trained on increasing proportions of the manually labeled sample.

2.4 Supervised Classification

2.4.1 Background

With enough manually labeled training data it is hypothesized that a model could be created to accurately predict all remaining labels.

Four common supervised classifiers were compared: Pairwise Distances (equivalently, K-Nearest Neighbors with k=1, using Euclidean distance), Support Vector Machine (SVM) [10][13], Bernoulli Naïve Bayes [11][13], and Logistic Regression [12][13]. These classifiers were expected to handle the binary classification task of predicting whether products are, or are not 3D printer filament well.

Two text embedding methods were also compared: using a sentence-transformer model, and using simple token vectors. The sentence-transformer model would have the advantage of capturing semantic meaning from the titles, while the token vectors would have the advantage of being simple to calculate and interpret.

The sentence-transformer model made use of the pretrained "all-MiniLM-L12-v2" model [14][15], which at time of writing was also the default embedding model of BERTopic. This is a general-purpose sentence-transformer model fine-tuned on one billion sentence pairs across a wide variety of topics. It produces 384-dimensional dense vectors capable of capturing the complex interactions between words.

The token vectors were generated by first tokenizing the titles, then creating a matrix with one row per title, and one column per possible token. A one or zero would indicate the presence or absence of each possible token in a given title, making it simple to interpret, but with a dimensionality set by the number of unique tokens (Ex: 1,354-dimensional for the sampled product titles in the following convergence study). The space and time complexities of many algorithms scale with increasing dimensions, which might be alleviated through a dimensionality reduction algorithm if needed, such as PCA or UMAP, but has been left as a topic for future work.

2.4.2 Manual Labeling

To perform manual labeling for the supervised approach, the categories generated through the unsupervised BERTopic model were reviewed. A simple GUI (see **Figure 3** in the Appendix) was created to streamline the process, and highlight where the titles in a category shared common words. This made it clear when something simple like filament color was the only variation, greatly reducing the time required to check each category. If all products in a category were uniformly relevant or irrelevant, then they were marked as such. Otherwise, the category was skipped (left unlabeled). Products from skipped categories were then re-categorized at higher granularity and the process repeated until a full sample was labeled.

One observation made during the manual labeling process was that many products were related to 3D-printer pens specifically, including filament refill sticks or spools, booklets, templates and the pens themselves. It may be worth isolating this group in future analysis since the properties that make pen filament desirable may differ from the properties for 3D printers, and this group seemed to make up a significant portion of the data.

2.4.3 Convergence Study

To characterize the performance of the supervised models and embedding methods, a convergence study was performed. The study involved manually labeling a random subset of the data, then training the supervised models on increasing portions of the labeled data. The expectation was that the performance (F1-Score) of the models would increase as more data was made available for training, identifying which models performed well with little data, which plateaued early, etc. F1-Scores were used as the comparison metric since in a heavily unbalanced dataset a high accuracy can often be achieved by always predicting one class. The F1-Score accounts for True Positives, False Positives, and False Negatives, producing a more robust indication of model performance.

20% of the dataset (536 datapoints) was manually labeled, then split into training and testing sets (80% and 20%, respectively). At 10% intervals the models were fit with a portion of the training data,

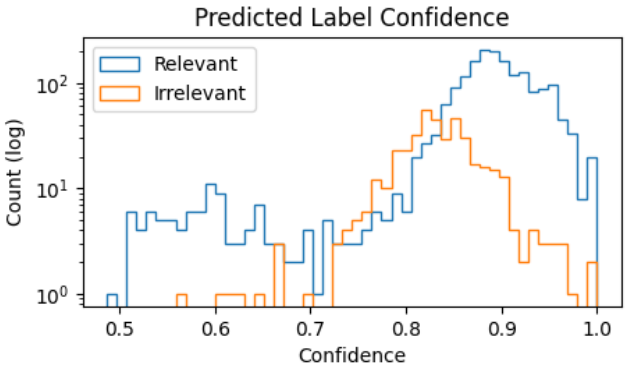

Figure 2: Classification probability ("confidence") of predicted labels by the ensemble model.

then their predictions evaluated against the full testing data. Repeating the process with re-randomized test-train splits produced significantly different results, prompting a set of ten trials to be run, then the means and standard deviations of the scores aggregated. The results are summarized in **Figure 1** (see **Table 1** and **Table 2** in the Appendix for tabulated values).

From **Figure 1** it can be observed that no one model outperformed the rest in all cases, though SVM with sentence-transformer encodings (SVM_sent) had the highest scores for most portions of training data and Binomial Naïve Bayes (Bayes_sent) performed the best at low portions. It can also be seen that the sentence-transformer embeddings tended to outperform the topic-vector embeddings in most cases, suggesting that the models benefitted from the encoded semantic information. One exception was the Logistic Regression model, which performed better with vector embeddings. This might be explained by the model's assumption of a linear relationship between independent input features and the output classification log-probability. Since the sentence-embeddings may have strong correlations and nonlinear relationships it makes sense that the model could exhibit relatively poor performance when using them.

To reduce the variability of predictions and try to improve mean scores, an ensemble of three models was evaluated: SVM, Binomial Naïve Bayes, and Pairwise Distances, all with sentence-transformer embeddings, using majority-vote rules. The ensemble's performance curve largely followed the SVM model's, but with improved means and standard deviations. The ensemble was chosen as the model to move forward with for predicting product classifications on the remaining, unlabeled data.

2.4.4 Classification Validation

The ensemble classified 2341 (84.1%) of the candidate products as "relevant". As expected, this ratio approximately matched the ratio of the manually labeled sample (84.0% relevant).

Using a model's prediction "confidence" (predicted class probabilities) can help identify where products have been misclassified. Probability calculations for the SVM and Bayes models were handled internally by their library. For Pairwise Distances the probability was estimated using the inverse normalized distance to the nearest labeled point:

$$P(x_i) = 1 - \frac{\min(d_{i,j}) - d_{min}}{d_{max} - d_{min}}$$

Where $\mathbf{x_i}$ is a given unlabeled title encoding from the set to be predicted, $\mathbf{d_{i,j}}$ is the distance between each labeled title encoding and $\mathbf{x_i}$, and $\mathbf{d_{min}}$ and $\mathbf{d_{max}}$ are the minimum and maximum distances of all pairs, respectively. The assumption is simply that the closer two points lie in embedded space the more confident we can be that they are describing a similar product, and should be labeled the same.

After averaging the probabilities across the three models in the ensemble (see **Figure 2**), 126 titles were identified with "low" confidence (< 0.75 for both classes). These titles were manually re-labeled revealing 16 misclassified products. The low confidence titles frequently included "bundles" where multiple products were present (e.g. 3D printer *with* filament), or a specific use case was called out (e.g. filament *for* MakerBot Replicator 3D Printer, filament *for* 3Doodler Pen, etc.). This reflects the expectation that the presence of multiple potential topics would make classification more difficult.

The ensemble was re-trained with the updated labels, resulting in 2291 products labeled as relevant (82.4%). Further iterations of correcting low-confidence predictions and re-training the model could be performed if desired, but they become increasingly inefficient as average confidence increases and expected error rate decreases.

3 Review Topic Modeling

Once the desired products had been isolated, the relevant reviews could be retrieved by matching the unique "Parent_asin" keys between the review and products data. Approximately 117k reviews were selected, from which topics could be extracted using BERTopic. Initial topic extraction trials produced topics that tended to be very generalized and difficult to identify a consistent theme from, likely due to the various topics that may be covered within a single review. Instead, topic modeling was performed on the individual sentences from the reviews (approximately 463k sentences). The tradeoff of increased execution time for increased subjective topic clarity was deemed worthwhile.

There doesn't exist a single "best" number of topics for a given corpus, since it depends on the needs of the user. For BERTopic, the number of topics generated depends on the number of clusters created by HDBSCAN, which is mainly influenced by the minimum cluster and sample size parameters. These parameters were varied manually and the resulting topics reviewed until a subjectively "good" set of topics was generated. A total of 25 topics were generated and one "uncategorized" topic ("-1_3d_print_silk_filament") for sentences that didn't meet HDBSCAN's clustering criteria (summarized in **Table 3** in the Appendix). Topics will be referred to either using their full name,

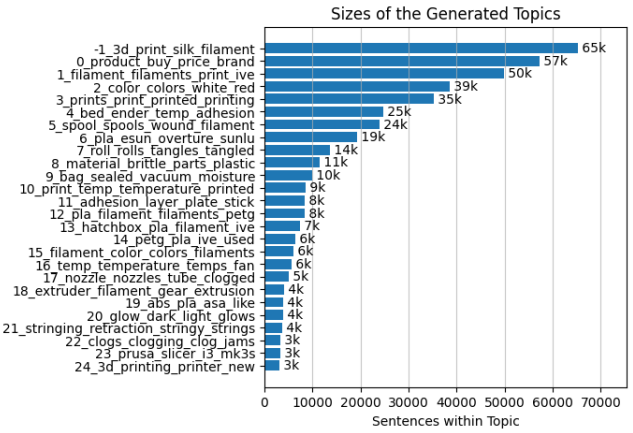

Figure 3: The generated topics and non-uniform number of sentences assigned to each topic.

or for brevity their index number, or a keyword-number pair (e.g. 2_color_colors_white_red, #2, or Color #2).

The number of sentences assigned to each topic are shown in **Figure 3**, which highlights the non-unform distribution, differing by over an order of magnitude from largest to smallest. Similar results were consistent over various HDBSCAN parameters, suggesting that a near-uniform class balance would be unrealistic with this method (combination of clustering and encoding methods).

The automatically generated topic names describe the words that occur disproportionately in each topic as calculated by cTF-IDF. These names help identify the core concepts of each topic, but lack information on how those concepts are applied, and their tone, for which manual review is required. For example, we could correctly assume that topic "12_pla_filament_filaments_petg" contains comments related to PLA filament, but what customers are saying *about* the filament must be extracted manually from example reviews.

Another approach, which has not been implemented for this project but could be a topic for future analysis, is to collect the reviews associated with a topic of interest and perform topic modeling on just that subset to extract its various sub-topics (a hierarchical approach).

3.1 Topic Summaries

Descriptions for the generated topics were created manually by reviewing 20 samples from each topic with the highest class-probabilities, as calculated by BERTopic (the "most representative" samples). The concepts present in the topics can be summarized as follows:

- Filament Appearance
  - Reviewers frequently mentioned disappointing color accuracy compared to images online, especially for metallic colors (gold, silver, etc.)
  - Filament designed to have a matte texture was repeatedly noted as having worse properties than normal filament
  - Customers found that surface finish of parts could be inconsistent while using the same filament and settings, suggesting manufacturing inconsistencies
  - For glow-in-the-dark filament, the brightness, "charge" time and glow duration were frequently commented on
- Filament Spools
  - Spools that were poorly wound, too tightly wound, or with short walls that could let filament slip off were mentioned as sources of tangling, or otherwise interfered with unspooling
  - Spool dimensions vary between suppliers, which can cause compatibility issues with spool holders
  - Customers mentioned reusing and recycling empty spools, including cardboard spools
  - Spool storage is a large concern, particularly for keeping filament dry
- Filament Packaging
  - Customers mentioned the use of vacuum bags with desiccant to keep filament dry during shipping
  - Shipping damage can lead to punctured vacuum bags and broken spools
- Filament Quality
  - Filament may become brittle and breaking during unspooling, or become brittle after printing
  - Poor dimensional accuracy or contamination can cause jams and irregular print quality
- Print Bed Adhesion
  - Various methods are described for improving adhesion of filament to the print surface, including tape, glue, hairspray and printing plastic "rafts"
- Print Settings
  - Customers frequently shared the settings that worked well for the filament on their 3D printer, often requiring adjustments to nozzle temperature, print bed temperature, printing speed, and fan settings
- Print Quality
  - Some filaments were noted as being unusually prone to forming blobs and strings
  - Some filaments had poor layer adhesion in which the plastic sticks poorly to itself

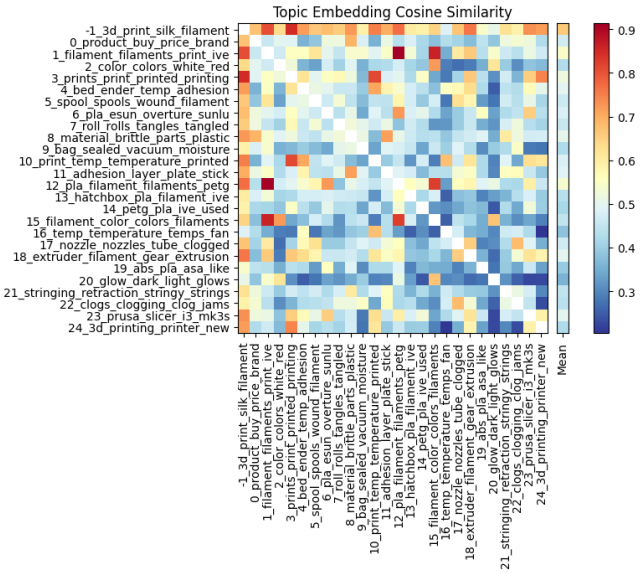

Figure 4: Comparison of topic sentence embeddings by cosine-similarity.

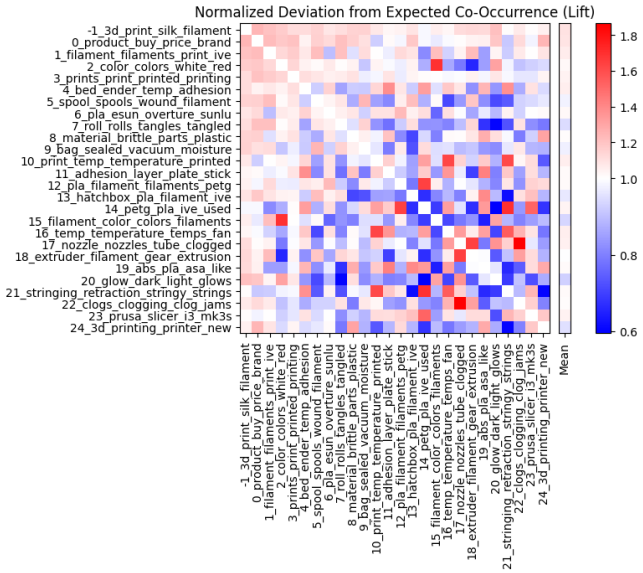

Figure 5: Comparing sentence topic co-occurrences rates within reviews using Lift.

- Bubbles forming during printing degrade surface finish and accuracy, which may result from wet filament, or poor manufacturing quality

- Filament Sellers
  - When mentioning brands customers frequently remarked on quality of customer service for resolving issues
  - Customers often mention issues with consistency of filament quality or performance when considering a different brand

3.2 Topic Comparison

Various methods can be used to compare the generated topics, allowing trends, patterns and outliers to be identified for further investigation.

3.2.1 Cosine Similarities

In **Figure 4** the cosine similarities of the topic embeddings have been compared, where each topic embedding is the average of all sentence embeddings assigned to that topic. A value of one would indicate topics with identical encodings, suggesting redundancy, or that they are very generalized and may benefit from being split into sub-topics. Values near zero indicate perpendicularity, or no correlation between embeddings, indicating that the topics probably cover very different concepts.

Most values in the plot have a similarity less than 0.5 (mean 0.468), which is reasonable, however there are some clear outliers near 0.9 (#1, 12, and 15). Manual review of these outliers revealed that the assigned sentences very frequently included the word "filament", but tended to describe different properties of the filament, or the same properties in different contexts. For example, the color of the filament might be mentioned in each, but in relation to availability with suppliers, color accuracy, or quality considering cost. They are therefore, not clearly redundant, but might benefit from being split into more specific sub-topics to clarify their differences.

3.2.2 Topic Pair Statistics

Since the topics were generated from the individual sentences found in each review, we can easily compare which topics appear together the most and least frequently. An appropriate method for this is to calculate the Lift, which estimates an expected co-occurrence rate for a pair, assuming independence of the topics, then compares it to the actual observed rate to highlight deviations:

$$Lift(A,B) = \frac{P(A \cap B)}{P(A) \cdot P(B)} = \frac{N_{A \cap B} \cdot N}{N_A \cdot N_B} \rightarrow \frac{Observed}{Expected}$$

Where **A** and **B** are topics, and **N** is the count of all observations.

These values are shown in **Figure 5**. Note that the scale is multiplicative, where a value of one indicates the expected co-occurrence rate based on the topic occurrences across all pairs. Lift is known to be sensitive to relatively rare classes (small topics), which can lead to exaggerated values for the associated pairs.

A statistical test should be performed to evaluate how confident we can be that these pairs are not just coincidences arising from the noise of the data. In this case contingency table X^2 tests were performed with an alpha of 0.05 and a Benjamini-Hochberg correction [16] to address possible false-positives in the large number of combinations (325) being evaluated. The null hypothesis was that occurrences of each topic are independent, while the alternative hypothesis was that occurrences are not independent. When collecting the pairs present in each review only unique combinations of topics were kept (no duplicate counts, or self-pairs), resulting in 2.92 unique topics per review on average, with a median of 2.0.

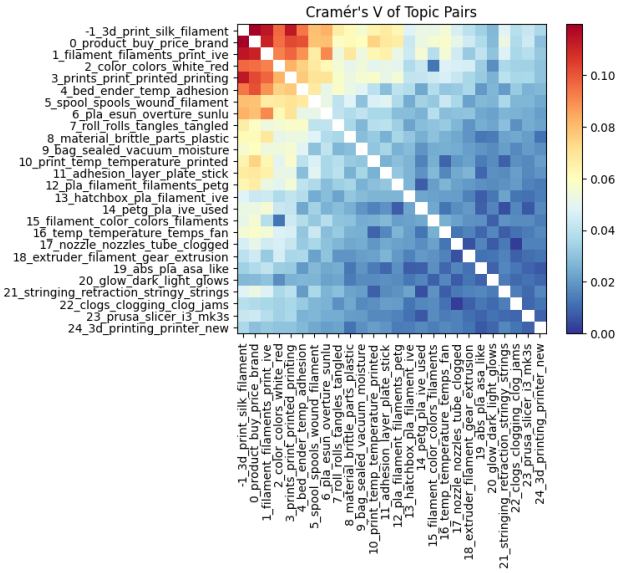

Figure 6: Comparing effect size using Cramér's V of the X^2 test for each topic pair.

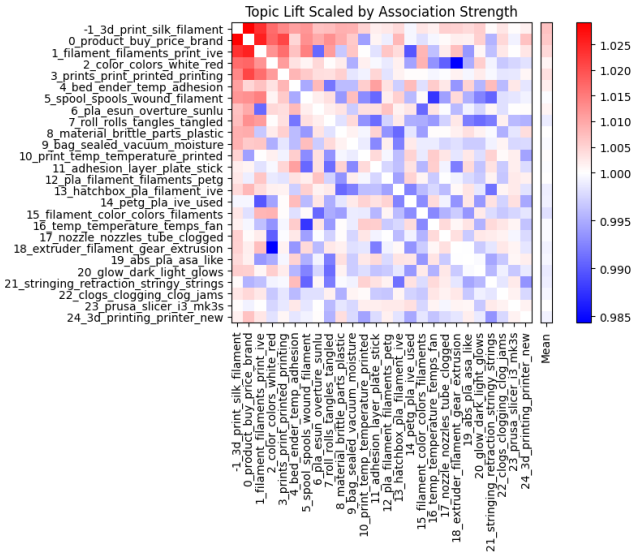

Figure 7: Lift of topic pairs scaled by association strength

As expected for a large population (approx. 547,000 observations), all but one of the pairs had very small p-values (<0.0006) well below the threshold, indicating rejection of the null hypothesis. The exception pair (Nozzles #17 and Clogging #22) had a p-value of 0.173, for which we fail to reject the null hypothesis. This is a surprising result since there *is* a logical reason for these topics to be paired frequently (of components that can become clogged, nozzles are the most typical). It's possible that the topics contain incompatible sub-topics, which interfere with their cohesion, or that the two topics are interchangeable with a simple rule splitting them (e.g. contains the word "nozzle"). Splitting, or merging the topics may help address this, though splitting them could increase the number of trivial pairings between nearly identical topics and merging simply makes them more generalized.

Applying the Benjamini-Hochberg correction revealed that for even a small false discovery rate (0.01) only the outlier pair exceed ed the calculated thresholds, indicating no change to the hypothesis conclusions.

While the Lift shows us which pairings have unusual frequencies and the X^2 text shows which are statistically significant, neither addresses the *strength* of the association between each topic. For this we can apply Cramér's V (or equivalently for 2x2 contingency tables, Cohen's ω), which is shown in **Figure 6**. Since Cramér's V can vary from zero (no association) to one (perfect association) it seems clear that all pairs are at best weakly associated (max 0.12).

Scaling the Lift by the association strength (**Figure 7**) highlights which relationships provide the most reliable and potentially interesting insights about the topic pairs.

$$Scaled\ Lift = V \cdot |Lift - 1| \cdot Sign(Lift) + 1$$

Where **V** is Cramér's V. We can see that the highest values are grouped in the top-left of the plot with the largest and most general topics. This makes sense if we assume that topics with heavily overlapping themes are more likely to be used together. Similarly, it would make sense for highly specific topics to rarely be used together, producing the wide spread of low values.

Using t-distributed Stochastic Neighbor Embedding (t-SNE) [17][13] we can produce a 2D representation of topic co-occurrences for easier comparison. Embeddings were generated from a distance matrix, which was calculated as the inverse of the Scaled Lift (high Lift pairs occur frequently together, so should be "closer"). The embedded positions are shown in **Figure 8**, where it can be seen that topics with high co-occurrences are roughly grouped. Domain knowledge reveals some logical clusters:

- Nozzles #17, Clogs #22, Extruder #18
    - The extruder assembly includes one or more nozzles, which are the most frequently clogged components. However, it was noted earlier that the pairing of topics #17 and #22 was not statistically significant.
- Color #2, Filament Color #15, Glow #20
    - All three are closely related to the appearance of the filament (color accuracy, surface finish, luminescence, etc.)
- Spool #5, Tangles #7, Bag #9
    - Filament tangling frequently originates from the spools, which are often shipped in vacuum-sealed bags to keep out moisture
- Temperature #16, Temperature #10, Bed #4, Stringing #21, Adhesion #11, PETG #14

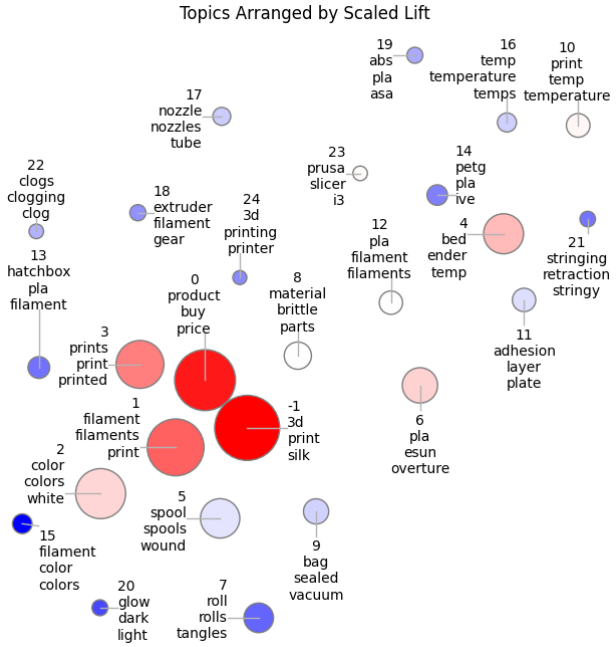

Figure 8: t-SNE encoded Scaled Lift for easier comparison by distance. Colors correspond to average topic Scaled Lift and size represents topic size.

- Various issues can be resolved through temperature adjustments, including print bed or print layer adhesion issues and stringing issues. PETG is known to be particularly prone to stringing.

It's also interesting to consider where topics have not been grouped, such as the color and temperature topics, which are as far separated as possible. One possible explanation is that they could have a near "expected" co-occurrence rate on average, or low association strength, allowing other, more strongly interacting topics to shift them apart.

For future analysis, it seems clear that large, generalized topics offer few interesting insights, but detecting meaningful interactions between small, specialized topics is not trivial. It may also be interesting to repeat this process, but with topic pairs grouped by product rather than review, which might help capture more interactions between small topics.

3.2.3 Topic Ratings

Understanding the tone of the topics can help us understand which facets of the products are perceived as the most and least desirable. Since each review includes a star rating from one to five, a simple method of estimating the tone of the topics is to average these ratings for each topic, where averages above the overall average are considered more "positive" and vice-versa. The values are shown in **Figure 9**, which highlights the distribution of ratings for each topic. For reference, the overall average is 3.7 stars, which most topics don't deviate far from.

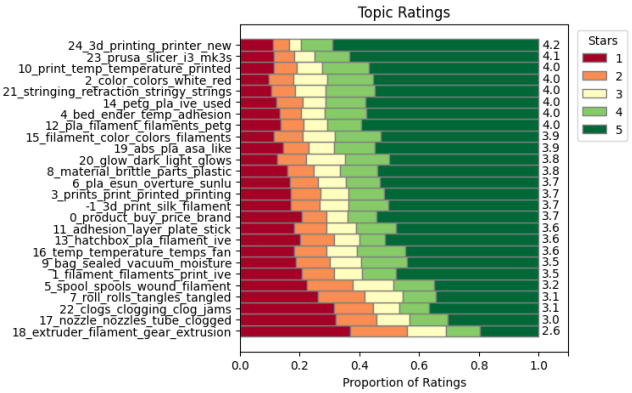

Figure 9: The distributions of topic ratings (1-5 stars) from reviews. Sorted by average ratings, which are shown to the right of the bars.

An interesting observation is that the one and five-star ratings usually make-up the largest proportions for each topic, suggesting that customers are easily polarized, or have a preference for these values. This raises the question of how well these values represent the true tone of the reviews, which could be estimated using sentiment analysis. Such analysis was not included in this project due to time constraints, but could be a suitable topic for future work.

When comparing the average ratings to the topic descriptions we can see that the intuitive expectations seem to match:

- #18, 17, 22, 7, and 5 are the five lowest rated topics and each include undesirable behaviors, such as tangling, jams and clogging.
- The highest-rated is #24, which tends to describe positive experiences with 3D printing, including buying items as gifts.

Extracting deeper insights and opportunities for improvement from these ratings may be aided by splitting them into sub-topics, which could reveal concepts like the aspect of extruder failures that's the most disliked. Comparing similar products with significantly different ratings could also be a useful method for identifying best-practices.

3.3 Insights

From the summaries and topic comparisons there are several insights and suggestions that can be extracted:

- One of the most significant properties for customers is the appearance of the filament, specifically the color. Judging the color of the filament from online photos can often be misleading, which might be improved by using a variety of lighting conditions, print samples, etc.
- Spools becoming tangled, or jammed significantly frustrates customers, requiring careful control of the winding process.

- Standardizing spool dimensions and offering recyclable cardboard spools may improve customer experiences.
- Filament that has absorbed moisture tends to have degraded qualities, but can be minimized through proper packaging in vacuum bags with desiccant.
- There may be opportunities to provide accessories with the filament, such as specialized tape and glue to improve bed adhesion, filament dryers, desiccant and dry-boxes to remove and prevent moisture accumulation, and filament spool holders.
- Print settings frequently require experimentation to dial-in, so providing a set of recommended settings to start with may improve user experience.
- Customers seem to prefer one-size-fits-all solutions where they can acquire all their filament from one supplier, but will switch to other brands if filament availability or quality becomes inconsistent. Improving these metrics may improve customer retention.

4 Conclusion

In this project a variety of statistical models were used to identify products related to 3D printer filament from a large dataset of Amazon products and accompanying reviews. The reviews associated with the products of interest were then split into sentences and the summarized common topics in those sentences extracted to identify concepts that are important to customers. Several actionable insights were then generated using various topic comparison methods, including evaluating topic pairing frequencies and manually reviewing high-probability samples. These insights could be used to aid filament sellers in improving their product offerings and customer experiences, which accomplished the main goal of the project.

This project highlighted many interesting trends and observations which would benefit from deeper exploration, making it a good starting point for further analysis.

Possible improvements and future work could include:

- Testing a variety of pre-trained LLMs to compare the performance of their encodings
- Testing different topic modeling techniques, such as NMF
- A method for sub-dividing product titles to identify where multiple products are present
- Additional filtering steps, such as separating the 3D printer pen products which may have different user expectations
- Specialized models for identifying the difficult-to-classify items, such as bundled products
- Dividing the topics into more specific sub-topics, particularly the largest and most general-purpose topics, which can be difficult to interpret
- Brand-specific, or product-specific analysis to identify unique strengths and weaknesses
- Sentiment analysis to identify topic tone and improve the quality of generated insights
- Analysis of topics over time to identify where issues have already been resolved and the emergence of new trends

REFERENCES

[1] Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, & Julian McAuley. (2024). Bridging Language and Items for Retrieval and Recommendation.

[2] Maarten Grootendorst. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure.

[3] Victor Sanh, Lysandre Debut, Julien Chaumond, & Thomas Wolf. (2020). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.

[4] Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res., 3(null), 993–1022.*

[5] Pentti Paatero; Unto Tapper (June 1994). "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values". Environmetrics. 5 (2): 111–126. doi:10.1002/ENV.3170050203. ISSN 1180-4009. Wikidata Q29308406.

[6] Abhinandan Krishnan, & Abilash Amarthaluri. (2019). Large Scale Product Categorization using Structured and Unstructured Attributes.

[7] Nils Reimers, & Iryna Gurevych. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.

[8] Leland McInnes, John Healy, & James Melville. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.

[9] Leland McInnes, John Healy, & Steve Astels (2017). hdbscan: Hierarchical density based clustering. *Journal of Open Source Software, 2(11), 205.*

[10] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

[11] Webb, G.I. (2011). Naïve Bayes. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_576

[12] Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, *20*(2), 215–232.

[13] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12, 2825–2830.*

[14] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, & Alexander M. Rush. (2020). HuggingFace's Transformers: State-of-the-art Natural Language Processing.

[15] Reimers, N., & Gurevych, I. (2021). *all-MiniLM-L12-v2* [Computer software]. Hugging Face. https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2

[16] Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple hypothesis testing. J R Stat Soc B 57:289–300

[17] L.J.P. van der Maaten and G.E. Hinton. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9(Nov):2579-2605, 2008.

Appendix

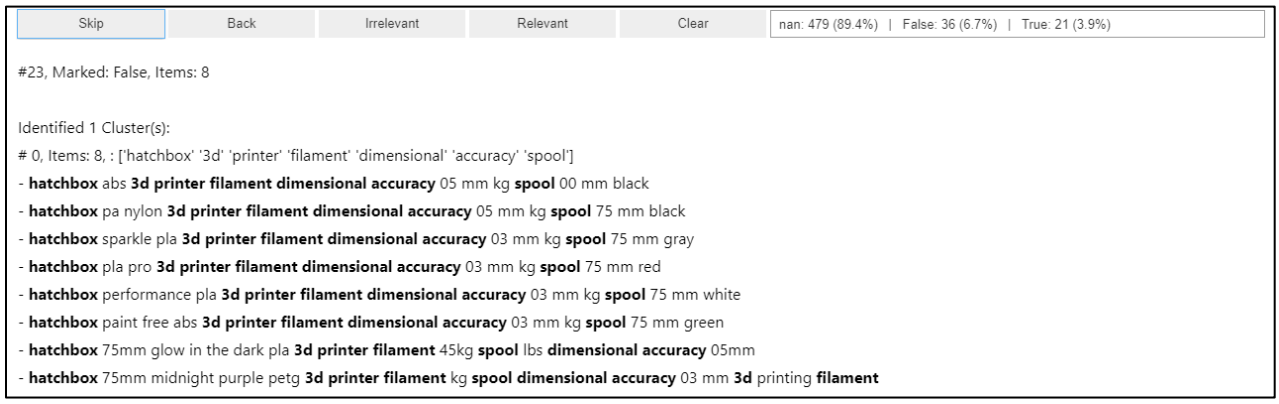

Figure 3: The GUI used to streamline manual cluster labeling.

Table 1: Mean F1-Scores of each classifier and text-encoding pair across each level of manual dataset labeling.

Encoding	Algorithm	Portion of Manually Labeled Training Data									
		10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
sentence transformer	Bayes	**0.9628**	**0.9757**	0.9709	0.9692	0.9698	0.9697	0.9702	0.9678	0.9677	0.9678
	Logistic	0.9137	0.9142	0.9180	0.9243	0.9300	0.9368	0.9478	0.9560	0.9637	0.9674
	PD	0.9395	0.9568	0.9627	0.9653	0.9673	0.9758	0.9785	0.9806	0.9828	0.9817
	SVM	0.9175	0.9488	0.9701	0.9790	0.9833	0.9855	0.9872	0.9877	0.9888	0.9894
topic vectors	Bayes	0.9137	0.9137	0.9156	0.9199	0.9290	0.9443	0.9548	0.9605	0.9711	0.9726
	Logistic	0.9189	0.9310	0.9447	0.9487	0.9574	0.9642	0.9688	0.9747	0.9785	0.9801
	PD	0.9174	0.9211	0.9251	0.9269	0.9299	0.9333	0.9368	0.9388	0.9418	0.9428
	SVM	0.9142	0.9228	0.9364	0.9453	0.9580	0.9656	0.9736	0.9757	0.9785	0.9790
sentence transformer	ensemble	0.9387	0.9625	**0.9737**	**0.9801**	**0.9849**	**0.9877**	**0.9888**	**0.9894**	**0.9905**	**0.9911**

Key
column max
column min
other

Table 2: Standard deviations of F1-Scores of each classifier and text-encoding pair across each level of manual dataset labeling.

Encoding	Algorithm	Portion of Manually Labeled Training Data									
		10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
sentence transformer	Bayes	0.0135	0.0168	0.0182	0.0142	0.0123	0.0142	**0.0160**	**0.0165**	**0.0189**	0.0137
	Logistic	0.0150	0.0156	0.0175	**0.0241**	0.0181	**0.0180**	0.0156	0.0159	0.0131	0.0123
	PD	0.0141	0.0172	0.0142	0.0167	0.0154	0.0123	0.0100	0.0097	0.0085	0.0091
	SVM	**0.0164**	**0.0268**	0.0205	0.0154	0.0121	0.0093	0.0083	0.0082	0.0087	0.0080
topic vectors	Bayes	0.0150	0.0150	0.0169	0.0189	0.0183	0.0136	0.0114	0.0115	0.0081	0.0062
	Logistic	0.0153	0.0177	0.0144	0.0171	0.0146	0.0148	0.0122	0.0116	0.0096	0.0106
	PD	0.0145	0.0166	0.0146	0.0153	0.0154	0.0153	0.0139	0.0135	0.0143	**0.0164**
	SVM	0.0156	0.0186	**0.0217**	0.0238	**0.0191**	0.0147	0.0121	0.0113	0.0105	0.0079
sentence transformer	ensemble	0.0140	0.0194	0.0163	0.0154	0.0109	0.0102	0.0083	0.0084	0.0087	0.0091

Key
column max
column min
other

Table 3: The extracted review sentence topics, their occurrence counts, representative words, and manually generated descriptions.

Count	Topic Name	Representation	Description
65332	-1_3d_print_silk_filament	3d, print, silk, filament, printing, prints, pen, printer	Uncategorized
57296	0_product_buy_price_brand	product, buy, price, brand, good, great, stuff, quality	Mixed tone, typically mentions whether they will buy again, or the price.
49910	1_filament_filaments_print_ive	filament, filaments, print, ive, prints, good, great, quality	Include the word "filament", mixed tone, mentioned topics include price, breaking filament, dimensional accuracy, color, and clogging.
38631	2_color_colors_white_red	color, colors, white, red, black, nice, green, blue	Mention texture/surface finish (matte and glossy), color accuracy, and transparency
35343	3_prints_print_printed_printing	prints, print, printed, printing, printer, great, quality, good	Describe how the filament interacts with the printer, such as binding, stringing, good/poor quality, adhering to print bed, adjusting settings.
24746	4_bed_ender_temp_adhesion	bed, ender, temp, adhesion, nozzle, glass, stick, pro	Frequently mention a Creality Ender printer (3, Pro, etc.), whether filament sticks to the print bed and temperature settings
23979	5_spool_spools_wound_filament	spool, spools, wound, filament, tangled, holder, ive, just	Comments about, or mentioning spools. Includes spool dimensions, filament tangling, spool reuse, and storage
19296	6_pla_esun_overture_sunlu	pla, esun, overture, sunlu, regular, ive, used, prints	Typically mention PLA filament and/or one of three brands (esun, overture, or sunlu). Mention customer service, switching brands, and temperature settings
13570	7_roll_rolls_tangles_tangled	roll, rolls, tangles, tangled, got, bought, just, half	Frequently negative comments about filament rolls, mentioning color accuracy, re-spooling, tangling/knotting
11449	8_material_brittle_parts_plastic	material, brittle, parts, plastic, strong, projects, flexible, gift	Mixed tone. Several comments about brittleness, several about giving as a gift
9840	9_bag_sealed_vacuum_moisture	bag, sealed, vacuum, moisture, dry, box, packaged, desiccant	Frequently mention moisture/drying/desiccant/humidity, packaging, vacuum bags
8641	10_print_temp_temperature_printed	print, temp, temperature, printed, printing, speed, prints, temps	Mention the printing temperature they used, often positive tone (e.g. "this is what worked:")
8467	11_adhesion_layer_plate_stick	adhesion, layer, plate, stick, layers, glue, build, warping	Adhesion to the print bed, warping, and methods to remedy, such as glue, tape, and hairspray. Also layer adhesion
8395	12_pla_filament_filaments_petg	pla, filament, filaments, petg, solutech, 3d, esun, overture	Typically mentions PLA filament and popular brands. Mention filament being in-stock
7322	13_hatchbox_pla_filament_ive	hatchbox, pla, filament, ive, used, filaments, better, brand	Frequently mention the hatchbox brand, with mixed comments about preference over other brands. Some complaints about their matte filament.
6485	14_petg_pla_ive_used	petg, pla, ive, used, print, printing, like, prints	Usually mention PETG filament. Mention stringing and sticking to print bed
5982	15_filament_color_colors_filaments	filament, color, colors, filaments, white, black, rainbow, matte	Filament color, color accuracy, multi-color
5618	16_temp_temperature_temps_fan	temp, temperature, temps, fan, tried, temperatures, heat, cooling	Various comments about printing temperature adjustments
4988	17_nozzle_nozzles_tube_clogged	nozzle, nozzles, tube, clogged, bowden, brass, clog, clogs	Comments about nozzles, such as clogging, swelling, feeding, stringing, temperature, nozzle size and hardened nozzles
4026	18_extruder_filament_gear_extrusion	extruder, filament, gear, extrusion, print, tube, bowden, extruders	Mention the extruder, jamming/clogging/grinding
3971	19_abs_pla_asa_like	abs, pla, asa, like, print, printing, filament, warping	Mention ABS, PLA, or ASA. Mention bubbling, humidity, layer strength, support removal, cost, color, warping
3918	20_glow_dark_light_glows	glow, dark, light, glows, glitter, bright, uv, glowing	Glow in the dark filament, mention duration and brightness
3750	21_stringing_retraction_stringy_strings	stringing, retraction, stringy, strings, speed, little, string, settings	Most mention stringing, or blobs. Mention delamination, print speed, temperature, clogging
3306	22_clogs_clogging_clog_jams	clogs, clogging, clog, jams, clogged, bubbles, hot, end	Mention clogging, bubbles, moisture, feeding issues, popping, brittleness
3286	23_prusa_slicer_i3_mk3s	prusa, slicer, i3, mk3s, mk3, settings, mini, using	Mention several 3D printer brands, slicing settings
3036	24_3d_printing_printer_new	3d, printing, printer, new, printers, im, years, works	Various, generally positive remarks about 3D printing