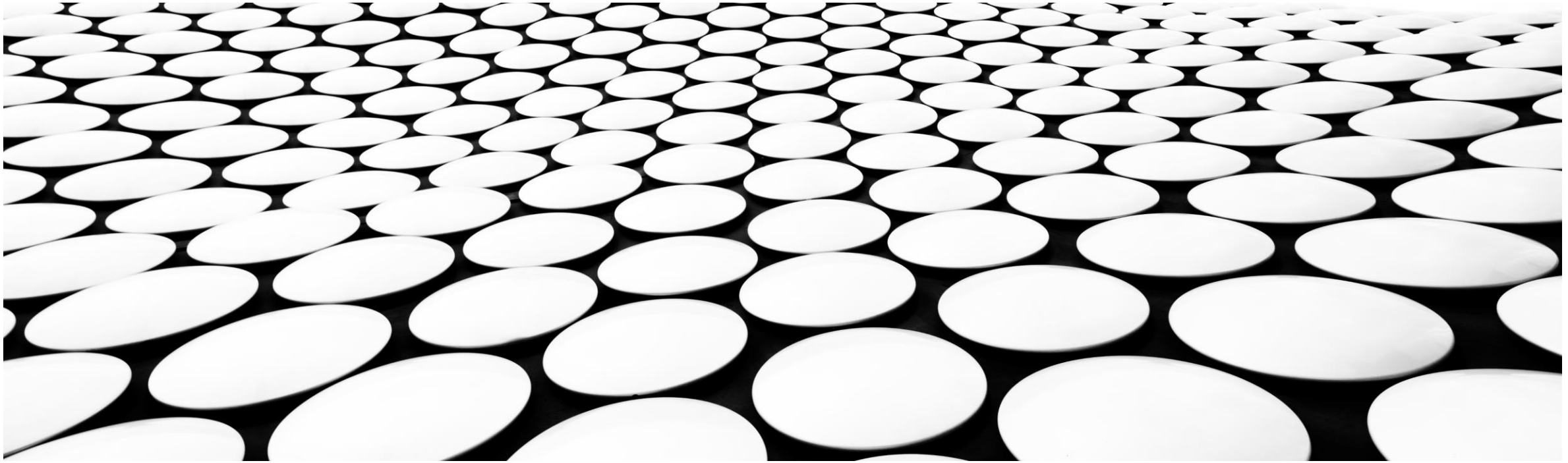
3D PRINTER FILAMENT REVIEW TOPIC ANALYSIS

DTSA 5506 DATA MINING PROJECT



BACKGROUND

- Goal: extracting useful insights from e-commerce data
- Dataset: large volumes of Amazon data are available
  - AMAZON REVIEWS 2023
    - Includes reviews and product metadata from 1996 – 2023
    - 48 million items, 570 million reviews
- Interested in just 3D printer filament products
  - How do we extract just these products? (not clearly categorized)
  - What are the common topics in reviews?

By Maurizio Pesce from Milan, Italia - 3D Printing Materials, CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid=51016982

RELATED WORK

- Topic modeling is a common task
  - Can be used for product categorization, as well as identifying review topics
  - Various models available (LDA, NMF, BERTopic, Multi-LSTM/CNN)
  - BERTopic preferred for “understanding” language and streamlined implementation
  - Assumes a single topic per document, but reviews can be split into sentences if needed

- BERTopic Overview:
  1. Document embedding using a pre-trained LLM (Sentence Transformer)
  2. Dimensionality reduction (UMAP)
  3. Creation of topic groups via clustering (HDBSCAN)
  4. Topic extraction (cTF-IDF)

RELATED WORK (CONT.)

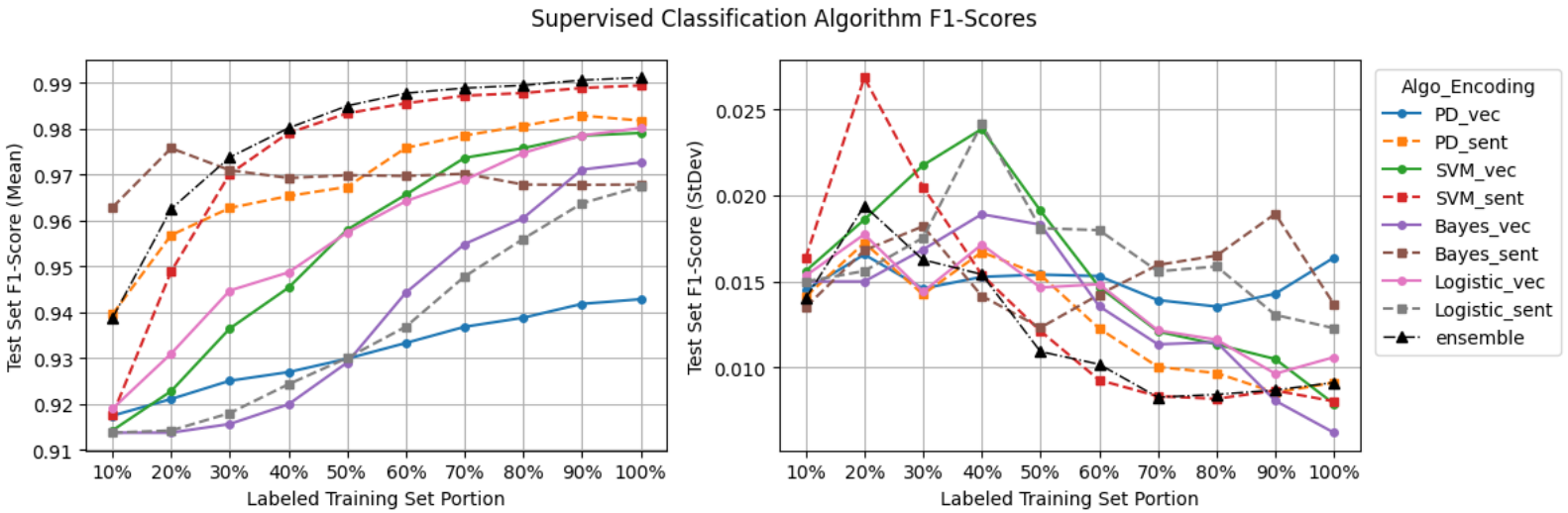
- Filtering the dataset to just 3D printer filament is not trivial
- Supervised models can be used to predict which products are relevant
  - Pairwise Distances, Support Vector Machines, Bernoulli Naïve Bayes, and Logistic Regression
- Text encoding method must be considered
  - Sentence-Transformers (language-model based)
  - Simple token-vector encoding (mark token as present or not)
- Experimentation required to determine best combination

DATASET FILTERING

- 427,000 products in “Industrial & Scientific” category reduced to:
  - 7,000 products by keyword filtering (“filament” and any known plastic type)
  - 2,800 products by minimum five associated reviews

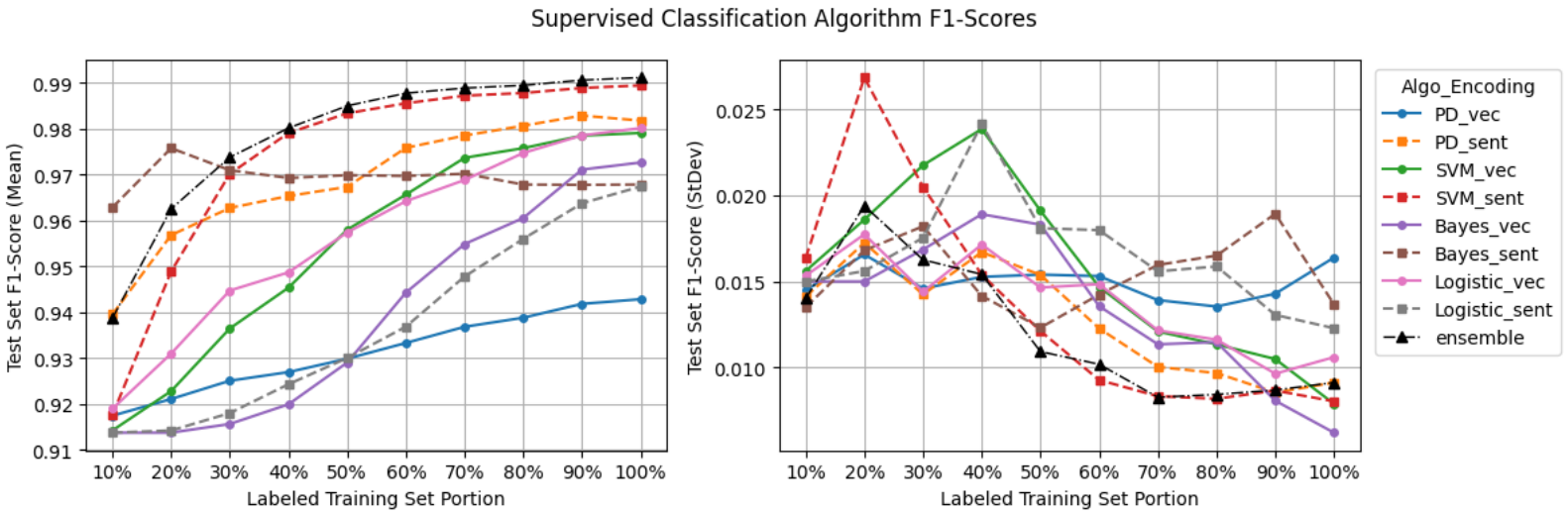
- Example product titles (difficult to manually filter):
  - *1.75MM Filament PLA Refills, Jekon PLA Filament for 3D Pen/3D Printer 1.75mm 20 Colors One Pack, Each Color 33 feet, 660 feet in Total*
  - *#1 Best Filament ABS Black 1.75 mm +/-0.02mm Top Accuracy, 3D Printer Spool Extruder Holder Stand, 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*

DATASET FILTERING (CONT.)



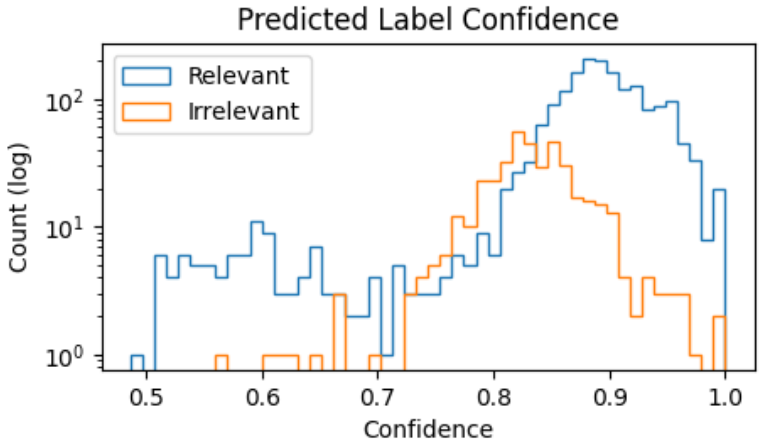
- Further filtering required supervised classification (relevant or irrelevant products)
  - Manually labeled 20% of data for training and validation
  - Training / testing split: 80% / 20%
- Training data divided into 10% intervals
- F1-Score used as evaluation metric
- Means and standard deviations aggregated from 10 randomized trials

DATASET FILTERING (CONT.)



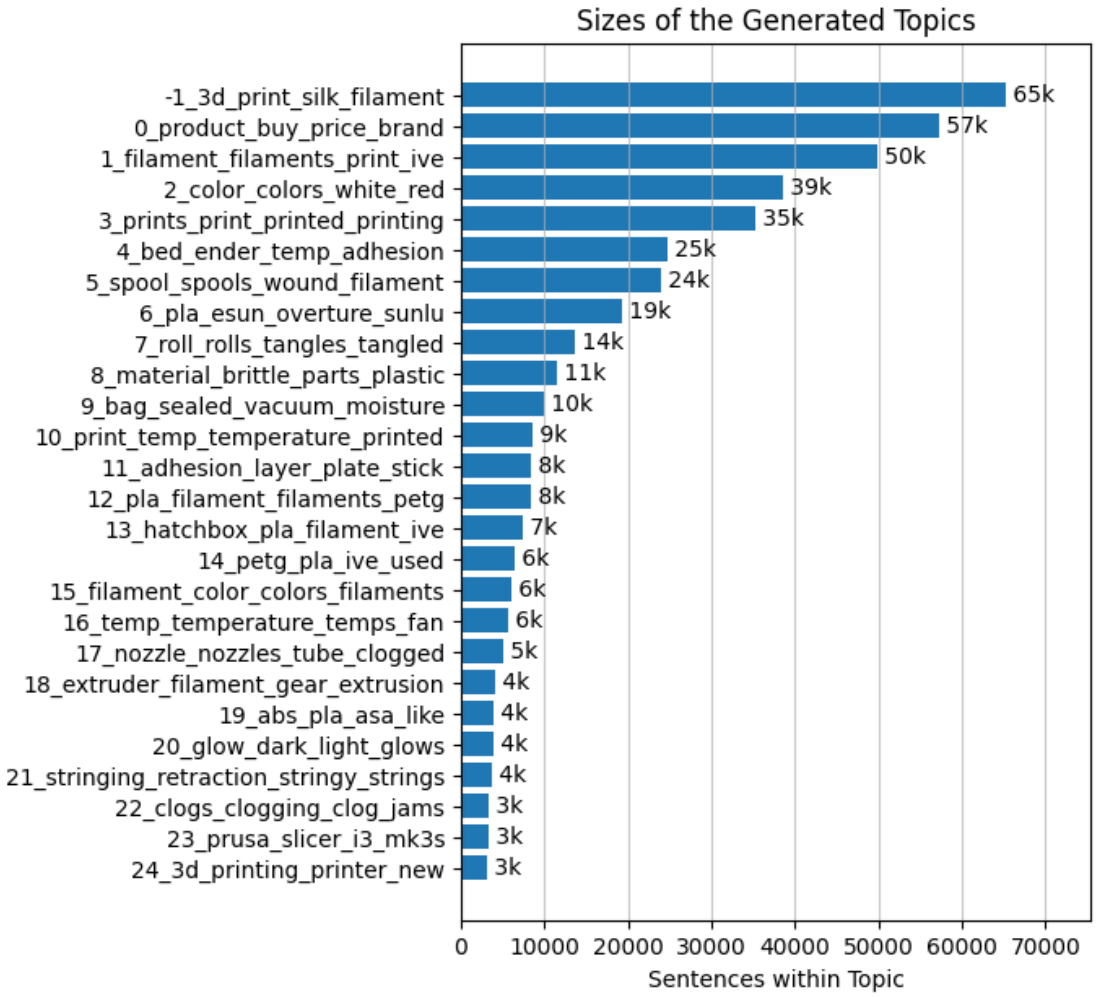
- Sentence Transformer encodings outperformed Token-Vector encodings
- Ensemble of Pairwise Distances, SVM, and Binomial Naïve Bayes selected
  - Used Sentence-Transformer encoding
- Dataset reduced further to 2341 products (84.1% of candidate products)

DATASET FILTERING (CONT.)



- Model prediction confidence used to identify likely misclassification
  - Confidence estimated as highest predicted class probability
- 126 titles with “low” confidence (<0.75)
  - Manual review found 16 misclassifications
- Re-trained model with additional manual labels
  - New prediction of 2291 relevant products (82.4% of candidate products)

REVIEW TOPIC MODELING



- 117,000 reviews associated with selected products
- Reviews split into 463,000 sentences to reduce complexity
- BERTopic used to extract topics
- Number of topics requires careful consideration
  - Could produce many very similar topics, few very general topics, or anything in-between
- Must inspect associated sentences to understand topic theme

COMMON CONCEPTS (*What are people talking about?*)

- Filament Appearance
  - Color accuracy compared to images online
  - Matte filament has worse properties
  - Inconsistent surface finish
  - Glow in the dark filament
- Filament Spools
  - Poor spool design and winding causes issues
  - Spool dimensions are not standardized
  - What to do with empty spools
  - Keeping out moisture during storage

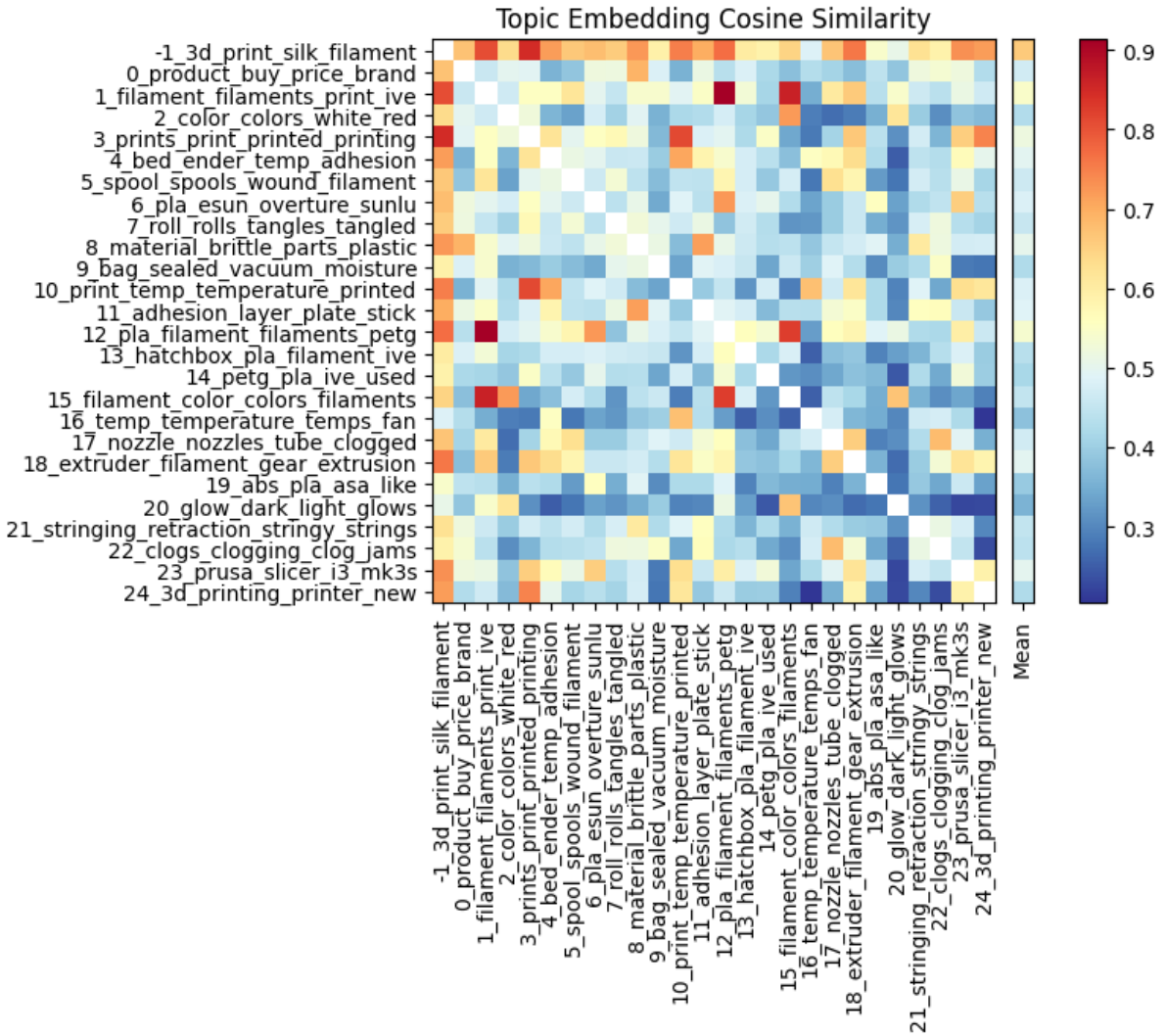
COMMON CONCEPTS (CONT.)

- Filament Packaging
  - Vacuum bags with desiccant
  - Shipping damage
- Filament Quality
  - Brittleness before, or after printing
  - Dimensional accuracy and contamination
- Print Bed Adhesion
  - Filament might need help sticking to print surface
- Print Settings
  - Shared their testing and adjustments for best results

COMMON CONCEPTS (CONT.)

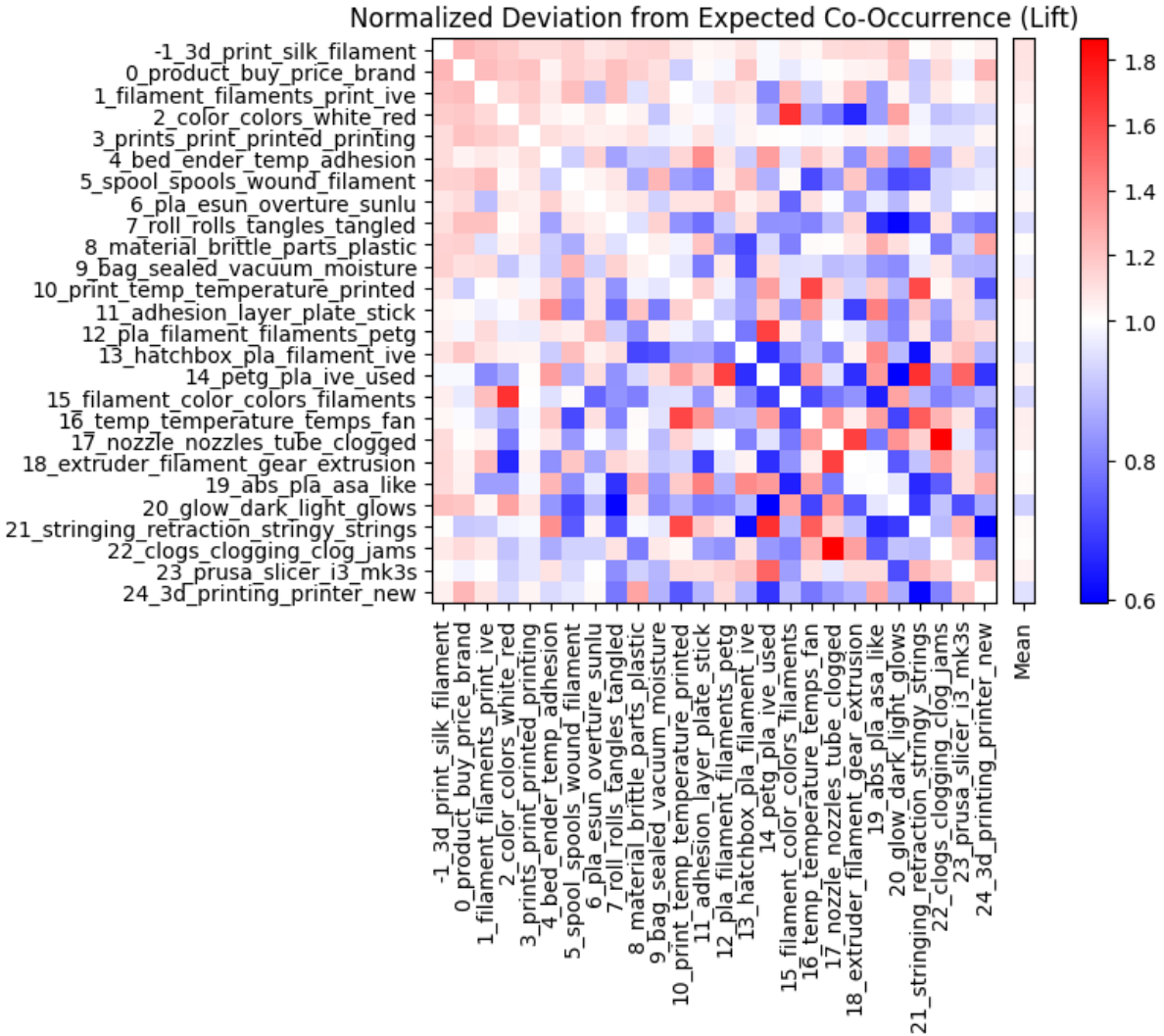
- Print Quality
  - Blobs and strings
  - Poor layer (self) adhesion
  - Bubbling filament
- Filament Sellers
  - Customer service
  - Brand loyalty

TOPIC INSPECTION



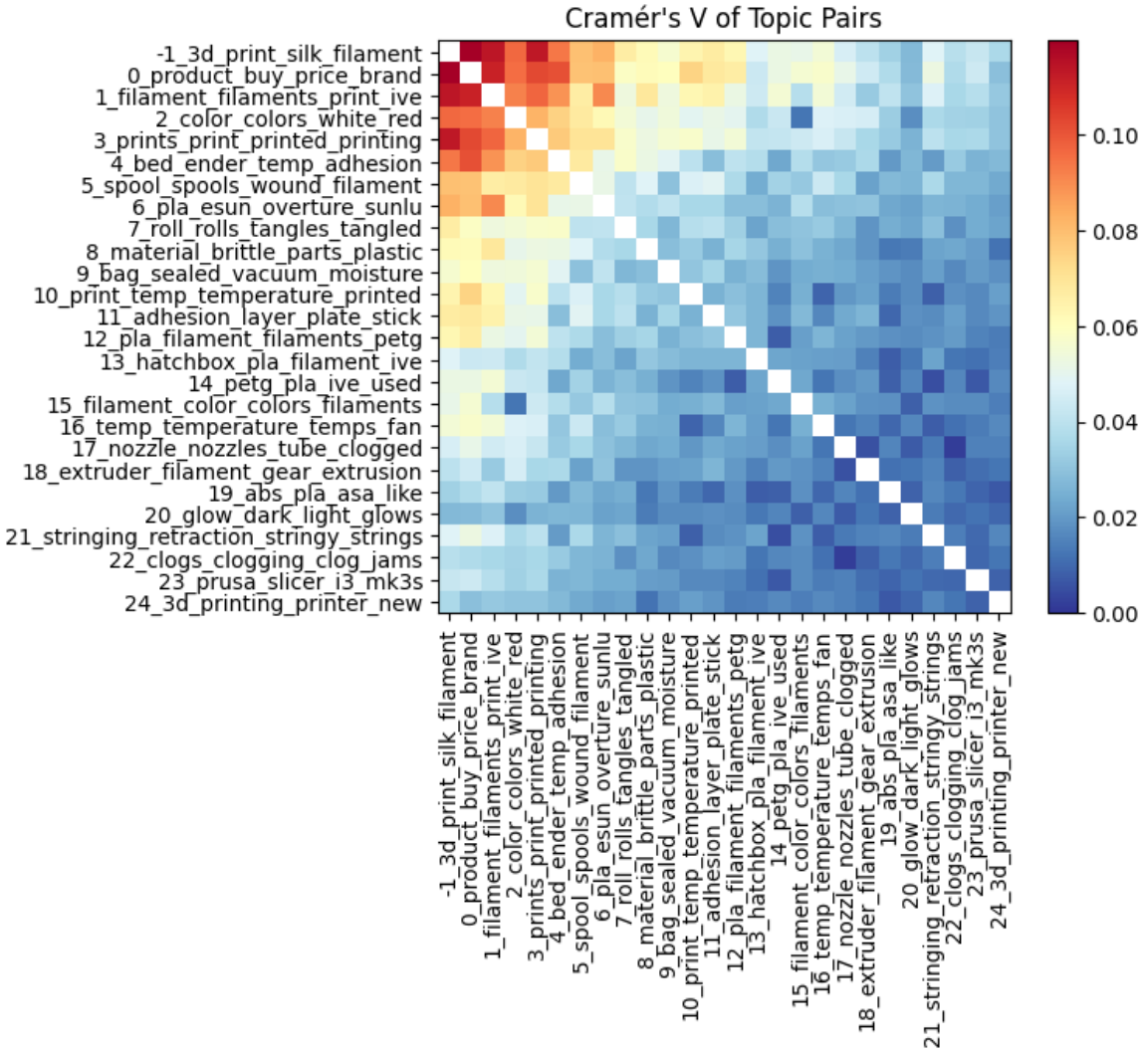
- How similar are the topic embeddings?
  - Compared by Cosine Similarity
  - Helps identify which topics might be too general, or have large overlap
- Most are reasonably dissimilar (≤ 0.5)
- Some have strong similarity (1, 12, 15)
  - Not necessarily redundant
  - They each mention color, for example, but in different contexts

TOPIC INSPECTION (CONT.)



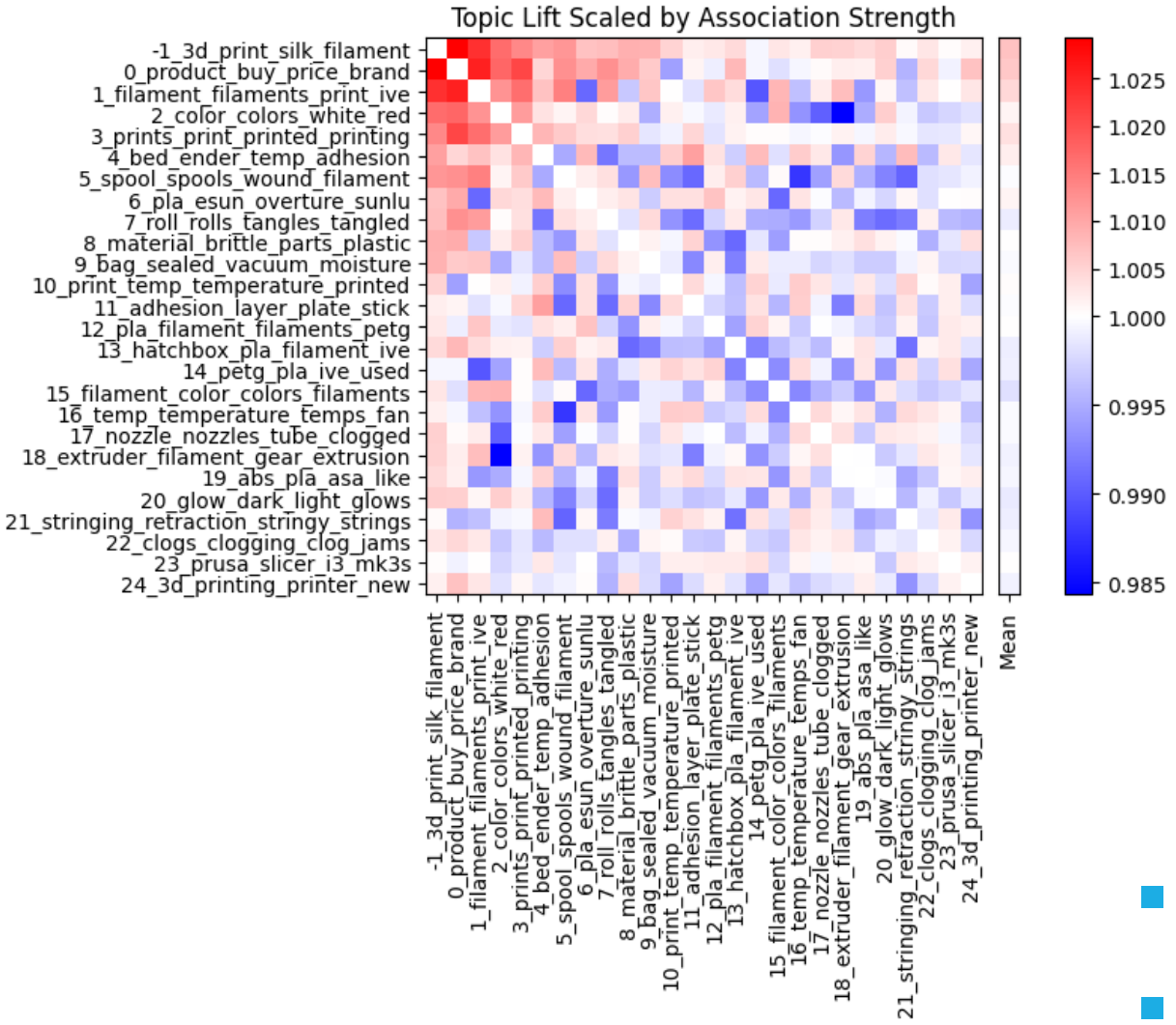
- Lift highlights which pairs occur more or less frequently than expected
- 1.0 represents the expectation
- Sensitive to small topics, which can be exaggerated

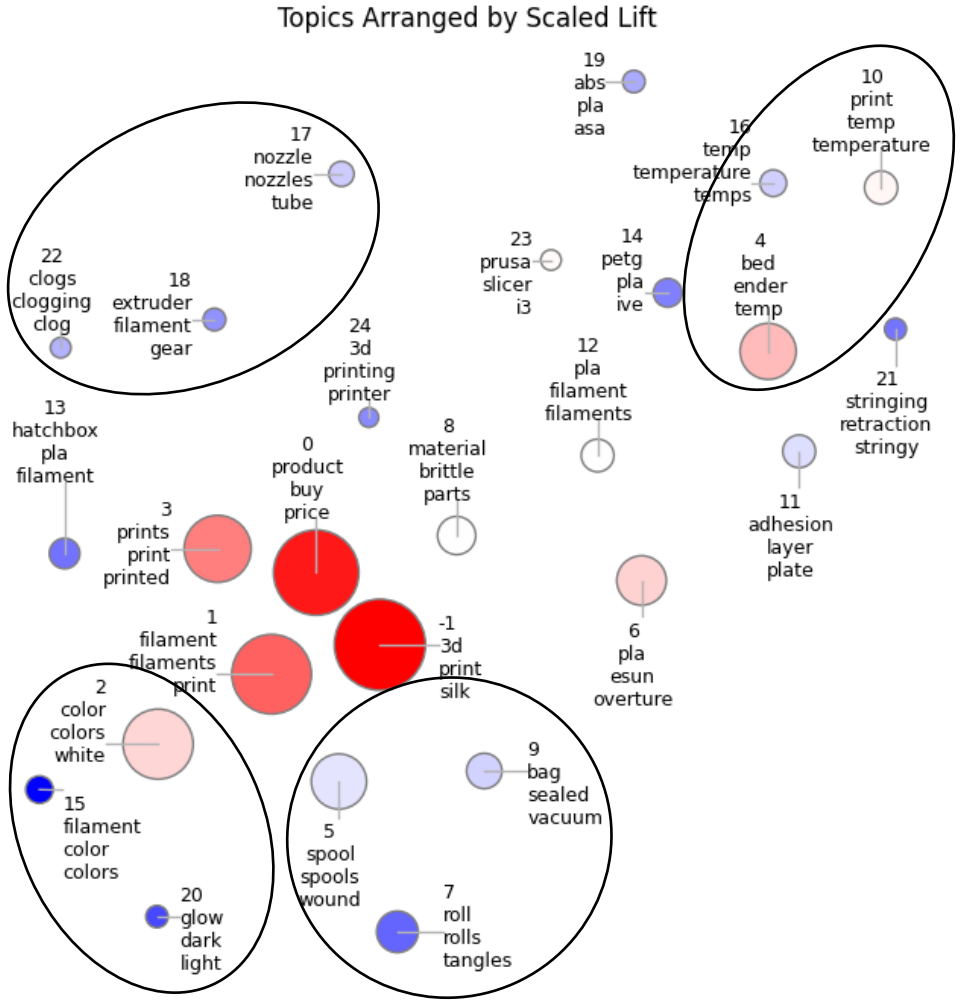
TOPIC INSPECTION (CONT.)



- Statistical significance of observed pairs
  - Contingency table χ^2 test of independence ($\alpha = 0.05$)
    - p-values <0.0006, one exception for pair (17, 22) with p-value 0.173
  - Benjamini-Hochberg Correction for large number of pairs (325), (Q = 0.01)
    - No change to hypothesis conclusions
  - Cramér's V to calculate strength of association
    - All relatively weak (< 0.12)
  - Observations make sense, but are not strong

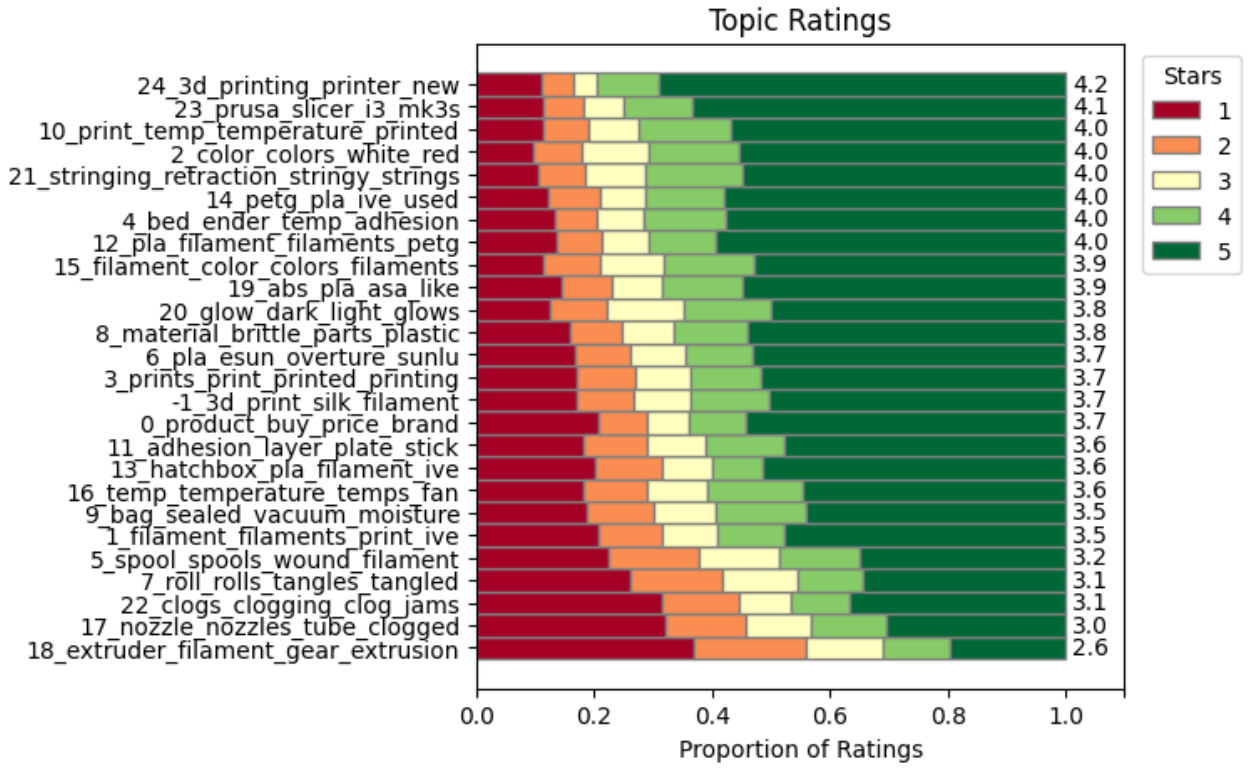
TOPIC INSPECTION (CONT.)





- t-SNE encoding to 2D representation
- Loose clusters that have general themes

TOPIC INSPECTION (CONT.)



- What are the tones of the topics?
  - Positive/negative/neutral?
- Can be estimated using review star ratings
- Lowest rated topics describe clearly undesirable behavior
  - Clogging, jamming, tangling, brittleness, etc.
- LLM-driven sentiment analysis could give a second opinion

INSIGHTS

- Appearance, particularly color, is important, but hard to judge.
- Customers are quickly frustrated by poorly wound spools and filament out of dimensional tolerance.
- Standardized, recyclable spools may improve customer experience.
- Moisture is a large factor in print quality, but can be mitigated with proper packaging.
- Additional products to address common problems could be offered (glue, filament driers, etc.)
- Providing suggested print settings could help users get the best performance.
- Consistent quality and availability may be vital for customer retention.

CONCLUSION

- The original goal was to extract actionable insights about 3D printer filament from product review data
- Relevant products were extracted from the dataset using manual filtering and supervised classification
- Topics were extracted from reviews, manually reviewed, compared through various statistics, then used to generate insights
- Further insights could be generated with deeper review

IMPROVEMENTS AND FUTURE WORK

- Other LLM encodings
- Other topic modeling techniques, such as NMF
- Identifying where multiple products are present in a title
- Separating 3D printer pen filament
- Specialized models to identify bundled products
- Finer topic granularity and size balance
- Comparing similar products and brands
- Sentiment analysis

END

Thank You