**Beauty product sentiment analysis and Recommendation Based on User Behaviour**

**Dibya Pandey Adhikari**

**MPS – Data Science**

**Prof.** [Muhammad Ali Yousuf](mailto:maliyou1@umbc.edu)

**Date: Dec 20, 2023**

# ABSTRACT

In the ever-evolving beauty industry, navigating through the myriad of available products is daunting. Recommender systems have emerged as crucial aids, guiding consumers to beauty products that align with their unique preferences and needs. Within this realm, recommendation systems and sentiment analysis of product reviews have garnered substantial interest among researchers, focusing on crafting systems that suggest products based on user preferences, simplifying the search process. Social data, especially reviews, plays a pivotal role in these recommendations.

This project focuses on developing an efficient deep learning-based recommendation system by handling text data from Sephora product reviews. It employs preprocessing techniques like stemming, lemmatization, and stop word removal to streamline data interpretation and utilization. These methods help eliminate inconsistencies and duplicates, reducing computational time and enhancing efficiency. Subsequently, the project leverages natural language processing techniques such as summarization, sentiment analysis, named entity recognition, and topic modeling to extract valuable insights from the top 3 moisturizer reviews.

The approach involves using unsupervised extractive summarization to generate concise overviews, highlighting key information. It then delves into contextual analysis to identify sentiments, entities, relationships, and latent topics. Visual aids like word clouds and knowledge graphs reveal linguistic patterns, while hierarchical clustering uncovers thematic structures within the reviews. Additionally, an ensemble model predicts sentiment from the summaries, efficiently gauging perspectives at scale.

These methodologies offer both comprehensive summaries and a detailed analytical understanding of complex review datasets, contributing to product recommendations based on user reviews. The project showcases the value of NLP in accelerating the discovery and comprehension of beauty products. This synthesis of disparate evidence enables swift, data-driven decisions in the industry. Overall, the project demonstrates a pipeline that integrates summarization, contextual analysis, relationship extraction, and predictive modeling to extract transformative insights from vast textual data in the beauty industry.

# INTRODUCTION

This project aimed to apply natural language processing techniques to analyze and extract key insights from a large corpus of reviews of beauty products. The objectives were to:

1. Provide concise overviews of reviews through summarization.
2. Identify contextual details like sentiments, entities, and topics.
3. Discover relationships and patterns in the data.
4. Build predictive models using the extracted information.
5. Recommend products on the basis of reviews.

The techniques delivered both high-level summarized views as well as a deeper contextual analysis of the complex review dataset.

# THE IMPORTANCE OF TEXT SUMMARIZATION

Text summarization, the technique of shortening long pieces of text while retaining the overall meaning, has become extremely valuable in the modern information age. With the explosion of online content, social media, and written information in all spheres of life, the ability to quickly distill the key ideas and salient points from lengthy documents saves a great amount of time and effort. Beyond direct time savings, it can serve as an unbiased tool for gathering information, better than biased approaches like surveys, focus groups, manually curated Q&A, or interviews. Digesting textual data from diverse sources over long periods eliminates recency bias and provides objective summaries. Text summarization provides tremendous value in efficiency and understanding across many professional domains. Whether it is summarizing research papers, news articles, legal contracts, medical records, or any form of communication, text summarization provides an efficient way to filter out the noise and focus only on the core ideas. For professionals like lawyers, doctors, researchers, analysts, and journalists who have to sift through massive amounts of text daily as part of their work, automated summarization tools help accelerate understanding and decision-making by pinpointing the essence of the material. For portfolio managers in finance, digesting earning reports and financial documents into their main takeaways accelerates investment decision-making. Academic researchers rely on summarization to grasp the essence of long research papers and extract key insights.

With artificial intelligence techniques like deep learning advancing the capabilities of text summarization, what was once only manually possible is now feasible automatically. This assists professionals, students, businesses, and ordinary people to save time and effort while improving the consumption and transmission of ideas and information in modern life full of digital text. The automated nature of text summarization makes it more neutral than manually curated Q&A or interviews. The wide applicability across domains combined with its unbiased nature make text summarization an invaluable asset for productivity and reliable insights in the modern data-driven world.

# THE IMPORTANCE OF BEAUTY PRODUCTS AND THEIR RECOMMENDATION

E-commerce websites have expanded globally, minimizing geographical limitations and time constraints. Predicting customer needs accurately on this scale is a daunting task, prompting these platforms to prioritize customer satisfaction through predictive approaches. Analyzing vast databases manually and summarizing textual content is time-intensive and computationally demanding. Thus, employing text mining and natural language processing techniques becomes crucial to extracting and scrutinizing customer sentiments or opinions from numerous reviews[1].

The beauty industry stands as a dynamic and ever-evolving landscape, characterized by an overwhelming array of products, each promising to fulfill the unique desires and needs of consumers [2]. Amid this profusion of choices, recommender systems have emerged as indispensable tools, aiding consumers in navigating the intricate world of beauty products. These systems provide personalized product recommendations, aligning offerings with individual preferences, and streamlining the shopping experience[3].

# ABOUT THE DATA

In this project, I embark on a data-driven journey, seamlessly weaving together data analysis and sentiment exploration to delve into the world of beauty product recommendations [4]. My project relies on a rich dataset encompassing an extensive array of beauty products, user feedback, and reviews [5]. This foundational dataset serves as the bedrock for the recommendation system. It allows us to not only understand the intricacies of the beauty product landscape but also to fine-tune our recommendation models for optimal personalization[6].

# METHODOLOGY

1. **Data Collection**

This dataset from Kaggle is like a big collection of info all about the stuff you can buy at Sephora, like makeup, skincare, and more. It's got details about what things cost, what they're made of, and what people think about them. People write reviews saying if they like something or not, and they give ratings too. This data helps us understand what people prefer and how different products are doing. It's super helpful for businesses to make better choices and make customers happier.

1. **Data pre-processing**

Text preprocessing like removal of stop words, punctuation, and lemmatization normalized the data by filtering out non-essential words and reducing remaining words to their root forms. This improved matching and analysis.

1. **Text Summarization**

The project implements an unsupervised extractive text summarization technique to distill key information from large texts into concise overviews. Text summarization is essential for obtaining representative summaries covering the main topics and diverse aspects of a source text.

The approach first preprocesses the input text through steps like lowercasing, lemmatization, and stop word removal to normalize the text and focus on meaningful keywords. Word frequencies are then generated and normalized to determine the information content of each non-stop word.

Sentence importance scores are calculated by aggregating normalized word frequencies. This enables quantitatively identifying sentences containing more salient information. By scoring sentences based on collective word weights, the model accounts for informative keywords that may be distributed across different parts of a sentence.

The core summarization function extracts the top N-ranking sentences by importance into the final summary. By extracting sentences directly from the source, the model retains key details and diverse coverage of topics in a condensed overview.

This unsupervised technique distills informative content relevant to the main themes and discourse in the text. The data-driven word frequency analysis impartially identifies key details without manual supervision or labeling. The extracted sentences provide a representative summary sample covering the salient information in the source text.

1. **Sentiment analysis and subjectivity detection**

Sentiment analysis and subjectivity detection are critical natural language processing techniques for extracting effective information from text data. Sentiment analysis involves computationally identifying and categorizing opinions in the text to determine the writer's attitude as positive, negative, or neutral. This is often performed by using lexicons or machine learning algorithms that associate words and phrases with sentiment orientation. Sentiment analysis provides crucial insights for applications such as brand monitoring, analyzing customer satisfaction, and gauging public opinion. The project utilizes the VADER sentiment analysis tool to discern the emotional tone of the generated summary.

Subjectivity detection refers to identifying text that expresses opinions versus objective facts. This can be achieved by using methods like Text Blob to calculate a subjectivity score ranging from 0 to 1, with scores below 0.5 indicating more objective language. Determining the degree of subjectivity within the text is beneficial for tasks like separating factual reporting from opinionated editorials or filtering out objective from subjective content. Overall, sentiment analysis and subjectivity detection allow for a deeper understanding of the affective content and perspectives expressed in textual data. These techniques provide actionable insights across many domains including business, politics, and social science[7].

1. **Named entity recognition (NER)**

NER is a critical natural language processing technique that automatically identifies and categorizes key nouns and proper names within unstructured text. NER typically leverages machine learning algorithms within NLP libraries like spaCy to label extracted entities as Person, Location, Organization, etc. Performing NER enables sophisticated information extraction by structuring text data around detectable entities. This provides an avenue for numerous downstream applications such as building knowledge graphs, question-answering systems, semantic search, and more. Overall, implementing NER is essential for unleashing the value of text data through enabling relation extraction, knowledge base construction, and advanced text analytics. The ability to automatically tag and classify named entities facilitates richer text understanding and allows organizations to derive enhanced business insights from unstructured data[8].

1. **Topic modeling**

Topic modeling is an unsupervised machine learning technique that reveals latent thematic structure within a collection of documents. Our project demonstrates topic modeling that is effectively implemented in Python using sklearn and NMF (non-negative matrix factorization). The steps involved are tokenizing the input text, converting it to a tf-idf weighted term-document matrix, applying NMF to factorize the matrix, and extracting the topics using the resulting term-topic and topic-document matrices.

Automatically discovering topics using algorithms like NMF provides crucial insights that would otherwise require extensive human analysis. Extracting semantic topics allows for a more nuanced understanding of document contents, trends, and relationships. This enables fruitful applications in fields like digital humanities, social science, marketing, and more. For example, topic modeling can help cluster news articles and customer feedback by subject, guide recommendations by interest, and summarize large volumes of text. Implementing topic modeling, as illustrated in the sample code, is a vital NLP technique for simplifying, navigating, and gleaning strategic knowledge from massive text corpora[3].

1. **Word frequency analysis**

Word frequency analysis provides simple yet valuable text mining insights by identifying the most common words and phrases in a document, this can be achieved by tokenizing the text, filtering out stop words and non-alphanumeric, counting word frequencies using Python's Counter, and visualizing the results as a word cloud.

Implementing basic word frequency analysis is a key NLP technique for understanding text content and highlighting top terms and themes. The project demonstrates an efficient approach by leveraging NLTK for tokenization and filtration along with data visualization libraries like Matplotlib and Word Cloud. The results reveal the core topics and concepts discussed within the text in an intuitive graphical format.

More broadly, identifying word and phrase frequencies enables the extraction of key talking points and summarized content. It complements and provides an entry point for more advanced NLP tasks like named entity recognition, topic modeling, and sentiment analysis. Word frequency analysis is especially useful for digesting long reports or articles, clustering documents by keywords, and monitoring shifts in terminology over time. This form of simple statistical NLP provides tremendous utility for content analysis and text mining with versatile applications across many domains[9].

1. **Word embeddings and word cloud**

Word embeddings represent words as high-dimensional numeric vectors that encode semantic meaning based on contextual usage. The project demonstrates an effective approach for training Word2Vec embeddings on text, reducing dimensionality with t-SNE, and visualizing the embeddings in a 2D scatter plot.

Specifically, tokenizing the text, removing stop words, training a Word2Vec model to generate word vectors, and selecting the embeddings for chosen keywords. t-SNE is then utilized to reduce the high-dimensional vectors to 2D for visualization. The resulting scatter plot provides an intuitive graphical view of how words are distributed based on semantic similarity. Words with related meanings cluster together.

Visualizing embeddings is tremendously useful for understanding model behavior and analyzing linguistic relationships. Best practices are implemented for visually exploring how words are encoded by the model, revealing nuances and connections. More broadly, creating and inspecting word embeddings is a vital NLP technique for tasks like identifying analogies, powering search systems, improving recommendations, and supporting transfer learning. The code demonstrates an effective approach to training, dimensional reduction, and visualization of word vectors for fruitful text analysis[6].

1. **Knowledge graph:**

Knowledge graphs provide an intuitive way to represent connections and relationships between entities extracted from text. spaCy's named entity recognition can identify key nouns and proper names, which become nodes in the graph. The find\_relationship function implemented in the project then detects verbs and prepositions linking entities to infer semantic relationships. These are encoded as labeled edges in a directed NetworkX graph.

Visualizing the knowledge graph with Matplotlib highlights how entities are related based on their co-occurrence and interconnections within the text. This enables a high-level overview of key entities and granular insights into specific relationships. Constructing knowledge graphs is tremendously valuable for unlocking and representing knowledge within documents. The graphs not only summarize entities and connections but also reveal deeper semantics that may not be directly stated.

More broadly, knowledge graphs are a vital AI technique for tasks like question answering, recommendation systems, searching, and identifying gaps or inconsistencies in knowledge. By encoding entities and relationships in an accessible graph format, knowledge graphs power more nuanced text understanding. The project demonstrates a practical approach to distilling unstructured text into a structured knowledge network. Automatically constructing knowledge graphs from text is crucial for connecting information and deriving actionable insights[10].

1. **Hierarchical Clustering:**

Hierarchical clustering is an unsupervised learning technique for identifying latent groups and structures within unlabeled data. The provided code demonstrates an effective implementation for text data using sci-kit-learn's Agglomerative Clustering. The keywords from the text summary are first converted into numerical features with CountVectorizer. Agglomerative clustering is then applied to group the keywords based on similarity. The resulting dendrogram visualization provides intuition about document clustering.

Hierarchical clustering is immensely valuable for digesting and navigating large collections of text. By identifying semantic clusters, related content can be discovered without needing explicit categorization or labels. The code implemented has best practices like TF-IDF vectorization and Ward's linkage method to extract meaningful hierarchies for text mining. The dendrogram intuitively displays the nested groupings, visualizing how the algorithm partitions the content[11].

More broadly, hierarchical clustering helps uncover thematic structures, relationships, and overlaps within-corpus data. The technique can power applications like document organization, recommendation systems, search, and identifying patterns. Implementing hierarchical clustering and inspecting the dendrograms enables nuanced text analysis without supervision. The project demonstrates a robust unsupervised learning approach for clustering text data and extracting insights[8].

1. **Predictive analytics based on sentiment.**

Sentiment analysis on automatically generated extractive summaries provides an effective pipeline for analyzing effects and perspectives within large text corpora. Recent research has explored predictive modeling approaches using supervised learning algorithms to classify sentiment based on summarized content. Key steps in this process include:

1) Generating a concise summary covering the most salient information. This reduces computational expenses while retaining indicative content.

2) Processing the summary sentences with spaCy for polarity scoring using VADER or other lexicons.

3) Structuring the data and training an ensemble model stacking logistic regression, random forests, SVMs, and a logistic regression meta-learner.

4) Evaluating model performance using cross-validation and metrics like accuracy, precision, recall, and F1 on held-out test data.

Predictive sentiment analysis on summaries has been shown to achieve high performance while minimizing computational costs for large corpora. This allows efficient analysis of opinions and effects at scale. Applications include tracking brand sentiment across news and social media, analyzing customer feedback, and monitoring public discourse. The code implemented in the project demonstrates the effective implementation of this predictive text analytics pipeline using standard NLP libraries and ensemble modeling. Further research can explore the impact of different summarization techniques and model architectures. Overall, this technique provides versatile capabilities for generating actionable insights from text data[12].

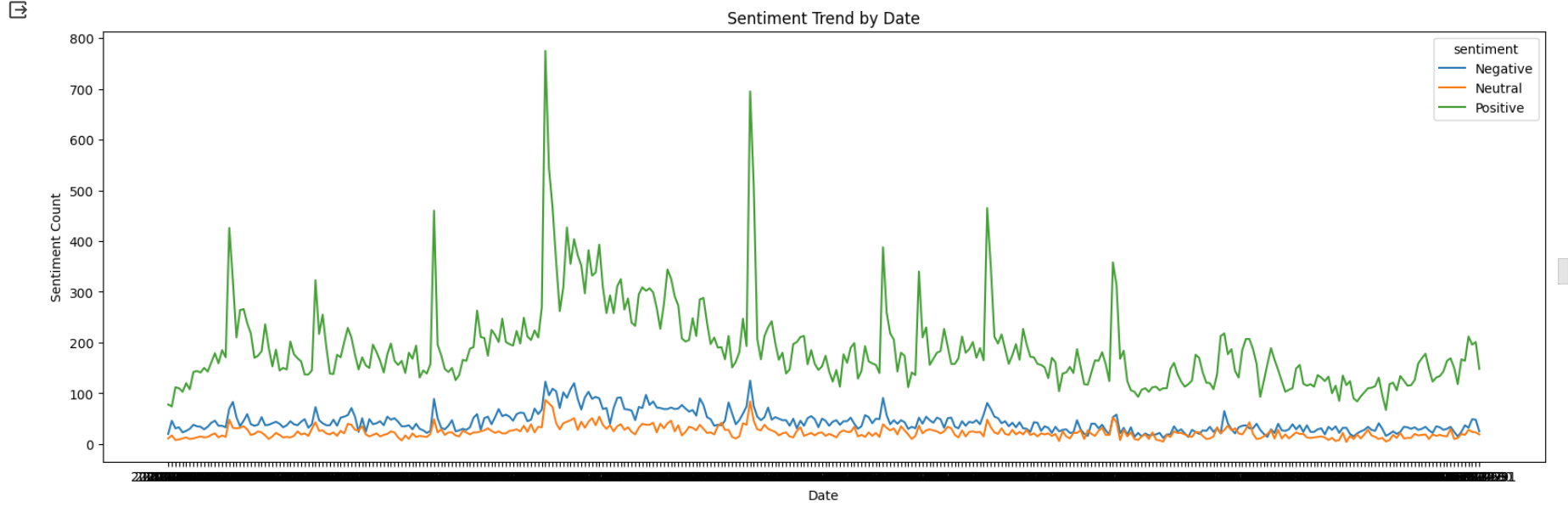
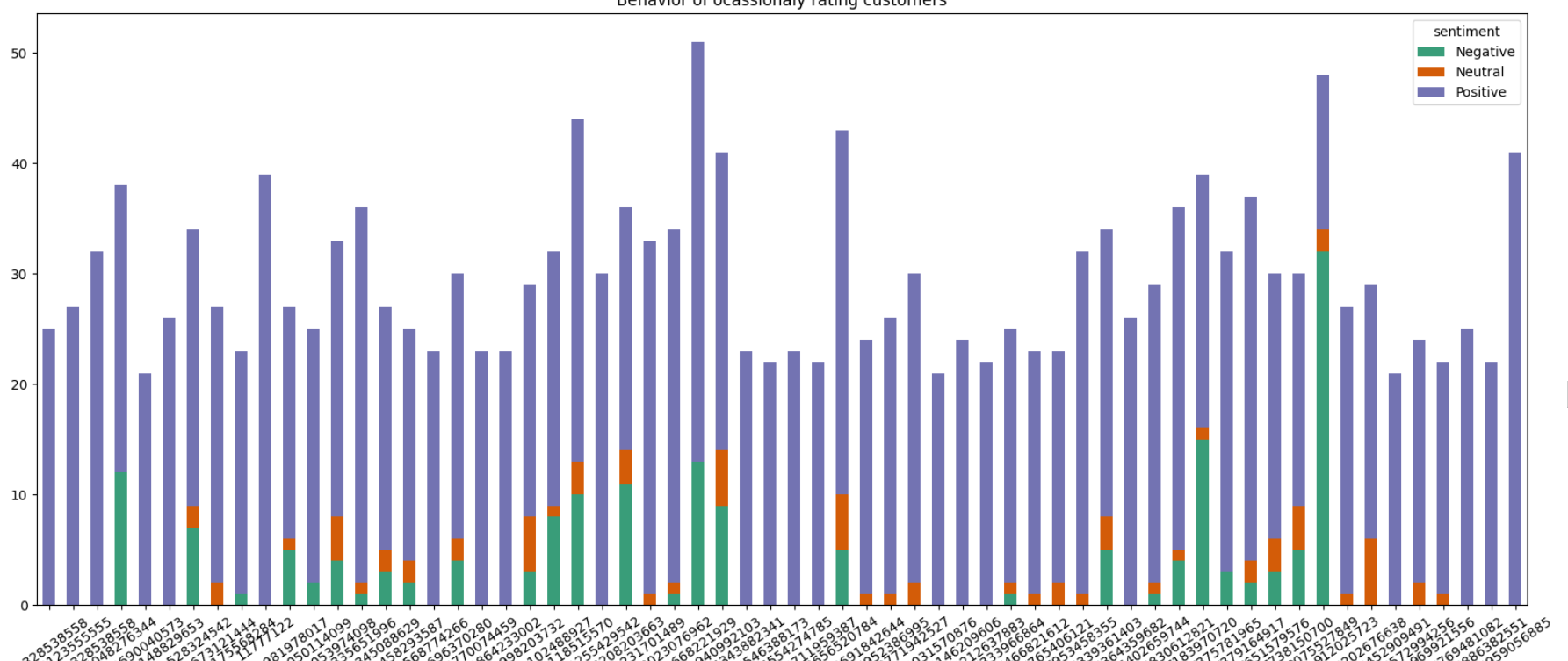
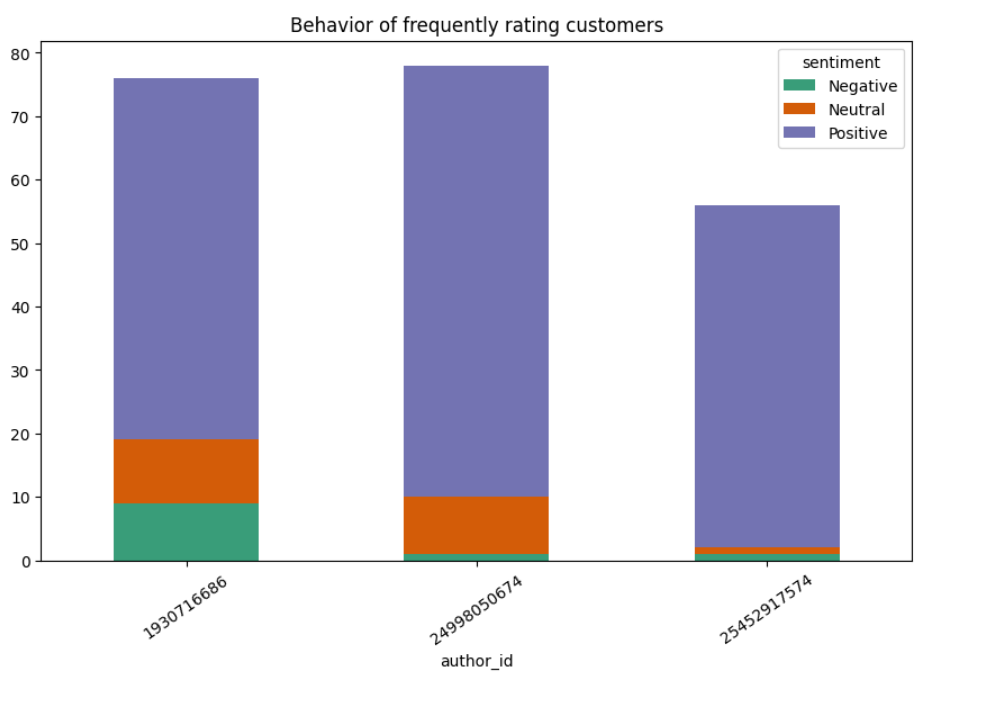
1. **Cosine similarity between summaries**

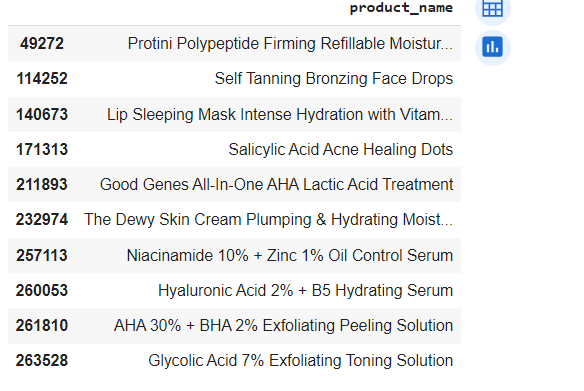
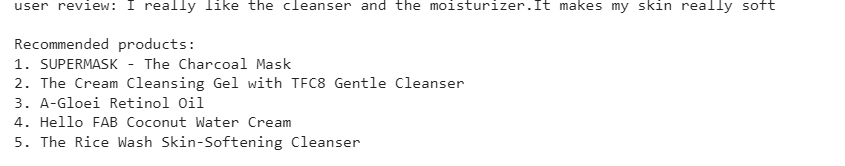
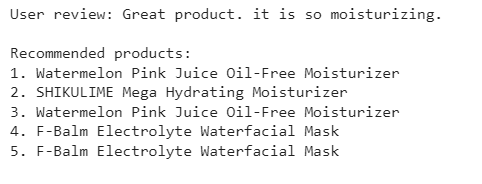
This research implements cosine similarity to compare semantic closeness between beauty product data. The review data from Farmacy, Tatcha, and Drunk Elephant moisturizer brands is preprocessed by lowercasing, removing punctuation, and stemming words to their root forms. Scikit-learn's TfidfVectorizer then converts the normalized text into TF-IDF vector representations encapsulating word frequencies. Cosine similarity is calculated between vector pairs using matrix multiplication to quantify similarity based on the angle between vectors.

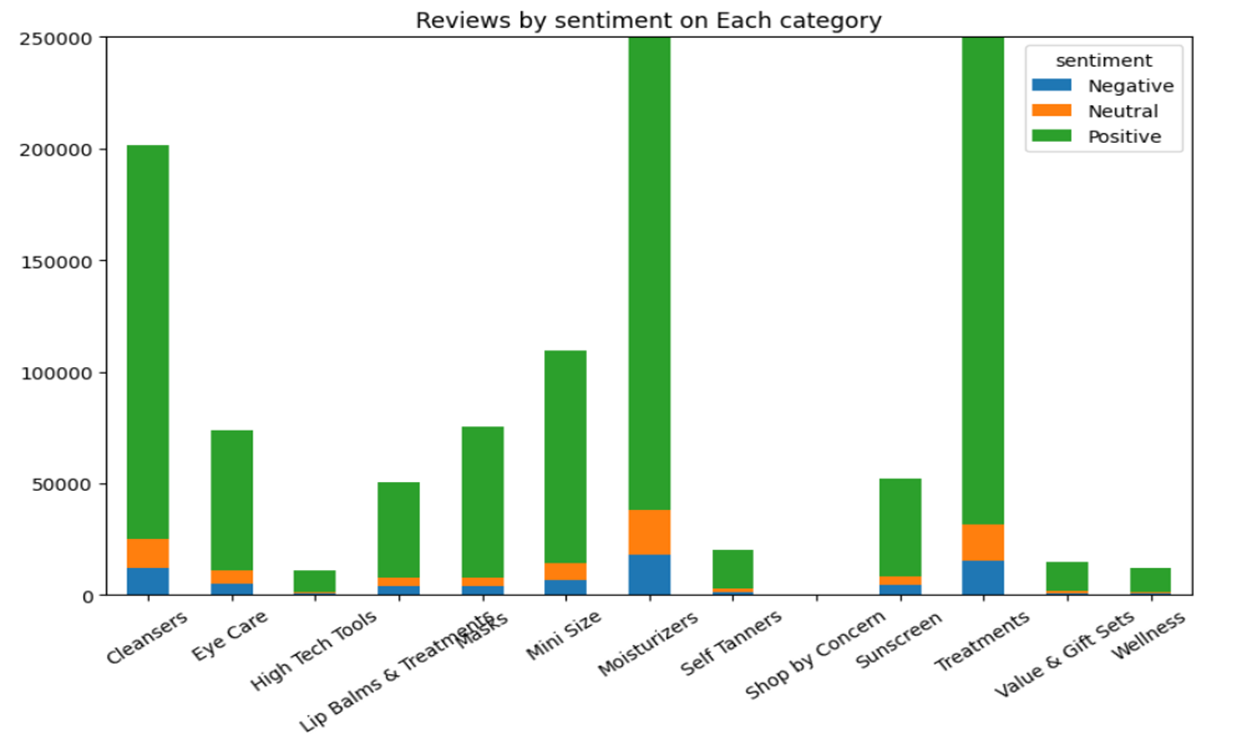
Computing similarity enables discovering relationships within the corpus to identify patterns, trends, and variations among products. The code demonstrates a robust process for textual similarity analysis using TF-IDF vectorization and cosine similarity. This unsupervised technique is foundational for numerous applications in information retrieval, natural language processing, and text analytics. Overall, cosine similarity provides an efficient method to compare documents based on semantic content rather than surface-level features. Assessing similarity is crucial for clustering documents, improving search relevancy, analyzing trends, and gaining insights from unstructured text data[10].

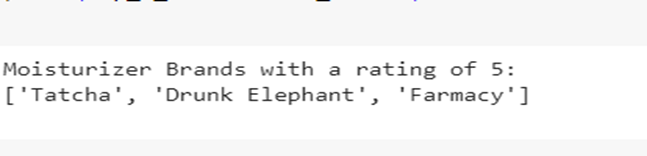
# RESULTS AND DISCUSSION

Some EDA

Recommendation system

**Product recommendation on the home page of new customers:****Product recommendation based on reviews:**

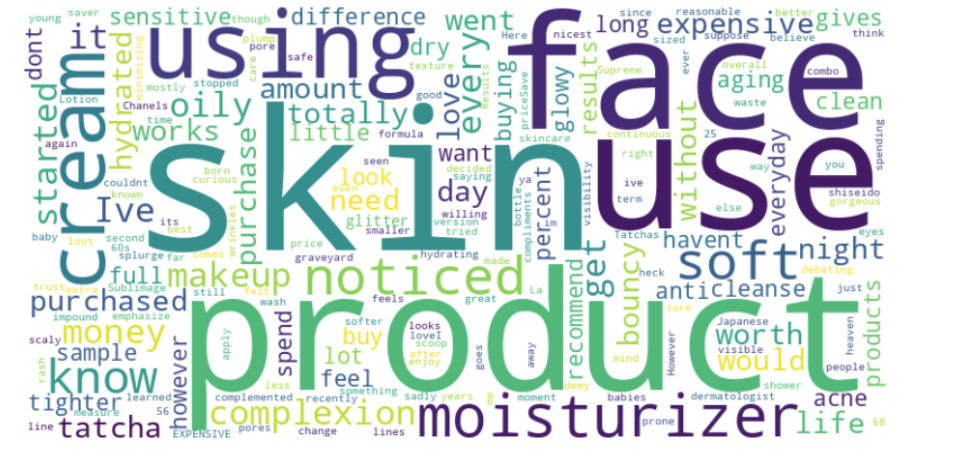
**Sentiment Analysis on top 3 moisturizers:**

****

|  | | **Tatcha** | **Farmacy** | **Drunk Elephant** | |
| --- | --- | --- | --- | --- | --- |
| **Sentiment Analysis** | | {'neg': 0.067, 'neu': 0.356, 'pos': 0.578, 'compound': 0.0258}    Subjectivity: 0.479284474206349  The text is more objective. | {'neg': 0.114,  'neu': 0.295,  'pos': 0.591, 'compound': -0.9463}    Subjectivity: 0.38706589706589706  The text is more objective. | {'neg': 0.143, 'neu': 0.31, 'pos': 0.548, 'compound': 0.9543}    Subjectivity: 0.3870967741935484  The text is more objective. | |
| **Percentages**  **Of Sentiment** | | The percentage of positive sentiment is **57.8%**  The percentage of neutral sentiment is 35.6%  The percentage of negative sentiment is 6.7% | The percentage of positive sentiment is **59.1%**  The percentage of neutral sentiment is 29.5%  The percentage of negative sentiment is 11.4% | The percentage of positive sentiment is **54.8%**  The percentage of neutral sentiment is 88.2%  The percentage of negative sentiment is 14.3% | |
| **Named Entities** | | 56 years - DATE  the day - DATE  25 percent - PERCENT  Japanese - NORP  Chanel - ORG  Tatcha - PERSON  second - ORDINAL  68 - MONEY  Tatcha - ORG | 19 - DATE  LABeautyologist - ORG  better - PERSON  Tatcha Dewy - PERSON  Curel - PERSON  about 2 months - DATE  Cali - GPE  Greens - NORP  Sephora - GPE | years - DATE  about 3 weeks - DATE  first - ORDINAL  54 - CARDINAL  moisturizer - PERSON  the day - DATE  1.69 - CARDINAL  4 months - DATE  first - ORDINAL | |
| **Topic Modeling** | | Topic 1:  buying, complexion, money, love, results, expensive, oily, started, soft  Topic 2: smaller, face, feel, like, known,recommend, does,nights  Topic 3: annoying, rash, directly, prone, face, loved, difference,sure | Topic 1: dry, weird,moisturizers, face,new, zone,acne, seeing,got,patches,everyday,moisturizer  Topic 2: try,really,tried,greasy,lot,texturized,does,results  Topic 3: mad,weird,super,lucky,like,skin,stick,fore,cali,disappear,glow,try | Topic 1: improvement,expected,floz,pores,clarity,thicker,im,decent,sink,works  Topic 2: pairing,takes,shiny,change,appearance,inside,feelingmeh,lie  Topic : cover,enjoy,skeptical,ingredients,realized,purchase,soft,spend | |
| **Word Frequency Analysis** | | skin (9)  product (6)  use (6)  face (5)  using (5)  cream (3)  moisturizer (2.45) | skin (8)  like (6)  oily (5)  just (4)  ve (3)  really (2) | moisturizer (3)  definitely (2)  sensitive (2)  cream (2)  super (2)  love (1) | |
| **Cosine Similarity** | For Tatcha and Drunk Elephant : 0.5239151678462538  For Tatcha andFarmacy:0.491527210203696  For Farmacy and Drunk Elephant:0.595662096899911 | | | | |
| **Predictive Analysis** | Accuracy: 0.88  Precision: 0.91  Recall: 0.88  F1 Score: 0.91 | | Accuracy: 1  Precision: 1  Recall: 1  F1 Score: 1 | | Accuracy: 1  Precision: 1  Recall: 1  F1 Score: 1 |

**Word Cloud**

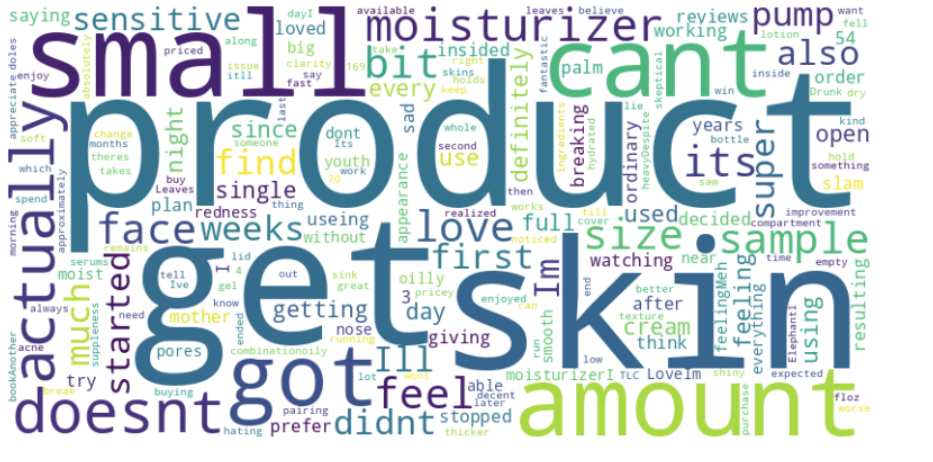
Tatcha

****

Farmacy

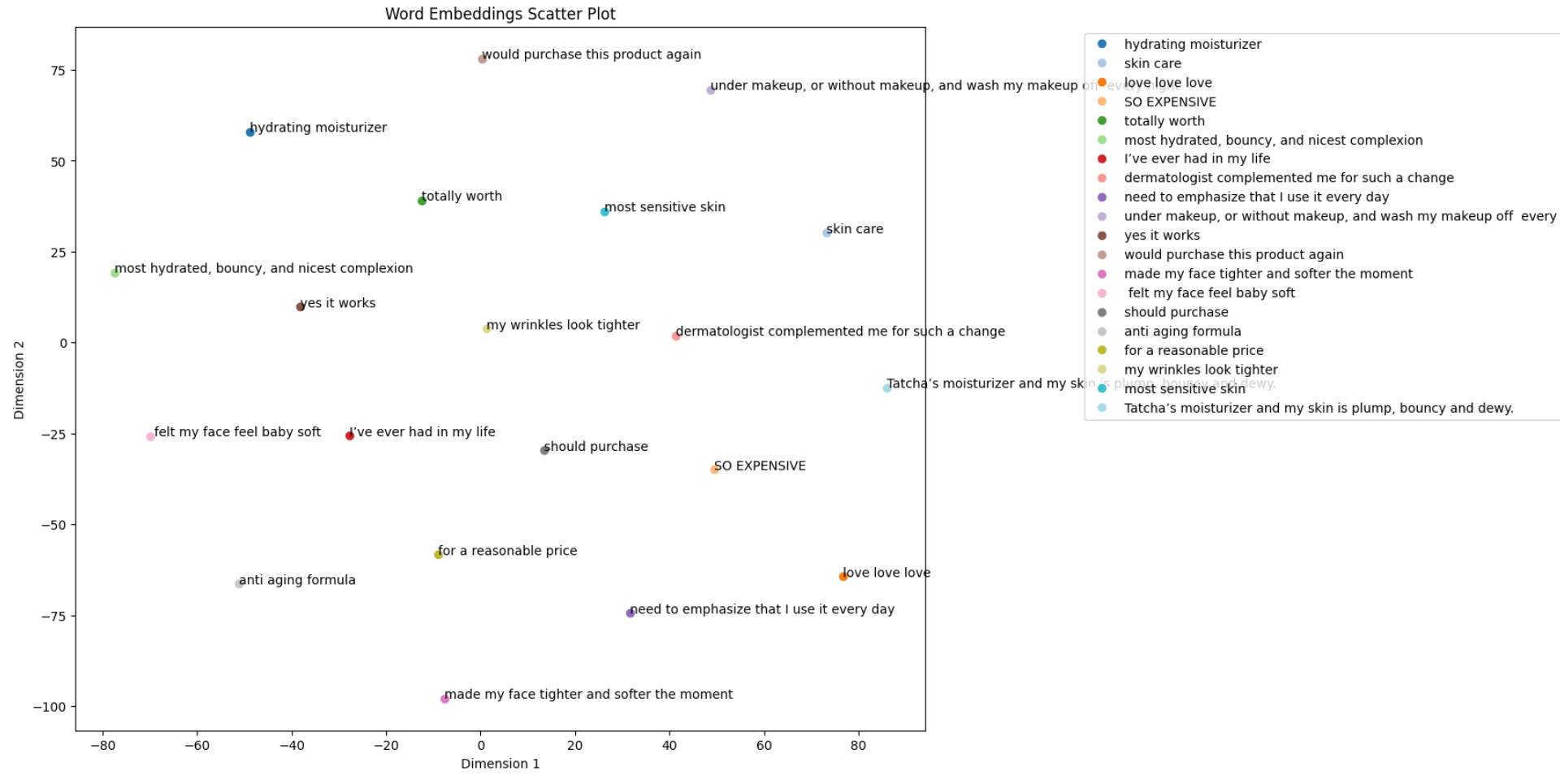


Drunk Elephant

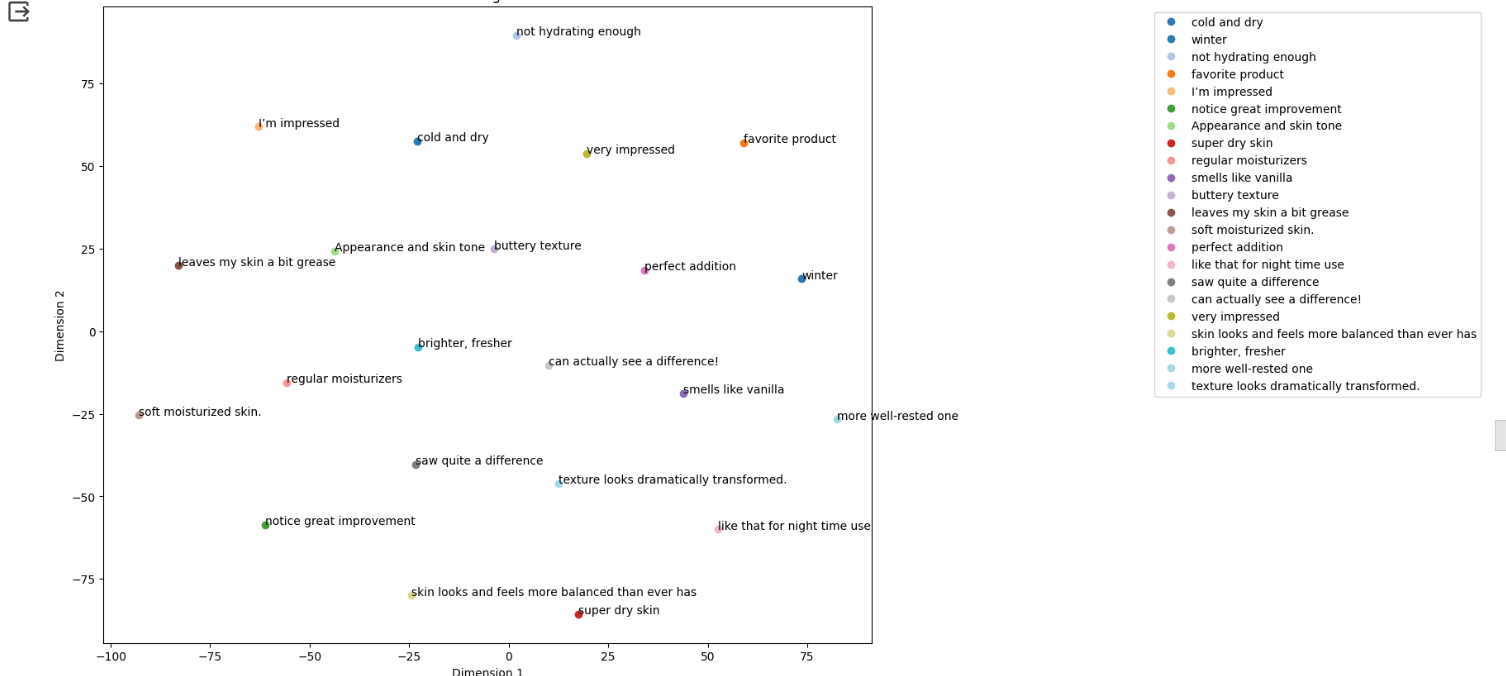


**Word Embeddings t-SNE scatterplot**

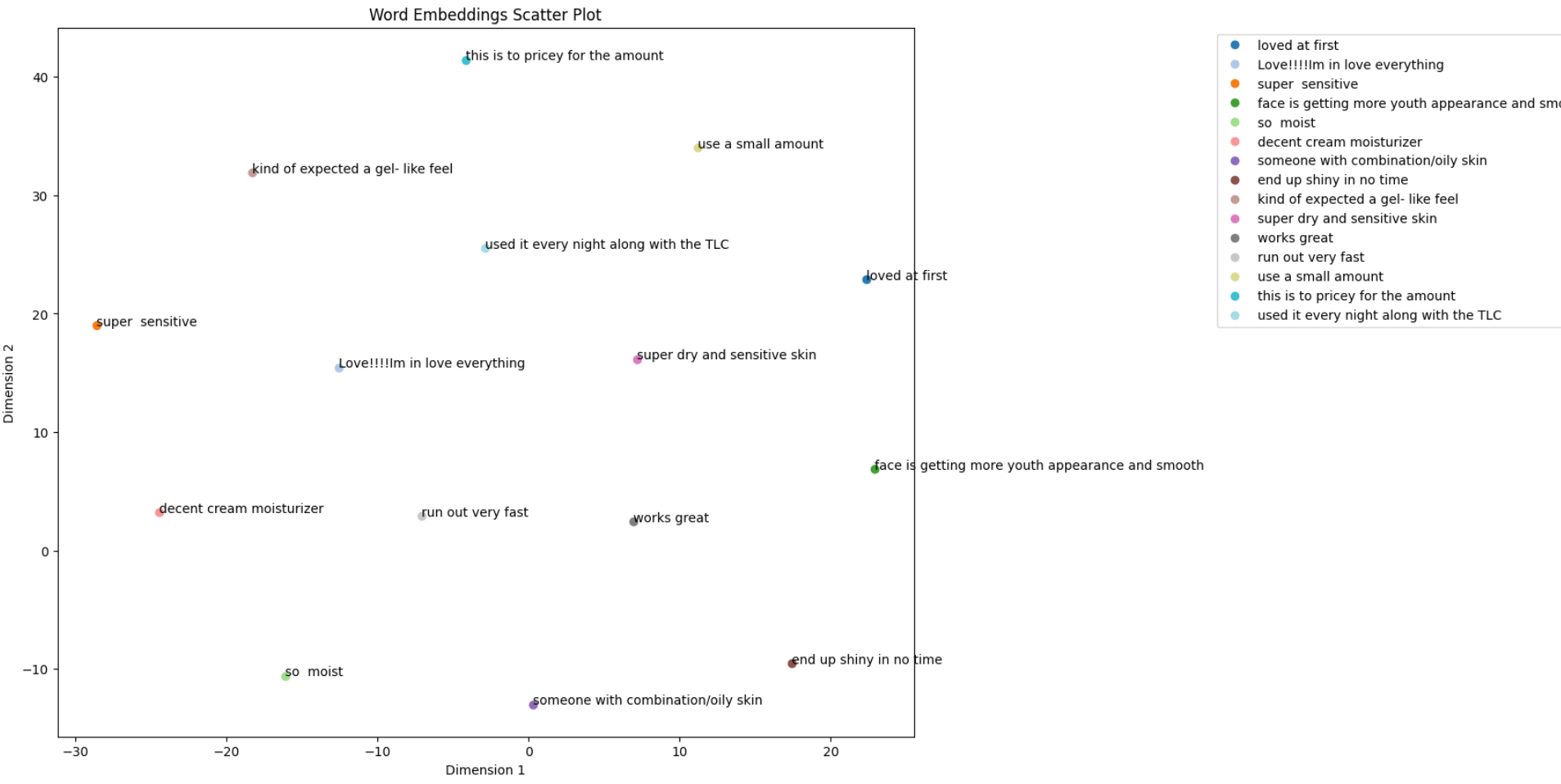
Tatcha



Farmacy

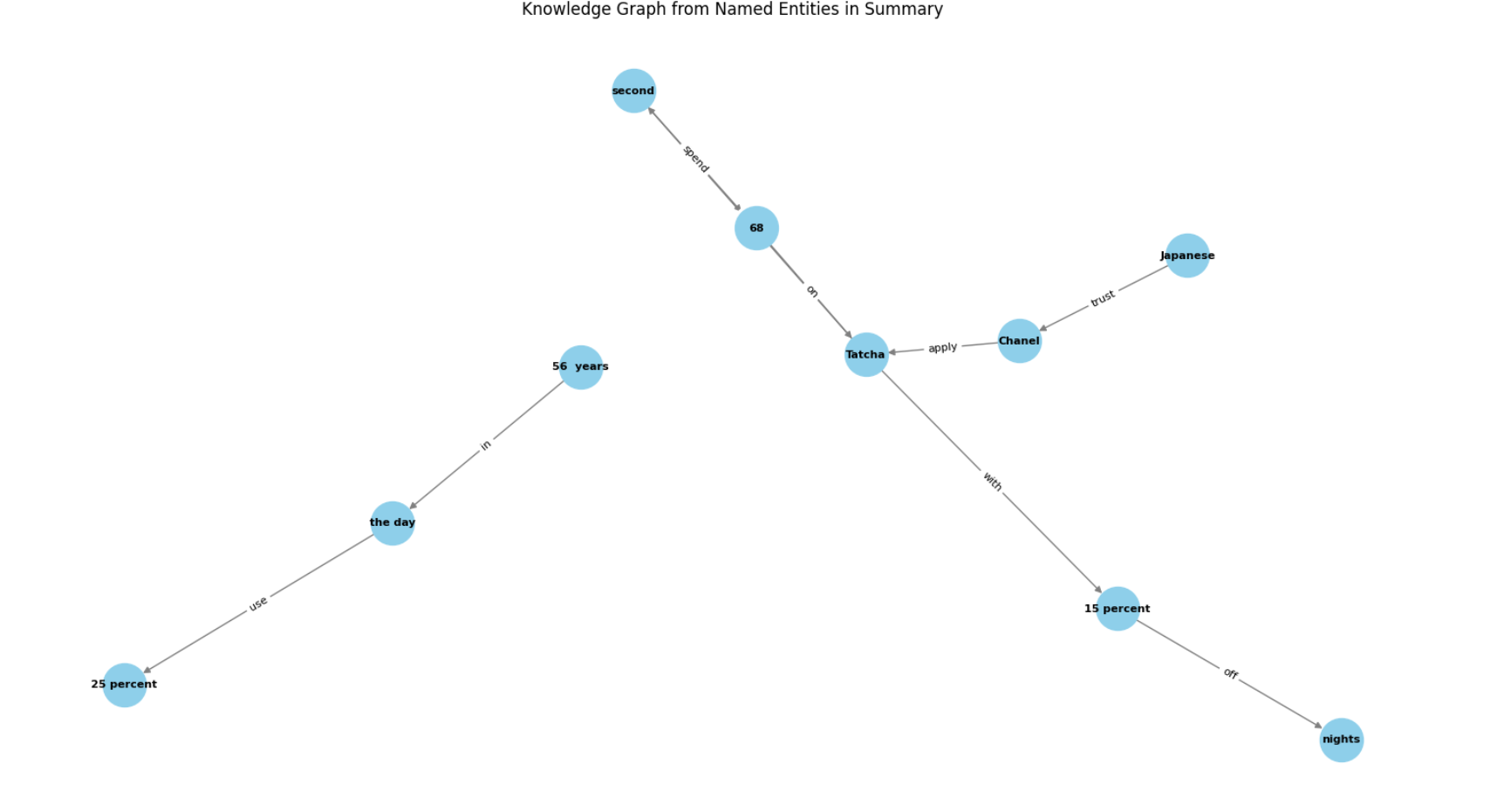


Drunk Elephant

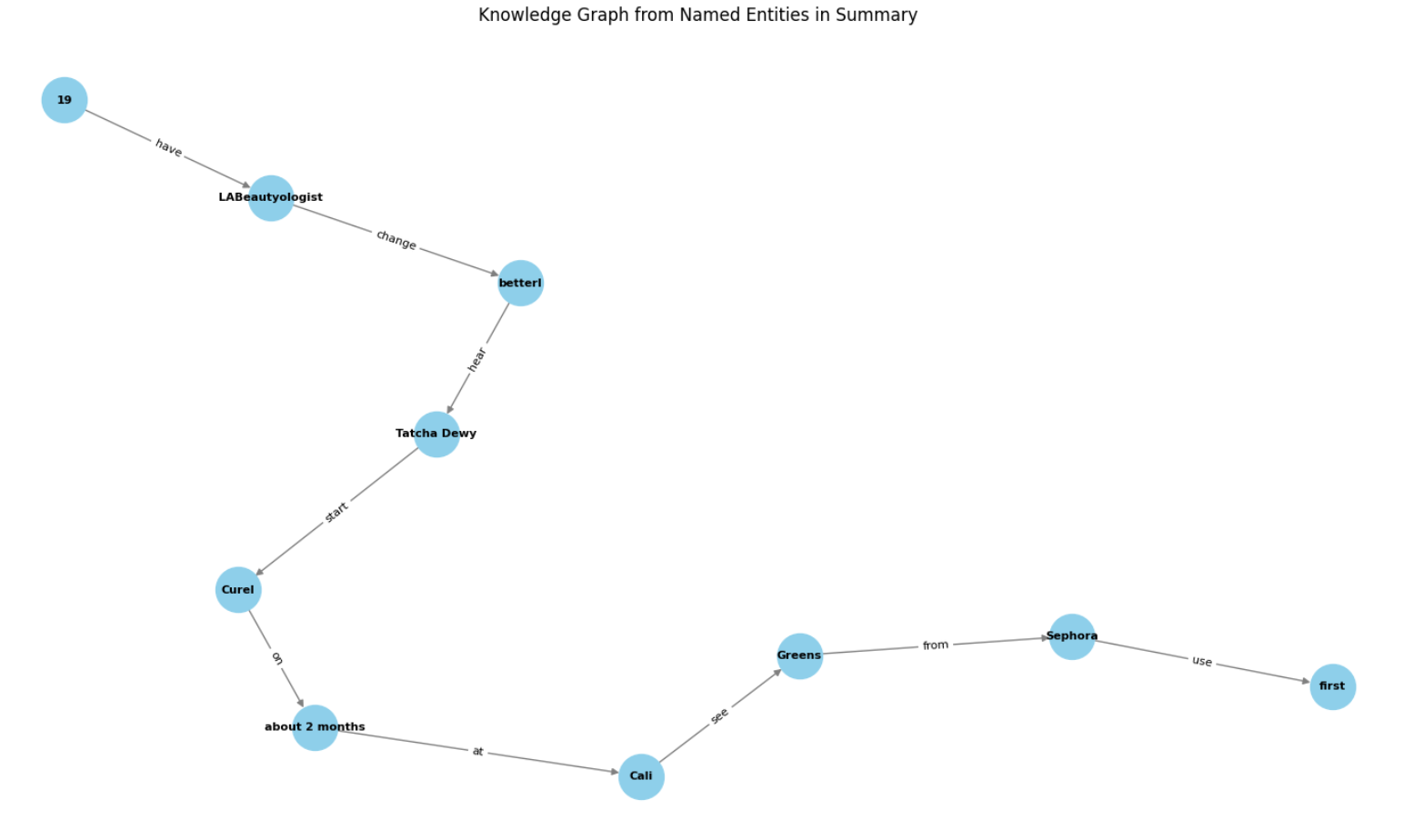


**Knowledge Graph**

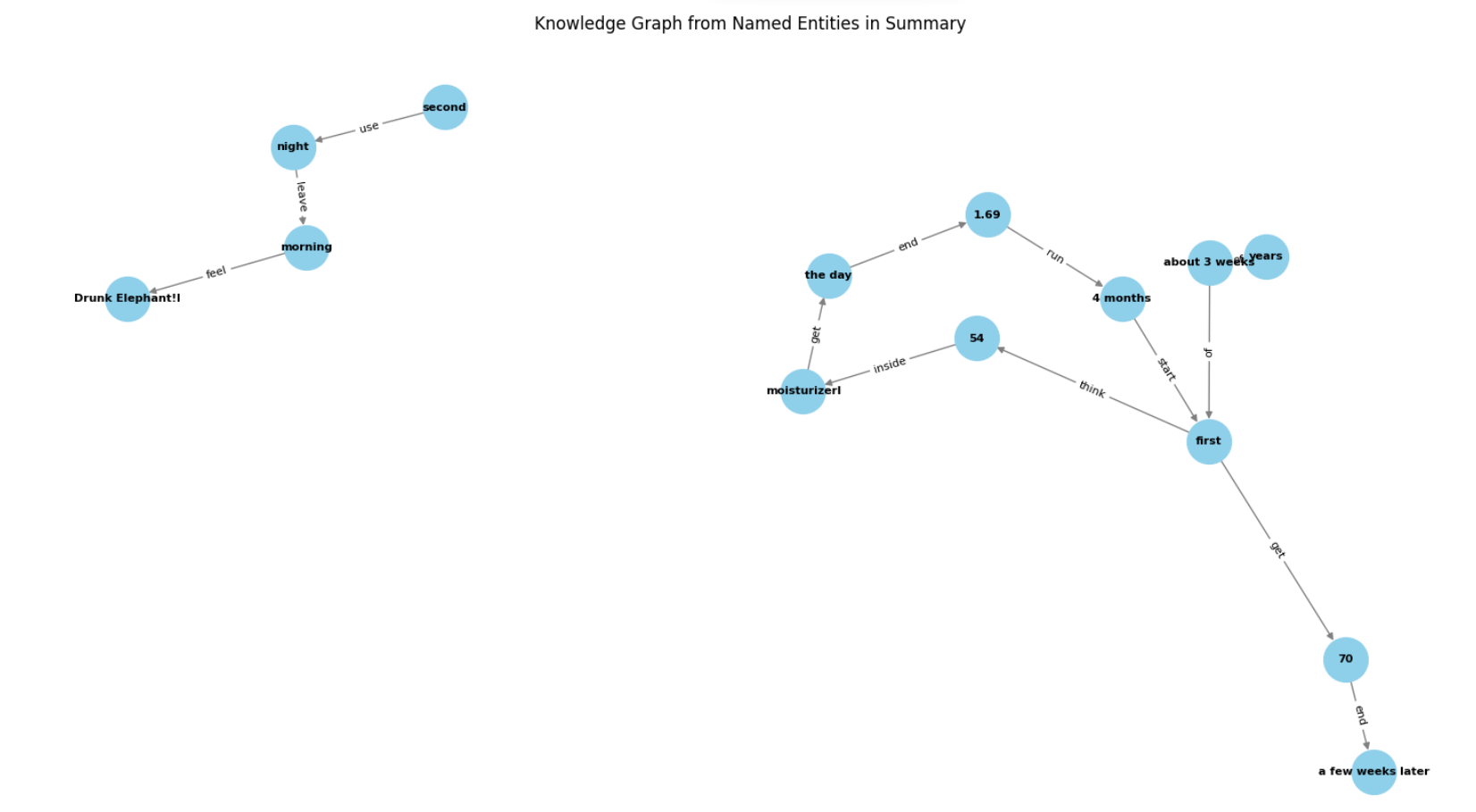
Tatcha



Farmacy

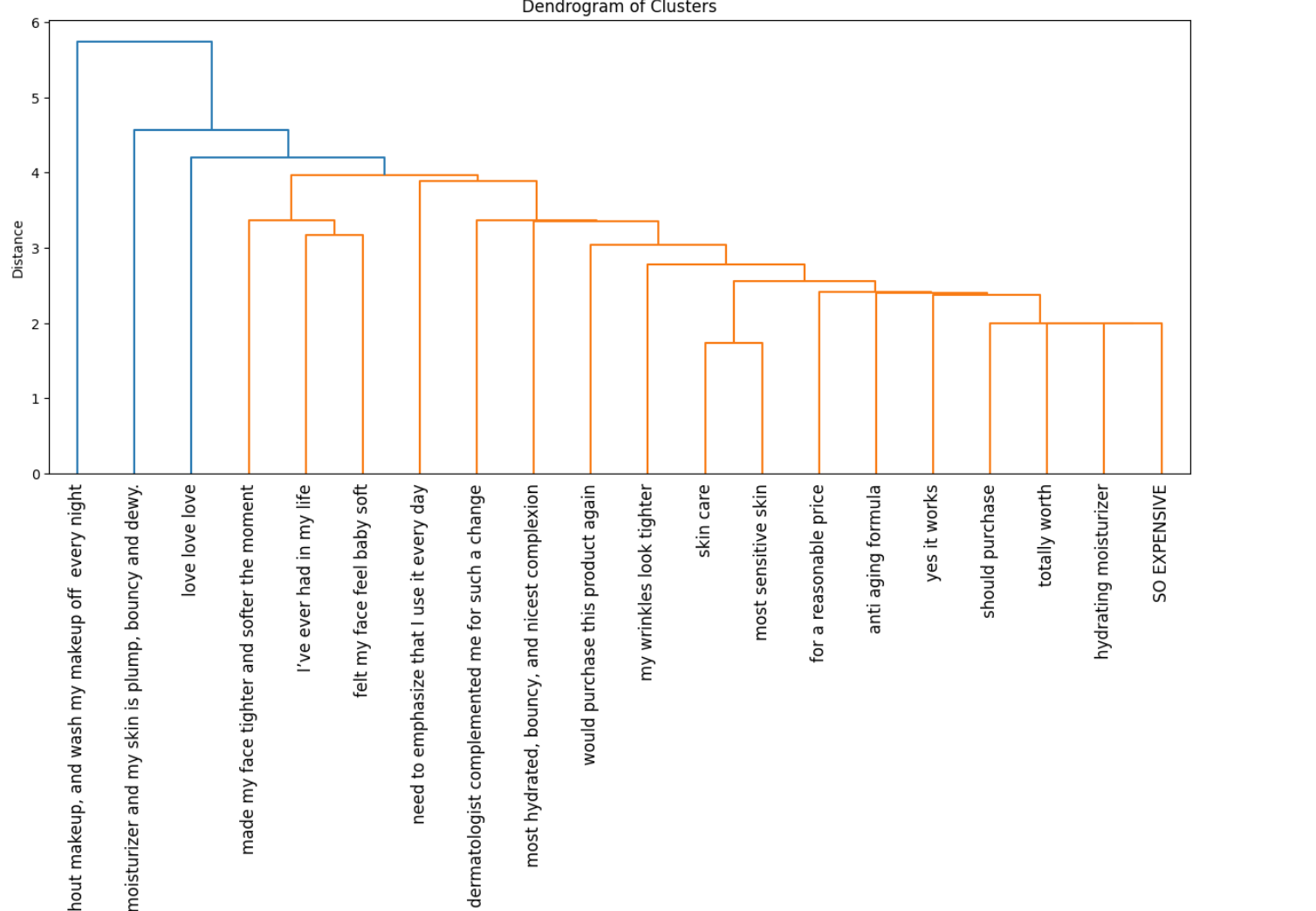


Drunk Elephant

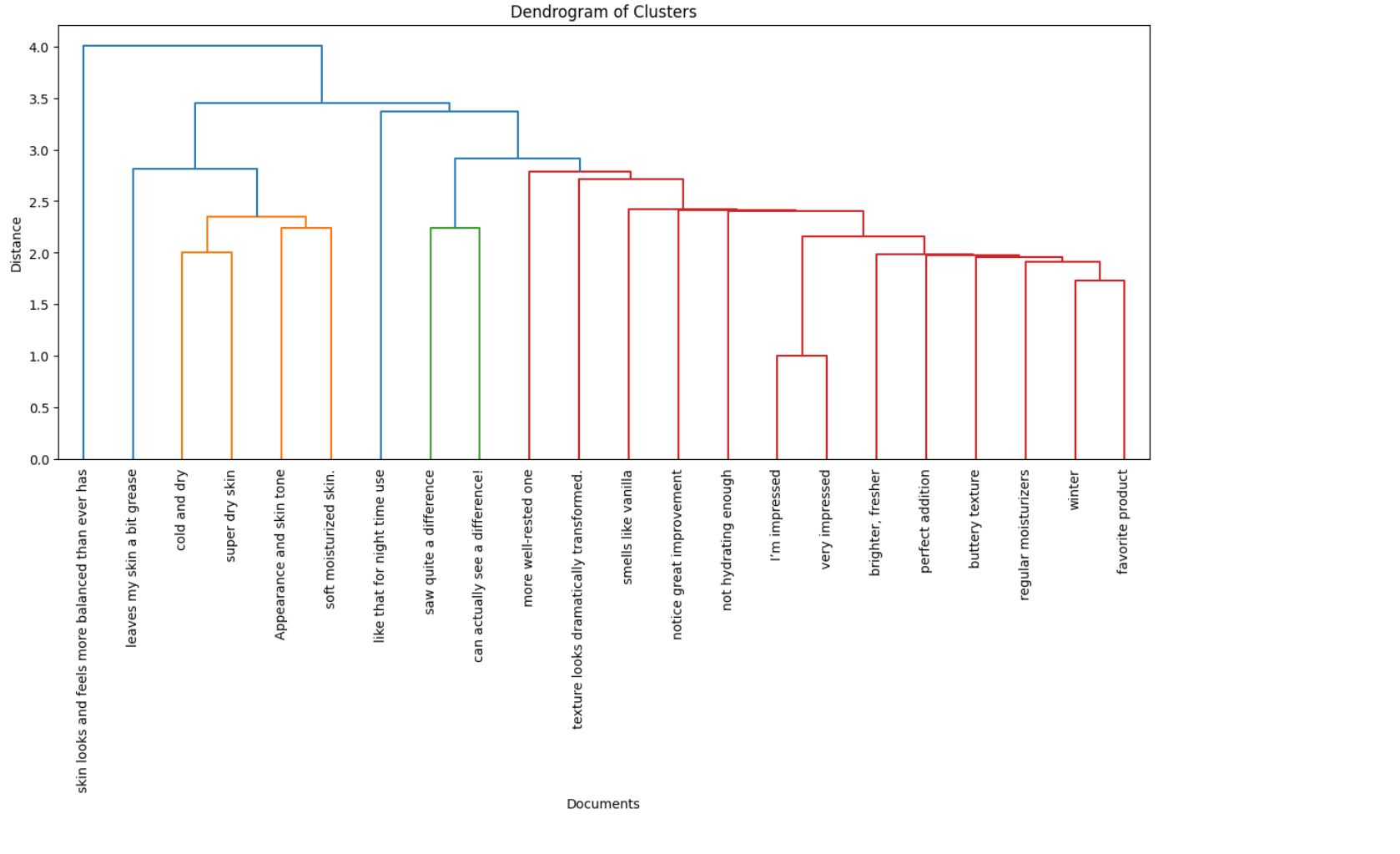


**Hierarchical Clustering**

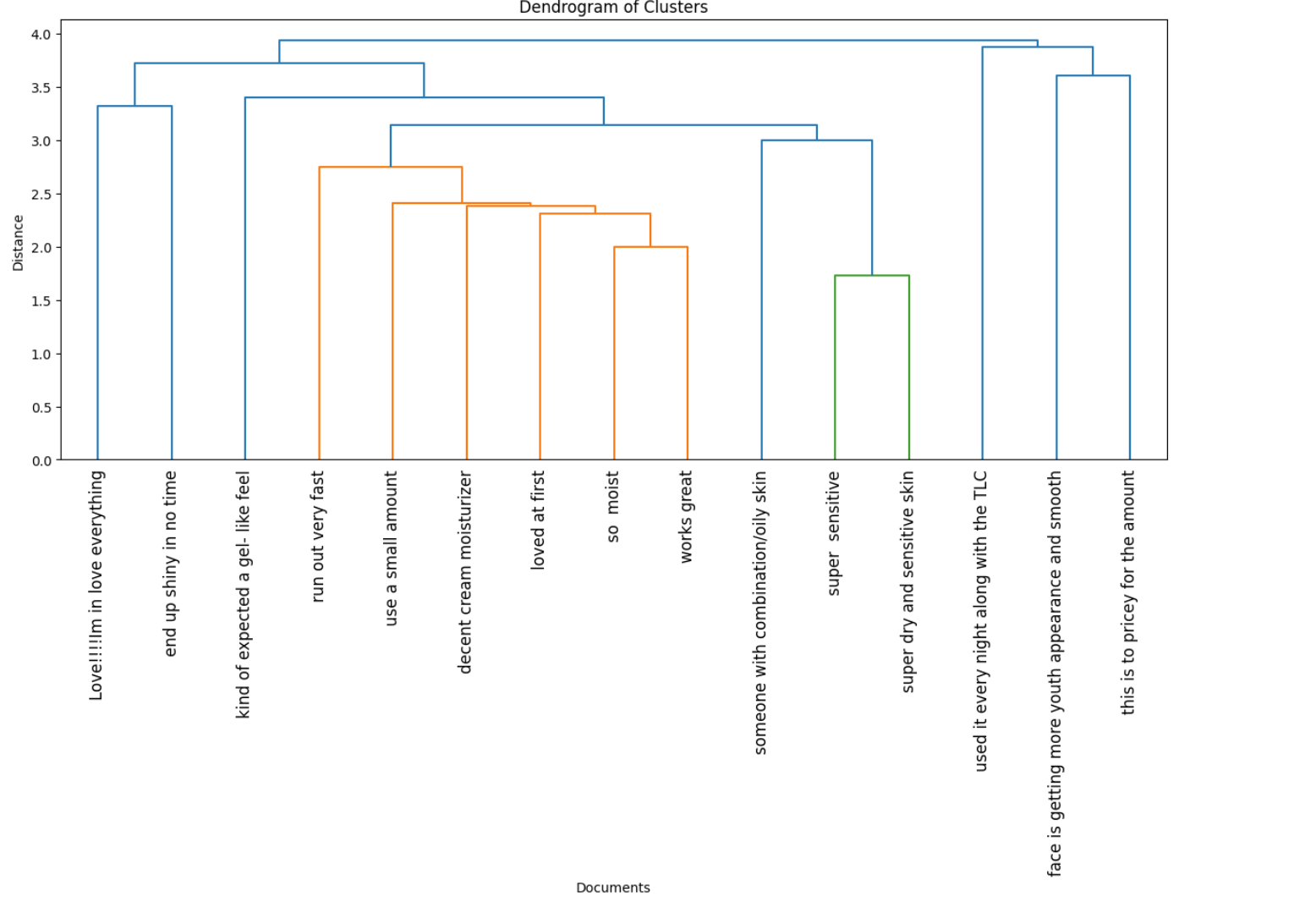
Tatcha



Farmacy



Drunk Elephant

The dendrogram shows the relationship between the number of clusters and their impact on skincare. The closer the clusters are together, the more likely it is that it unlock a deeper understanding of moisturizers and navigates the vast skincare landscape with confidence.

# CONCLUSION

Text summarization along with Sentiment classification, Topic modeling, Hierarchical clustering techniques, and predictive analytics based on sentiment are invaluable for review data analysis by distilling the key findings and conclusions from this vast corpus of product data and research. These methodologies also help make complex product information more accessible to the public through summarized explainers. In the race to adapt policies before irreversible product tipping points, text summarization, product recommendation along with Sentiment classification, Topic modeling, Hierarchical clustering, and predictive analytics based on sentiment is an indispensable tool for researchers to benefit from the latest insights emerging across diverse sources of products data and literature.

# FUTURE DIRECTIONS

This research on applying NLP techniques to a product data corpus has unveiled immense potential for further advancing discovery and comprehension. An immediate next step is expanding the dataset to include more varied sources from different brands, stores, and social media to encompass broader perspectives. More refined preprocessing like coreference resolution could enrich the analysis. Advanced deep learning approaches like BERT and GPT-3 may extract deeper semantic insights versus the implemented ML models.

Exploring diverse NLP tasks like semantic role labeling to identify product benefits, usability, and actions could reveal new angles. Finetuning language models on product reviews may improve topic extraction and sentiment prediction. Ongoing research into the product, skin, health benefits from it, ingredients used discourse trends, and framing using these techniques can support communications. Combining products, environmental, skin, ingredients used, and health data could spotlight interconnections. More robust validation frameworks would reinforce reliability. Overall, this research highlighted the tremendous potential of NLP and ML to transform product science analysis but also uncovered numerous promising directions for innovation. Advances in models, datasets, and techniques will enable even more powerful review data mining, stronger product insights, and accelerated progress on this critical global challenge.

# REFERENCES:

1. Huang, Y., Wang, N.N., Zhang, H., Wang, J.: A novel product recommendation model consolidating price, trust and online reviews. Kybernetes 48 (6), 1355–1372 (2019). <https://doi.org/10>. 1108/K-03-2018-0143

2. Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. Computer, 42(8), 30-37.

3. Wang, J., Zhang, Y., Li, J., & Wang, J. (2023). An Intelligent Hybrid Neural Collaborative Filtering Approach for True Recommendations. [Online] Available at: https://www.researchgate.net/publication/371891225\_An\_Intelligent\_Hybrid\_Neural\_Collaborative\_Filtering\_Approach\_for\_True\_Recommendations.

4. Ricci, F., Rokach, L., & Shapira, B. (2015). Recommender Systems Handbook. Springer.

5. McNee, S. M., Riedl, J., & Konstan, J. A. (2006). Being accurate is not enough: How accuracy metrics have hurt recommender systems. In CHI'06 Extended Abstracts on Human Factors in Computing Systems (pp. 1097-1101).

6. Zhan, D., Cui, L., & Xu, H. (2016). A survey of collaborative filtering-based social recommender systems. Computer Communications, 89, 1-10.

7. Sparks, E. R., & Talwalkar, A. (2013). HOGWILD!: A lock-free approach to parallelizing stochastic gradient descent. In Proceedings of the 30th International Conference on International Conference on Machine Learning (Vol. 28, No. 3).

8. Nguyen, Q. V. H., Le, T. T., & Nguyen, L. M. (2016). Sentiment analysis of Vietnamese text using machine learning approach. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 1115-1123).

9. Winastwan, R. (2020, October 2). Visualizing word embedding with PCA and T-Sne. Medium. <https://towardsdatascience.com/visualizing-word-embedding-with-pca-and-t-sne-961a692509f5>

10.Brownlee, J. (2019, August 7). A gentle introduction to text summarization. MachineLearningMastery.com. <https://machinelearningmastery.com/gentle-introduction-text-summarization/>

11.Lavanya, K. C., Sivamani, C., Tomy, L., & Paul, A. R. (January 2019). Two-level text summarization with sentiment analysis for multi-document summarization.

12.Selvaraj, G., & Selvaraj, G. (May 2023). Improved driven text summarization using PageRanking algorithm and cosine similarity.