STAT 432

Drug Recommendation via Efficacy and Adverse Event Modeling

Aniket Tathe (atathe2)
Amitabh Swain (amitabh5)

Problem Introduction :

Personalized medicine, a rapidly advancing field, aims to tailor medical treatment to individual patients based on their unique characteristics, such as genetic profiles, medical history, or personal preferences. A critical aspect of this field is recommending the most suitable drugs for patients. This process involves analyzing extensive and complex data to understand patient-specific factors that influence drug efficacy and safety.

Traditionally, supervised learning approaches have been employed for such tasks, requiring labeled data to train predictive models. However, in the context of personalized drug recommendations, supervised methods can be inefficient and challenging due to the need for large-scale labeled datasets and the inherent diversity in patient profiles.

To address these challenges, we propose an unsupervised approach leveraging topic modeling techniques such as **BERTopic** and **Latent Dirichlet Allocation (LDA)**. Using the publicly available **Drug Reviews dataset** (Druglib.com) from the **UCI Machine Learning Repository**, we developed a methodology for recommending the top 10 drugs tailored to patients' preferences or medical history. This dataset contains patient reviews on various drugs, including their benefits, side effects, and overall comments, along with ratings for effectiveness and satisfaction. By analyzing this rich textual data using unsupervised techniques, we aim to uncover latent patterns that align with patient preferences without relying on predefined labels.

Our approach focuses on extracting meaningful insights from patient reviews to identify patterns in drug performance and satisfaction. These insights are then used to recommend drugs that best match individual patient profiles. This methodology not only enhances the efficiency of personalized drug recommendations but also demonstrates the potential of unsupervised learning in addressing complex healthcare problems.

Basic terminology

The *Drug Reviews (Druglib.com)* dataset is a comprehensive collection of patient reviews on various drugs, sourced from online pharmaceutical review sites. Released on October 1, 2018, this dataset comprises *4,143 instances* and *8 features*, enabling the analysis of patient experiences regarding drug effectiveness and side effects. The dataset is structured to facilitate various tasks, including classification, regression, and clustering, making it a valuable resource for research in personalized medicine.

Dataset Features

reviewID : A unique identifier for each review. This is an integer type variable and does not contain missing values.

urlDrugName : This categorical variable represents the name of the drug being reviewed. It allows for the identification of specific medications within the dataset.

Rating : An integer variable that captures the overall satisfaction rating given by patients on a scale of 1 to 10. Higher values indicate greater satisfaction with the drug.

Effectiveness : A categorical variable that categorizes the perceived effectiveness of the drug on a 5-step scale. This variable helps in understanding how patients rate the drug's efficacy.

sideEffects : Similar to effectiveness, this categorical variable classifies the side effects experienced by patients into a 5-step rating system. It provides insights into the tolerability of the drug.

Condition : A categorical feature indicating the medical condition for which the drug was prescribed. This variable is crucial for correlating drug efficacy with specific health issues.

benefitsReview : A text-based feature where patients describe the benefits they experienced while using the drug. This qualitative data can be analyzed to extract common themes and sentiments.

sideEffectsReview : Another text feature where patients detail their experiences with side effects. Analyzing this data can help identify prevalent adverse reactions associated with specific medications.

commentsReview : This text variable contains overall comments from patients about their experience with the drug. It serves as a rich source of qualitative feedback that can be used for sentiment analysis and topic modeling.

Data Characteristics

The dataset contains well-defined features, with no missing values. Its combination of numerical ratings and qualitative reviews allows for a multifaceted analysis of patient experiences, making it suitable for an unsupervised approach to personalized drug recommendations based on patient preferences and medical history.

By leveraging topic modeling techniques such as BERTopic and LDA, we aim to extract meaningful patterns from this dataset to inform recommendations tailored to individual patient profiles, ultimately enhancing personalized medicine initiatives.

Literature Review

Topic Modeling in NLP

Topic modeling has emerged as a powerful tool in natural language processing (NLP) for extracting latent themes from textual data. Traditional methods like Latent Dirichlet Allocation (LDA) are widely used due to their simplicity and interpretability. However, recent advancements in transformer-based models, such as BERTopic, have significantly improved the quality of topic modeling by leveraging contextual embeddings

Comparative Studies on Topic Modeling

1. **Performance Comparison of LDA and BERTopic**
   Egger and Yu (2022) [1] compared LDA, BERTopic, Top2Vec, and NMF on datasets like Twitter data and customer feedback. BERTopic consistently outperformed LDA in terms of semantic coherence and relevance. While LDA effectively grouped documents into distinct clusters, BERTopic's use of transformer-based embeddings (e.g., MPNet and MiniLM) enabled it to capture richer semantic relationships between topics.
   - **Key Insight**: BERTopic demonstrated higher coherence scores, making it more suitable for complex datasets with nuanced text.
2. **Applications in Customer Feedback Analysis**
   Research comparing topic modeling methods on customer call data revealed that BERTopic excelled in quantitative metrics like coherence scores, while LDA sometimes aligned better with human interpretation for simpler datasets (**Umu DiVA Portal [2]**). This suggests that dataset characteristics play a crucial role in determining the optimal topic modeling approach.

Citations:
[1] : https://www.frontiersin.org/journals/sociology/articles/10.3389/fsoc.2022.886498/full
[2] : https://umu.diva-portal.org/smash/get/diva2:1763637/FULLTEXT01.pdf

Applications in Healthcare and Drug Recommendation Systems

1. **Unsupervised Learning for Drug Design**
   Recent advancements in unsupervised learning have shown significant potential for drug design and recommendation systems. A review highlighted the efficiency of unsupervised architectures in mapping molecular representations, indicating their applicability in drug discovery (**Egger & Yu, 2022 [1]**).
2. **Machine Learning for Drug Reviews**
   A study leveraging patient reviews for drug recommendation highlighted the potential of machine learning techniques to predict the best medications based on user satisfaction and side effects (**Bajorath et al., 2021 [2]**). These findings align with our methodology of using unsupervised topic modeling to extract insights from patient feedback.
3. **Predicting Drug-Related Side Effects**
   Ensemble machine learning models integrating chemical, biological, and phenotypic features have been used to predict drug-related side effects effectively (**Nature [3]**). These studies emphasize the importance of combining diverse data sources for robust predictions, which complements our approach to personalized drug recommendations.

Citations:
[1] : https://www.frontiersin.org/journals/sociology/articles/10.3389/fsoc.2022.886498/full
[2] : https://pubmed.ncbi.nlm.nih.gov/34393317/
[3] : https://www.nature.com/articles/s41573-019-0024-5

Performance Metrics and Visualization

Egger and Yu (2022) [1] provide a detailed comparison of topic modeling approaches, highlighting the performance differences between LDA and BERTopic. Their study presents results in tabular format, showing coherence scores, diversity metrics, and perplexity values for each model across various datasets.

	20 NewsGroups		BBC News		Trump	
	TC	TD	TC	TD	TC	TD
LDA	.058	.749	.014	.577	-.011	.502
NMF	.089	.663	.012	.549	.009	.379
T2V-*MPNET*	.068	.718	-.027	.540	-.213	.698
T2V-*Doc2Vec*	.192	.823	.171	.792	-.169	.658
CTM	.096	.886	.094	.819	.009	.855
BERTopic-*MPNET*	.166	.851	.167	.794	.066	.663

BERTopic [2]

Traditional Approach with LDA

Latent Dirichlet Allocation (LDA) has long been a foundational technique in topic modeling due to its ability to discover latent topics in document collections through unsupervised learning. Its key strengths include:

- Modeling each document as a mixture of latent topics.
- High computational efficiency and relatively fast processing times.

However, LDA has significant limitations:

- The **bag-of-words approach** disregards semantic relationships between words.
- It fails to account for **word context** within sentences.
- It does not leverage modern **contextual representations** of language.

Transition to BERTopic

BERTopic emerged as a modern alternative to address the shortcomings of LDA. Key improvements include:

- **Utilizing pre-trained transformer-based language models** for document embeddings.
- Applying a **class-based TF-IDF procedure** for topic representation.
- Decoupling **document clustering** from **topic representation generation**, enabling greater flexibility.

Modern Implementations

	20 NewsGroups		BBC News		Trump	
	TC	TD	TC	TD	TC	TD
BERTopic-*USE*	.149	.858	.158	.764	.051	.684
BERTopic-*Doc2Vec*	.173	.871	.168	.819	-.088	.536
BERTopic-*MiniLM*	.159	.833	.170	.802	.060	.660
BERTopic-*MPNET*	.166	.851	.167	.792	.066	.663

BERTopic [2]

The advancements in BERTopic leverage state-of-the-art language models such as MiniLM and MPNET, which provide distinct advantages:

Model	Topic Coherence	Topic Diversity	Performance Notes
BERTopic-MiniLM	0.159	0.833	Efficient with limited GPU capacity
BERTopic-MPNET	0.166	0.851	Highest overall performance

Key Insights:

- **BERTopic-MiniLM** provides an excellent trade-off between speed and performance, making it suitable for resource-constrained setups.
- **BERTopic-MPNET** achieves superior topic coherence and diversity, making it the preferred option when computational resources are available.

Both implementations significantly outperform traditional LDA, which achieved a coherence score (TC) of 0.058 and a diversity score (TD) of 0.749. BERTopic also demonstrates consistent performance across various datasets, leveraging contextual embeddings to enhance topic modeling applications.

Citations:

[1] : https://www.frontiersin.org/journals/sociology/articles/10.3389/fsoc.2022.886498/full

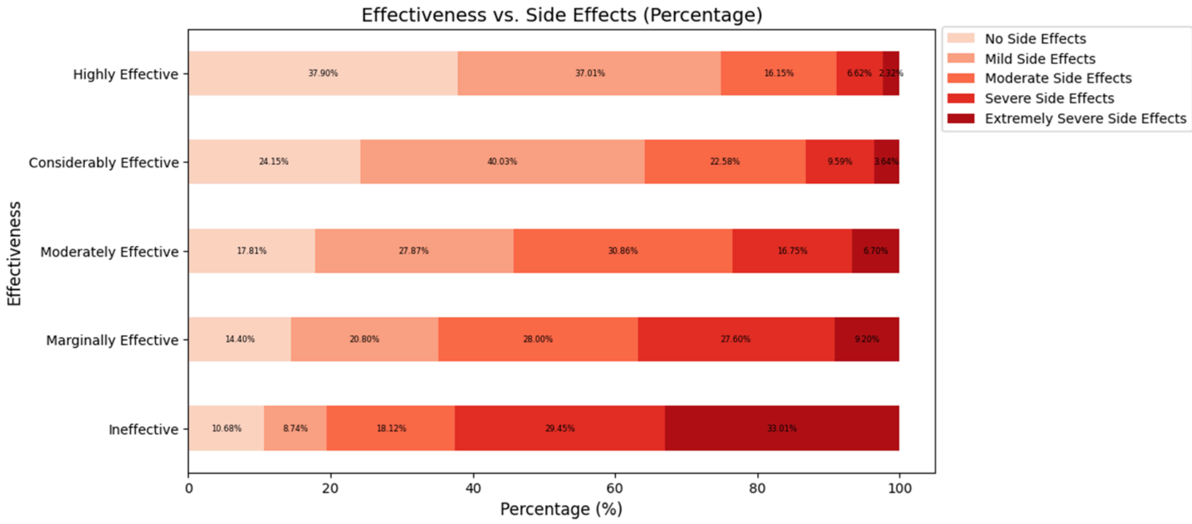
[2] : https://arxiv.org/pdf/2203.05794

Exploratory Data Analysis

This is how the dataset looks :

	Unnamed: 0	urlDrugName	rating	effectiveness	sideEffects	condition	benefitsReview	sideEffectsReview	commentsReview
0	2202	enalapril	4	Highly Effective	Mild Side Effects	management of congestive heart failure	slowed the progression of left ventricular dys...	cough, hypotension , proteinuria, impotence , ...	monitor blood pressure , weight and asses for ...
1	3117	ortho-tri-cyclen	1	Highly Effective	Severe Side Effects	birth prevention	Although this type of birth control has more c...	Heavy Cycle, Cramps, Hot Flashes, Fatigue, Lon...	I Hate This Birth Control, I Would Not Suggest...
2	1146	ponstel	10	Highly Effective	No Side Effects	menstrual cramps	I was used to having cramps so badly that they...	Heavier bleeding and clotting than normal.	I took 2 pills at the onset of my menstrual cr...
3	3947	prilosec	3	Marginally Effective	Mild Side Effects	acid reflux	The acid reflux went away for a few months aft...	Constipation, dry mouth and some mild dizzines...	I was given Prilosec prescription at a dose of...
4	1951	lyrica	2	Marginally Effective	Severe Side Effects	fibromyalgia	I think that the Lyrica was starting to help w...	I felt extremely drugged and dopey. Could not...	See above
...	...	...	...	...	...	...	...	...	...
3102	1039	vyvanse	10	Highly Effective	Mild Side Effects	adhd	Increased focus, attention, productivity. Bett...	Restless legs at night, insomnia, headache (so...	I took adderall once as a child, and it made m...
3103	3281	zoloft	1	Ineffective	Extremely Severe Side Effects	depression	Emotions were somewhat blunted. Less moodiness.	Weight gain, extreme tiredness during the day,...	I was on Zoloft for about 2 years total. I am ...
3104	1664	climara	2	Marginally Effective	Moderate Side Effects	total hysterctomy	---	Constant issues with the patch not staying on....	---
3105	2621	trileptal	8	Considerably Effective	Mild Side Effects	epilepsy	Controlled complex partial seizures.	Dizziness, fatigue, nausea	Started at 2 doses of 300 mg a day and worked ...
3106	2748	micardis	4	Moderately Effective	Moderate Side Effects	high blood pressure	The drug Micardis did seem to alleviate my hig...	I find when I am taking Micardis that I tend t...	I take Micardis in pill form once daily.

Distribution of Drug Effectiveness and Side Effects :



The stacked bar chart reveals important patterns in the relationship between drug effectiveness and associated side effects across the dataset. Here are the key observations:

Highly Effective Drugs

- 37.90% of highly effective drugs reported no side effects
- The majority experienced mild to moderate side effects (approximately 75%)
- Only 4.21% reported extremely severe side effects

Considerably Effective Drugs

- Show a similar pattern with 34.15% reporting no side effects
- Higher proportion of moderate side effects compared to highly effective drugs
- Slightly higher incidence of severe side effects at 9.59%

Moderately to Marginally Effective Drugs

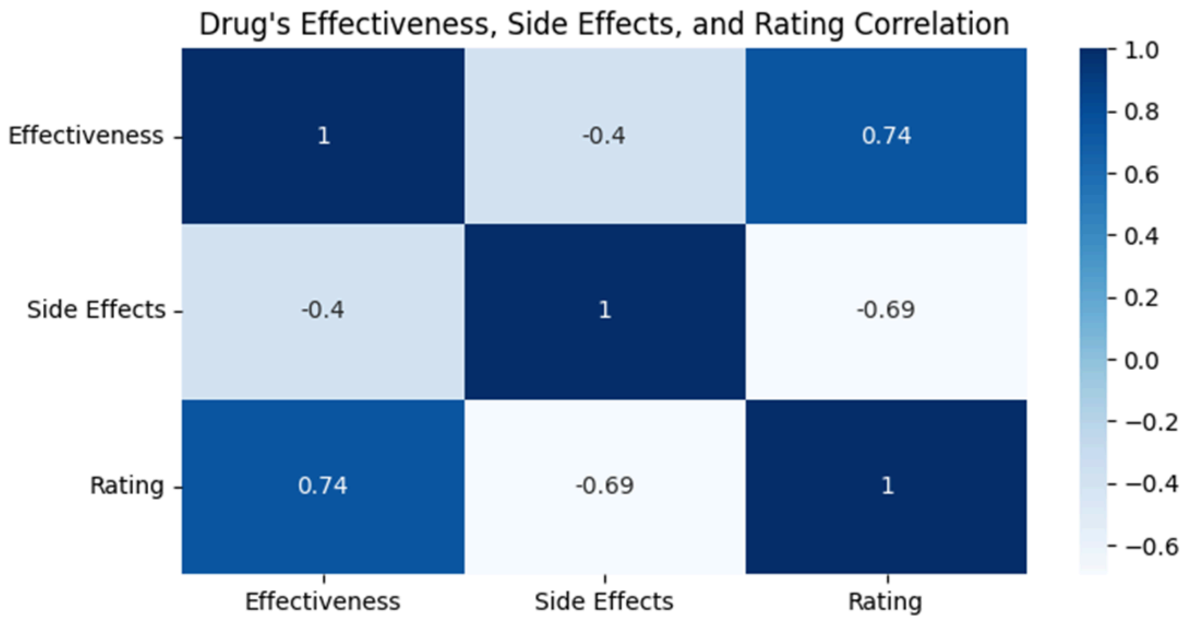
- Show a gradual increase in reported side effects
- Lower percentage of patients reporting no side effects (17-16%)
- Higher proportion of moderate to severe side effects

Ineffective Drugs

- Show the highest proportion of severe and extremely severe side effects
- Only 10.68% reported no side effects
- Nearly 29.31% reported extremely severe side effects

This analysis suggests a notable inverse relationship between drug effectiveness and the severity of side effects, with more effective medications generally associated with fewer and less severe side effects.

Correlation Analysis



The correlation heatmap reveals important relationships between three key variables in our drug review dataset:

Effectiveness and Rating

- A strong positive correlation of 0.74 exists between effectiveness and rating.
- This indicates that drugs perceived as more effective consistently receive higher overall ratings from patients.

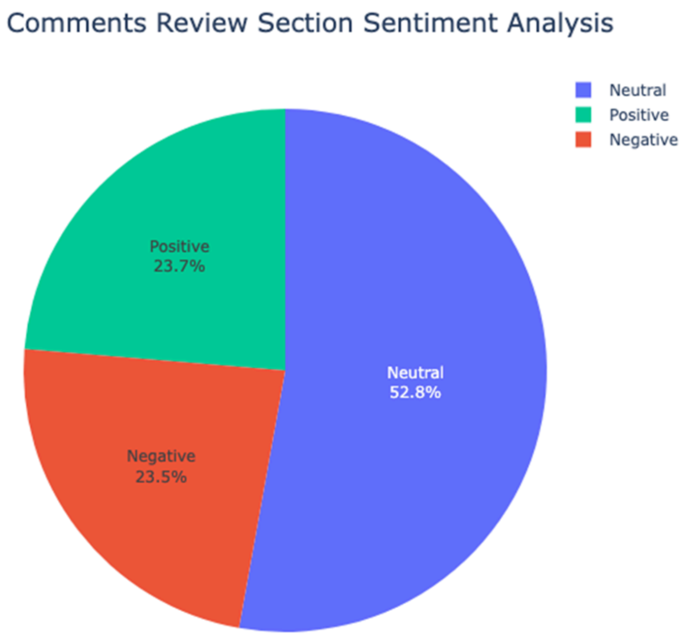
Side Effects and Rating

- A strong negative correlation of -0.69 between side effects and rating.
- Shows that drugs with more severe side effects tend to receive lower ratings from patients.

Effectiveness and Side Effects

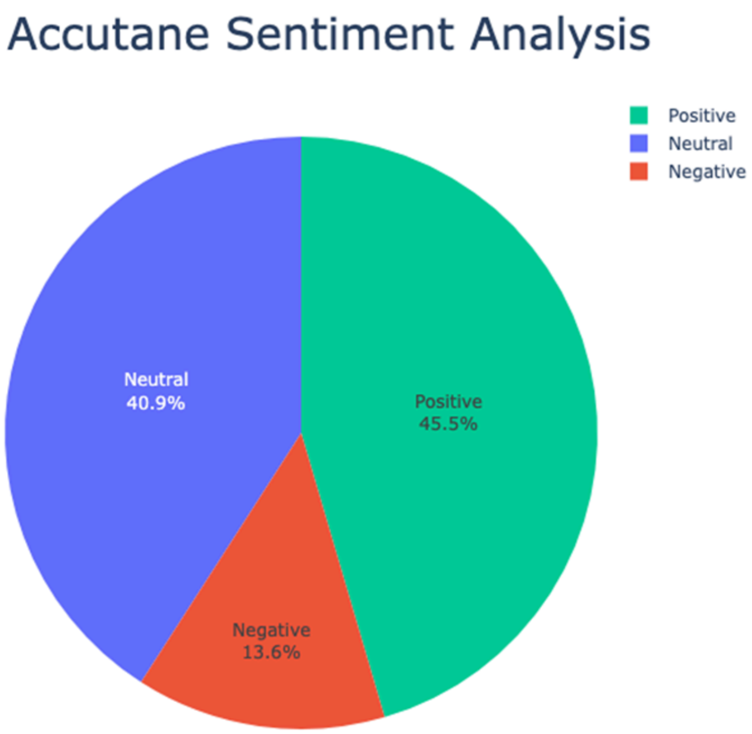
- A moderate negative correlation of -0.4 between effectiveness and side effects.
- Suggests that more effective drugs generally tend to have less severe side effects, though the relationship is not as strong as the other correlations.

Sentiment Analysis on *Comments Review*:



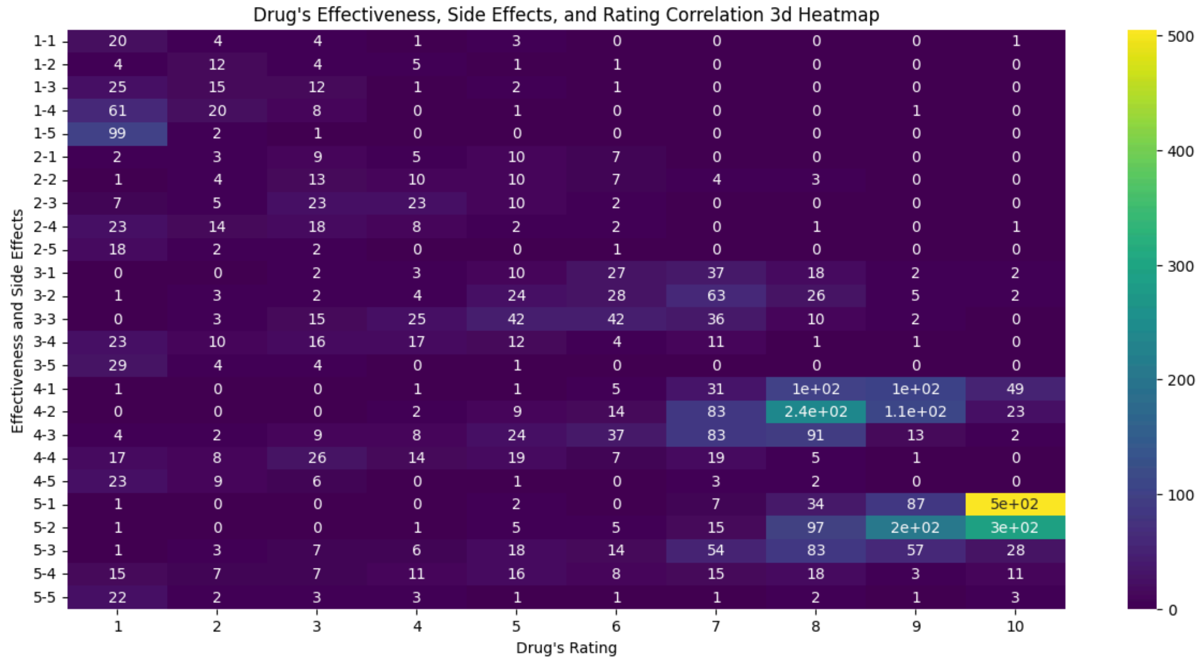
The sentiment analysis of the Comments Review section reveals that the majority of patient reviews (52.8%) express neutral sentiment, while positive and negative sentiments are almost equally distributed at 23.7% and 23.5% respectively.

Now for a particular drug **Accutane**:



Using sentiment analysis, we can observe that Accutane received predominantly favorable reviews, with 45.3% of patients expressing positive sentiments about their experience with the drug. A significant portion (40.9%) of the reviews maintained a neutral stance, while only 13.6% of patients reported negative sentiments. This distribution suggests that despite Accutane's known side effects, most patients found the treatment beneficial or at least satisfactory for their condition.

Drug's Effectiveness, Side effects, and Rating Correlation 3D Heatmap



It was seen that **rating has a positive correlation with effectiveness** while having a **negative correlation with side effects**. Also, the effectiveness of a drug has a slightly stronger correlation than side effects, meaning that Highly Effective drugs will probably guarantee a rating of no less than 7.

Accutane world cloud analysis :

Using word cloud visualization on Accutane data, it appears that the drug primarily affects the skin and causes dryness as a major side effect. The prominent words "dry," "skin," "lips," and "eyes" suggest that patients commonly experience dryness in these areas while using Accutane. The word cloud also indicates other effects like "depression" and mentions "sensitivity," highlighting both the treatment's effectiveness and its potential side effects.

Model training

We employed two approaches:

1) Latent Dirichlet Allocation (LDA)
2) BerTopic using two models for generating embeddings:
    a) "all-MiniLM-L6-v2"
    b) "all-mpnet-base-v2"

Approach 1: LDA Topic Modeling:

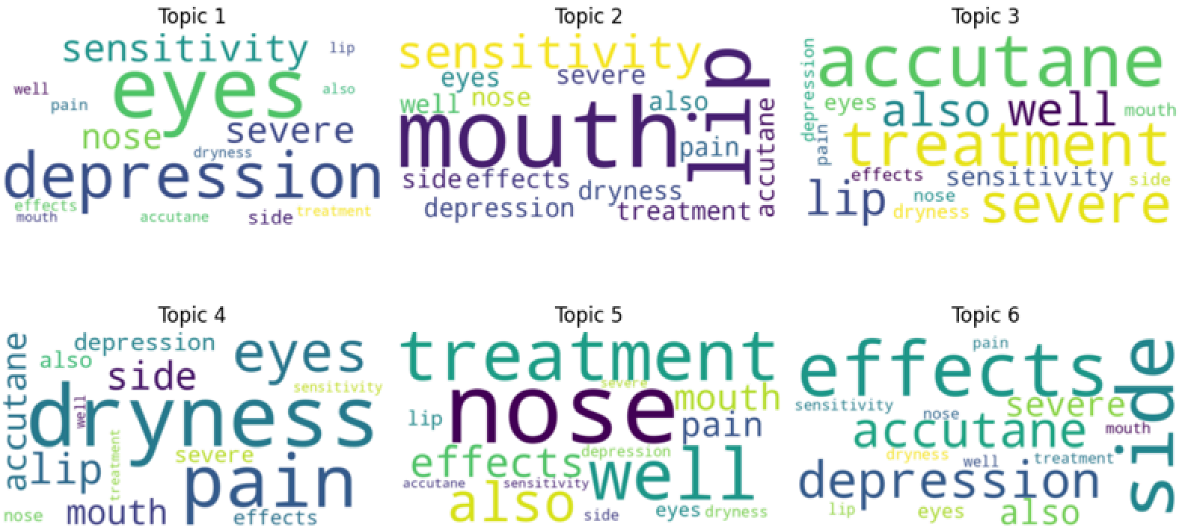
To develop our drug recommendation system, we first categorized reviews based on specific conditions (e.g., depression, dry skin) using the condition column. For each condition, we sorted drugs according to their ratings (1 to 10) using rating column to identify the top 10 most effective medications. To gain deeper insights into each top-rated drug, we conducted a detailed topic analysis using Latent Dirichlet Allocation (LDA).

Taking Accutane as an example, we employed LDA topic modeling on the preprocessed review data. The preprocessing stage involved cleaning and tokenizing the reviews, followed by creating a dictionary of terms. To enhance the quality of our analysis, we filtered out terms that appeared in fewer than 5 reviews or in more than 50% of the reviews, effectively reducing noise in the dataset. The LDA model was configured with 5 topics and trained with 10 passes over the corpus to ensure optimal convergence.

The analysis revealed distinct themes related to Accutane's effects and treatment experience. The most prominent topics centered around skin-related effects, particularly dryness affecting lips and eyes, which aligns with the word cloud visualization showing "dry," "skin," and "lips" as frequently mentioned terms. This systematic approach allowed us to understand both the primary effects and common experiences associated with each recommended drug.

This was the result of Accutane after LDA:

```
(0, '0.307*"eyes" + 0.223*"depression" + 0.138*"sensitivity" + 0.127*"severe" +
0.110*"nose" + 0.016*"side" + 0.010*"pain" + 0.009*"effects" + 0.009*"accutane" +
0.009*"well"')
(1, '0.320*"mouth" + 0.233*"lip" + 0.232*"sensitivity" + 0.018*"dryness" + 0.018
*"depression" + 0.018*"eyes" + 0.018*"treatment" + 0.018*"nose" + 0.018*"also" +
0.018*"severe"')
(2, '0.258*"accutane" + 0.176*"treatment" + 0.131*"severe" + 0.095*"lip" + 0.095
*"well" + 0.094*"also" + 0.031*"sensitivity" + 0.023*"eyes" + 0.015*"dryness" +
0.014*"depression"')
(3, '0.338*"dryness" + 0.247*"pain" + 0.136*"eyes" + 0.065*"lip" + 0.053*"side" +
0.041*"mouth" + 0.040*"accutane" + 0.016*"depression" + 0.014*"severe" + 0.014*"a
lso"')
(4, '0.242*"nose" + 0.205*"treatment" + 0.204*"well" + 0.107*"also" + 0.059*"effe
cts" + 0.058*"mouth" + 0.058*"pain" + 0.009*"lip" + 0.008*"eyes" + 0.008*"depress
ion"')
(5, '0.315*"side" + 0.297*"effects" + 0.107*"depression" + 0.100*"accutane" + 0.0
62*"also" + 0.062*"severe" + 0.011*"eyes" + 0.006*"pain" + 0.006*"well" + 0.006
*"dryness"')
```



The LDA topic modeling analysis of Accutane reviews revealed several distinct themes, supported by both probability distributions and visual representations. The model identified five main topics, with each topic characterized by specific word clusters and probabilities. Topic 1 prominently features terms related to sensitivity and eye-related side effects, while Topic 2 centers around mouth-related symptoms and treatment effects. Topic 3 focuses on the medication itself and treatment outcomes, with "accutane" and "treatment" being dominant terms. Topic 4 highlights the primary side effect of dryness, particularly affecting multiple body areas including eyes, lips, and nose. Topic 5 emphasizes the overall effects and side effects of the treatment. These findings are further reinforced by the word cloud visualization, which prominently displays terms like "dry," "skin," "lips," and "eyes" as central themes in patient experiences. The hierarchical organization of these terms in both the topic models and word cloud suggests that while Accutane is effective for skin treatment, patients consistently report dryness as a primary concern across multiple body areas, along with other side effects including potential mood changes and sensitivity issues.

Limitation of this approach: While this method provides valuable insights into drug effects and patient experiences, it is based on aggregated data from the entire observed dataset. As such, it lacks personalization for individual patients. To address this limitation and make recommendations more tailored, we implement BERTopic, which offers improved performance over LDA. Additionally, we incorporate each patient's specific preferences and medical history into the recommendation process, allowing for more personalized drug suggestions.

Approach 2 : Topic Modelling using BerTopic and patient's preferences/medical history:

To give a brief about BerTopic this is how it works:

Embedding Generation :
BERTopic starts by using a pre-trained language model (we used "all-MiniLM-L6-v2" and "all-mpnet-base-v2") to generate embeddings for each document. These embeddings are high-dimensional vector representations that capture the semantic meaning of the text.

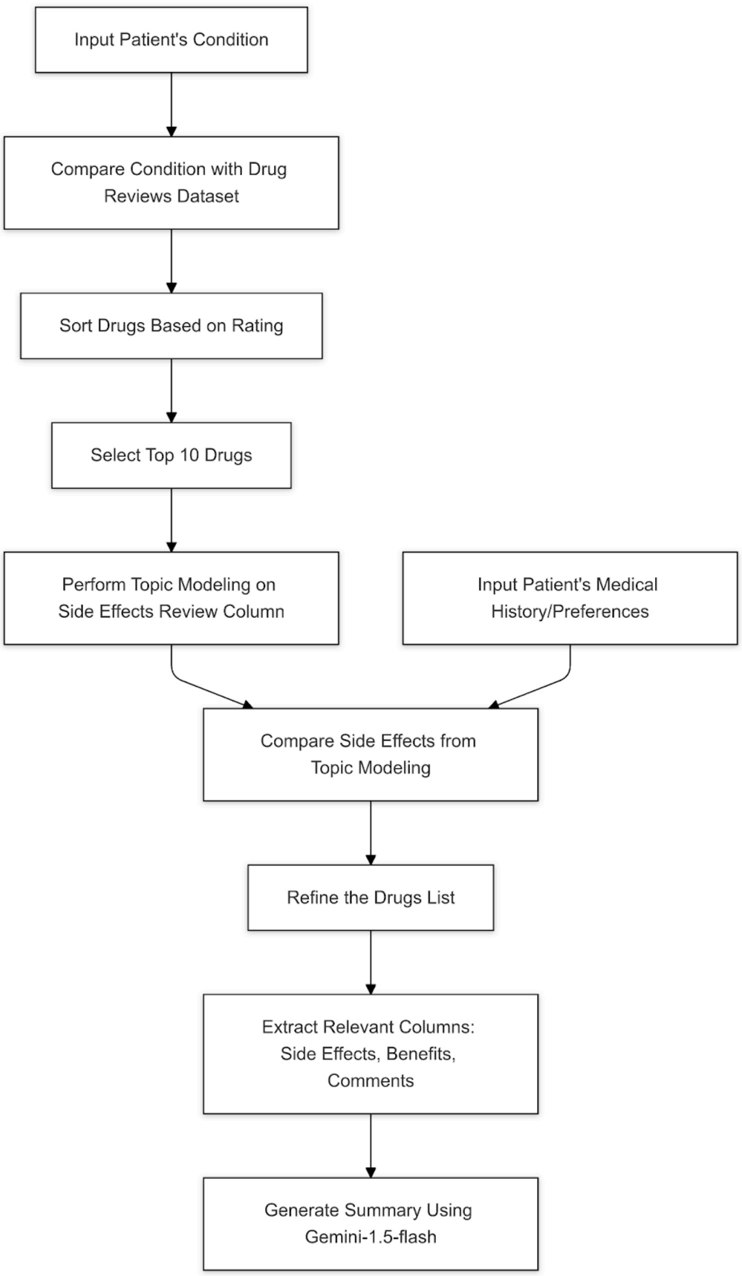
Dimensionality Reduction with UMAP :
The high-dimensional embeddings are then reduced to a lower-dimensional space using UMAP (Uniform Manifold Approximation and Projection). UMAP preserves both the local and global structure of the data, which is crucial for identifying meaningful topics.

Clustering :
The reduced embeddings are clustered using HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise). This step groups similar documents together, forming the basis for topics.

Topic Representation :
For each cluster, BERTopic calculates a c-TF-IDF (class-based Term Frequency-Inverse Document Frequency) representation. This highlights the most important words for each topic relative to other topics.

Topic Labeling :
Each topic is labeled using the top words from its c-TF-IDF representation.
Reference paper: https://arxiv.org/abs/2203.05794

Below is the flowchart for this approach:



The process begins by taking the patient's condition as input and matching it against a drug reviews dataset. The steps are as follows:

1. Sort drugs based on their ratings to identify the top 10 most effective medications for the given condition.
2. Perform topic modeling on the side effects review column to understand common patient experiences.
3. Incorporate the patient’s medical history and preferences to ensure personalized recommendations.
4. Compare identified side effects from topic modeling with the patient's specific conditions and preferences to refine the drug list.
5. Analyze the refined drug list across multiple review dimensions (side effects, benefits, and comments).
6. Generate a comprehensive summary for each drug using the Gemini-1.5-flash model.

This systematic approach ensures recommendations are not only based on general effectiveness but are also tailored to individual patient characteristics and concerns.

Topic modelling using BerTopic on side-effects column:

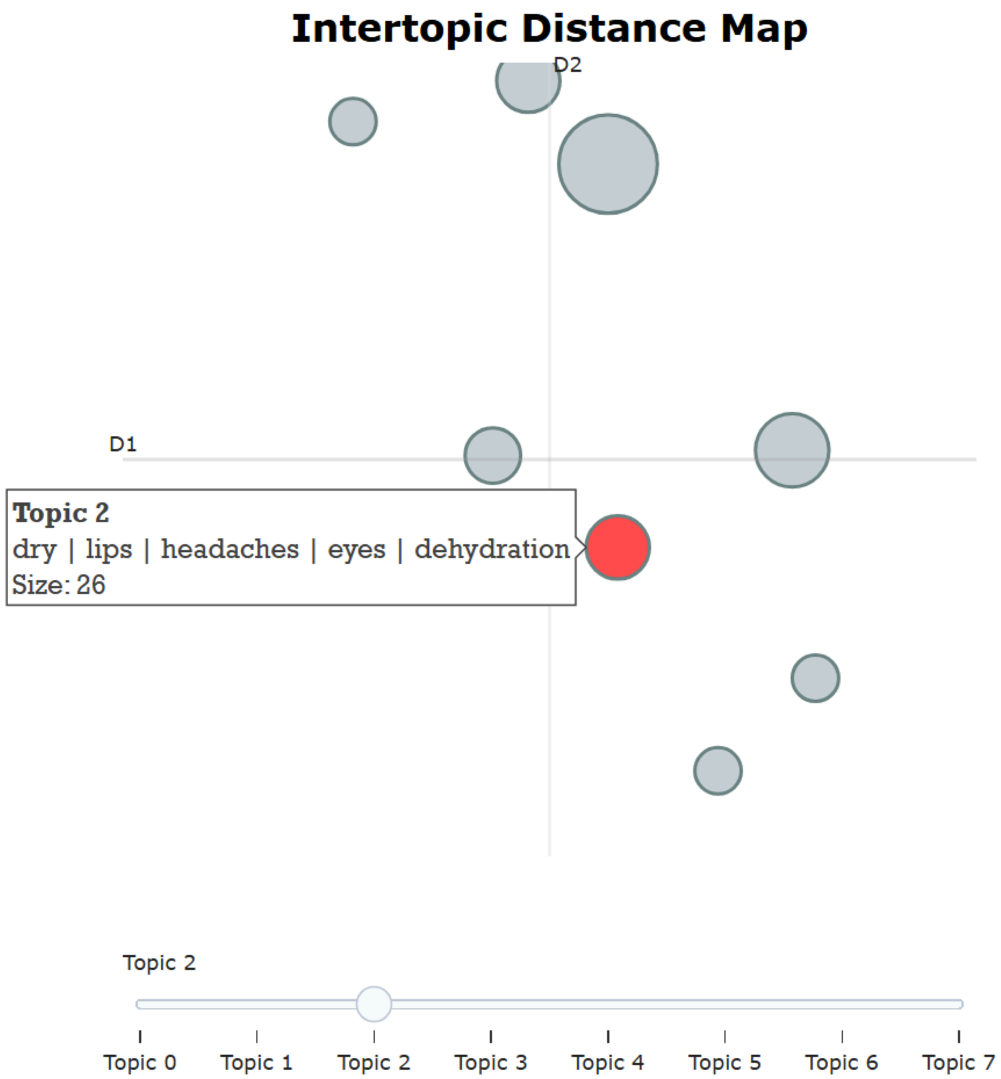
1) **Using “all-MiniLM-L6-v2”** :

```
Model: all-MiniLM-L6-v2
   Topic  Count                                  Name  \
0     -1     20         -1_bad_wax_accutane_depression
1      0     59               0_effects_did_years_time
2      1     34               1_skin_red_use_extremely
3      2     26              2_dry_lips_headaches_eyes
4      3     26                3_face_skin_taking_used
5      4     21  4_peeling_sensitivity_region_difficulty
6      5     20       5_irritation_sun_skin_moisturizer
7      6     14                                6_far___
8      7     14              7_retin_acne_strong_micro

                                      Representation  \
0  [bad, wax, accutane, depression, cracking, 12,...
1  [effects, did, years, time, experienced, treat...
2  [skin, red, use, extremely, photosensitivity, ...
3  [dry, lips, headaches, eyes, dehydration, sure...
4  [face, skin, taking, used, normal, aging, time...
5  [peeling, sensitivity, region, difficulty, dry...
6  [irritation, sun, skin, moisturizer, quickly, ...
7                             [far, , , , , , , , , ]
8  [retin, acne, strong, micro, bumps, worse, got...

                                 Representative_Docs
0  [I would have spent my last $2 on lip balm, ev...
1  [I had no side effects because I was applying ...
2  [I did not put it on near my eyes but somehow ...
3  [Dry lips, dry skin, dry hair, dry eyes and se...
4  [It does cause the skin to flake and peel but ...
5  [Peeling, dryness, sun sensitivity, peeling; d...
6  [My breakout were more frequent and severe whe...
7              [As above, none so far, none so far]
8  [I got my prescription for Retin-A Micro refil...
```

Clusters generated with the model can be viewed below: (Topic -1 doesn't get plotted as it's outlier)



Identified 8 distinct topics with document counts ranging from 14 to 59

Key Topics Identified :

- Topic -1: Depression and general side effects (20 documents)
- Topic 0: Long-term effects and treatment timeline (59 documents)
- Topic 1: Extreme skin sensitivity (31 documents)
- Topic 2: Dryness affecting multiple areas (26 documents)
- Topic 3: Skin aging and changes (24 documents)
- Topic 4: Regional sensitivity and peeling (21 documents)
- Topic 5: Sun sensitivity and moisturization (20 documents)
- Topic 6: Far (which didn't matched with others) (14 documents)
- Topic 7: Retinoid treatment effectiveness (14 documents)

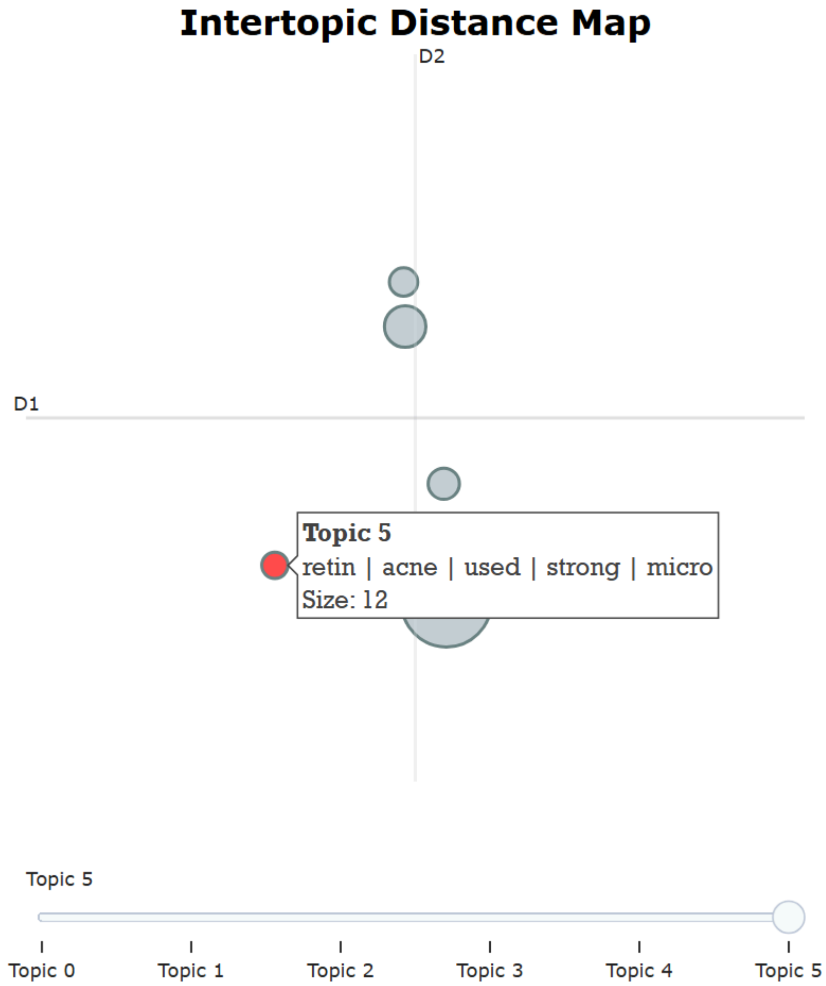
2) **Using "all-mpnet-base-v2"** :

```
Model: all-mpnet-base-v2
   Topic  Count                                 Name  \
0     -1      8   -1_differin_breast_apply_vaseline
1      0    140             0_skin_dry_lips_effects
2      1     30  1_sensitivity_sun_skin_moisturizer
3      2     16   2_effects_adverse_food_discomfort
4      3     14                             3_far___
5      4     14            4_irritation_skin_dry_sun
6      5     12             5_retin_acne_used_strong

                                      Representation  \
0  [differin, breast, apply, vaseline, breasts, c...
1  [skin, dry, lips, effects, dryness, treatment,...
2  [sensitivity, sun, skin, moisturizer, sunscree...
3  [effects, adverse, food, discomfort, drink, me...
4                           [far, , , , , , , , , ]
5  [irritation, skin, dry, sun, moisturizer, aggr...
6  [retin, acne, used, strong, micro, tolerance, ...

                                 Representative_Docs
0  [With the increase of breast size also came ex...
1  [First, let me say that I didn't experience th...
2  [It does cause the skin to flake and peel but ...
3  [I don't believe I experienced any side effect...
4              [As above, none so far, none so far]
5  [dry skin and skin irritation, Dry skin was ea...
6  [Peeling which can chafe the skin and interfer...
```

Clusters generated with the model can be viewed below: (Topic -1 doesn't get plotted as it's outlier)



BERTopic Model Analysis

Our analysis utilized two BERT-based models for topic modeling on the drug review dataset, revealing distinct patterns and themes:

MPNet Model (all-mpnet-base-v2)
Generated 7 distinct topics with varying document distributions (8-140 documents)
Key Topics Identified:

- Topic -1: Breast-related effects and vaseline application (8 documents)
- Topic 0: Skin dryness and lip effects (140 documents)
- Topic 1: Sun sensitivity and moisturizer usage (30 documents)
- Topic 2: Adverse effects and food discomfort (16 documents)
- Topic 3: General treatment experiences (14 documents)
- Topic 4: Skin irritation and sun sensitivity (14 documents)
- Topic 5: Retinoid treatment for acne (12 documents)

Comparative Analysis

Both models effectively captured key themes but with different emphases:

- The MPNet model showed stronger clustering around physical side effects and treatment applications.
- The MiniLM model provided more granular separation of psychological and long-term effects.
- Both models consistently identified important themes around skin sensitivity, dryness, and treatment effectiveness.
- The representation and representative documents in both models align well with known Accutane effects and patient experiences.

This dual-model approach provides comprehensive insight into patient experiences, with each model contributing unique perspectives to our understanding of drug effects and treatment outcomes.

Above were the topic modelling results on the entire side effect column, below we have performed topic modelling on a specific drug like Accutane using all-MiniLM-L6-v2 model:

```
bertopic_results['accutane']['all-MiniLM-L6-v2']['model'].get_topic_info() # depression,redness
```

	Topic	Count	Name	Representation	Representative_Docs
0	0	14	0_skin_lotion_dry_lips	[skin, lotion, dry, lips, bad, 12, water, use,...	[Dry skin and lips, increased susceptibility t...
1	1	8	1_depression_skin_dry_hands	[depression, skin, dry, hands, hips, worst, cr...	[Hair loss! Not so much from my head, but othe...
2	2	6	2_effects_sensitive_pain_accutane	[effects, sensitive, pain, accutane, joint, to...	[None - I took accutane and it changed my life...
3	3	4	3_sensitivity_dry_light_sunlight	[sensitivity, dry, light, sunlight, lips, hair...	[Chapped lips, dry skin, sensitivity to light....
4	4	4	4_eyes_little_chaped_dry	[eyes, little, chaped, dry, lip, nose, dryness...	[lip and eyes dryness, Very very dry skin. Lip...
5	5	4	5_birth_great_dermatology_control	[birth, great, dermatology, control, company, ...	[Headaches, dry lips, dry skin, elevated trigl...
6	6	4	6_redness_red_skin_extreme	[redness, red, skin, extreme, dryness, dry, , ...	[red skin, dry skin, Extreme skin dryness and ...

The BERTopic analysis of Accutane reviews using the all-MiniLM-L6-v2 model revealed 7 distinct topics, providing insights into patient experiences. The largest cluster (Topic 0) focused on skin-related effects, with 14 documents discussing dry skin and lips. Other significant topics included psychological impacts like depression (Topic 1), overall treatment effects (Topic 2), and environmental sensitivity (Topic 3). The analysis also captured localized effects on eyes and lips (Topic 4), aspects of medical monitoring (Topic 5), and severe skin reactions (Topic 6). Each topic was characterized by relevant key terms and representative documents, effectively capturing the diverse range of patient experiences with Accutane, from physical side effects to psychological impacts.

Different Methods Tried for Patient Input and Challenges Faced

Topic modeling, while appearing straightforward, presents significant challenges when processing patient preferences and medical history. For example, when analyzing a sample patient preference: "I do not want to experience dry skin, headache, or nausea while using these medications. I'm also concerned about potential hair loss and mood changes," different techniques yielded varying results:

- **LDA** extracted basic terms: ['want', 'using', 'skin', 'potential', 'nausea'].
- **BERTopic** proved ineffective for short inputs, requiring longer text passages for meaningful analysis.
- **YAKE** identified key phrases: ['experience dry skin', 'dry skin', 'experience dry', 'headache', 'skin', 'medications', 'potential hair loss', 'experience'].

Through experimentation, we found that using explicit keywords representing side effects to avoid (e.g., ["dry skin", "headache", "nausea", "depression"]) provided the most effective approach. When applying sentiment analysis to side effect reviews, we encountered the challenge of medical term specificity For eg: dry skin and dry lips even though they seem similar when we think in terms of the word “Dry” but in medical terms they do differ significantly and so do their treatment procedure and drug recommendation for cure.

The effectiveness of this approach largely depends on dataset size - a larger dataset provides better coverage of conditions and their variations. While we cannot capture every possible condition, a comprehensive dataset helps improve the accuracy and relevance of drug recommendations based on patient preferences.

For Patient preferences (side effects the patient wants to avoid): **["dry skin", "headache", "nausea","depression"]** these were the results.

We found that when we added “**depression**” in the input the drug accutane was getting eliminated so it confirmed that our approach was working which can be seen below.

```
bertopic_results['accutane']['all-MiniLM-L6-v2']['model'].get_topic_info() # depression,redness
```

	Topic	Count	Name	Representation	Representative_Docs
0	0	14	0_skin_lotion_dry_lips	[skin, lotion, dry, lips, bad, 12, water, use,...	[Dry skin and lips, increased susceptibility t...
1	1	8	1_depression_skin_dry_hands	[depression, skin, dry, hands, hips, worst, cr...	[Hair loss! Not so much from my head, but othe...
2	2	6	2_effects_sensitive_pain_accutane	[effects, sensitive, pain, accutane, joint, to...	[None - I took accutane and it changed my life...
3	3	4	3_sensitivity_dry_light_sunlight	[sensitivity, dry, light, sunlight, lips, hair...	[Chapped lips, dry skin, sensitivity to light....
4	4	4	4_eyes_little_chaped_dry	[eyes, little, chaped, dry, lip, nose, dryness...	[lip and eyes dryness, Very very dry skin. Lip...
5	5	4	5_birth_great_dermatology_control	[birth, great, dermatology, control, company, ...	[Headaches, dry lips, dry skin, elevated trigl...
6	6	4	6_redness_red_skin_extreme	[redness, red, skin, extreme, dryness, dry, , ...	[red skin, dry skin, Extreme skin dryness and ...

```
Recommended drugs (side effects not matching patient's concerns):
- differin
- retin-a
- spironolactone
- minocycline
- retin-a-micro
- sotret
- renova
- avita

Filtered out drugs (potential side effects matching patient's concerns):
- doxycycline
- accutane
```

We can see **accutane** being filtered out in the above list. After this for the remaining drugs which will be recommended to the patient we pass this to the **Gemini-1.5-flash** model to generate a summary for every drug in the recommended drug list to help the patient choose the drug as per his requirement based on effectiveness, it's most observed side effect, rating. The drugs are sorted in descending order according to their rating which can be seen below.

Summary using Gemini-1.5-flash:

```
Based on the provided data, here's a summary of each acne medication and a recommendation of the top 10 drugs:


**Drug Summaries:**

* **Minocycline:**  This antibiotic is often effective in reducing acne lesions, but it carries a risk of severe side effects s
uch as joint pain, headaches, hives, and swelling.  The effectiveness varies significantly, with some reporting only marginal i
mprovement or no benefit at all.  While some users experienced a decrease in pimples and cysts, others noticed no change or eve
n a worsening of their condition. Chronic yeast infections are also a reported side effect.  The average rating is 5.2.

* **Retin-A Micro:** This topical retinoid can improve skin texture and reduce acne, especially initial use, but it is also kno
wn to cause dryness, redness, peeling and increased sun sensitivity. Effectiveness varies, with some experiencing a significant
improvement, while others saw no benefit or even an increase in acne severity. The average rating is 7.4.

* **Retin-A:**  Similar to Retin-A Micro, this topical retinoid can be effective but causes dryness, irritation, peeling, sting
ing, and increased sun sensitivity.  Some users found it exacerbated their rosacea. Effectiveness also varies. The average rati
ng is 7.8.

* **Differin:** This topical retinoid can help reduce acne inflammation and improve skin texture, but some users found it ineff
ective or even made their acne worse, possibly due to an initial purging effect. It's usually associated with mild side effect
s, primarily dryness and irritation. The average rating is 8.1.

* **Avita:** This topical tretinoin cream is reported as moderately effective in clearing acne and improving skin texture.  The
most common side effect is dryness, and some experienced redness and peeling. The average rating is 7.6.

* **Spironolactone:** This medication may reduce acne severity, especially when combined with other acne treatments. It is, how
ever, noted for potentially causing significant side effects, including extreme fatigue and frequent urination.  The reported e
ffectiveness varies significantly. The average rating is 8.8.

* **Sotret (Isotretinoin):** This is a powerful medication that is very effective for many patients at permanently clearing up
acne.  However, it's associated with more serious side effects, including severe dryness, depression, and other potential issue
s.  It is vital that this medication is prescribed and monitored closely by a dermatologist.  The average rating is 9.3.

* **Renova:** This topical retinoid is often used for acne and wrinkles.  It can reduce acne but may cause dryness, redness, pe
eling, and skin sensitivity. The average rating is 8.2.


**Top 10 Acne Medications Based on Overall Rating:**

1. Sotret (Isotretinoin)
2. Spironolactone
3. Retin-A
4. Differin
5. Retin-A Micro
6. Renova
7. Avita
8. Minocycline


**Important Disclaimer:**

This information is for general knowledge and does not constitute medical advice.  The effectiveness and side effects of acne m
edications can vary greatly from person to person.  It is crucial to consult a qualified healthcare professional before startin
g any new medication, including those listed above.  They can help you determine the best treatment plan based on your individu
al needs and medical history.  While these drugs show generally positive results for acne treatment, individual responses can b
e unpredictable.
```

The prompt given to Gemini-1.5-flash to generate this response was:
Prompt used for Gemini:

instructions =f"You are a medical information assistant. I have a dataframe containing information about {**user_info[0]**} condition. The columns include 'urlDrugName', 'rating', 'effectiveness', 'sideEffects', 'condition', 'benefitsReview', 'sideEffectsReview', 'commentsReview', and 'overall_rating_score'.For each drug in the dataframe:Summarize the key information, including:The drug's name. Its average rating.Its effectiveness.Common side effects.A brief overview of its benefits.A concise summary of user-reported side effects.After summarizing each drug, provide a recommendation of the top {**len(recommended_drugs)**} drugs based on their overall rating scores in descending order and patient's preferences.Conclude with a brief statement about the importance of consulting a healthcare professional before starting any new medication.Present this information in a clear, easy-to-read format that a patient could understand. Use bullet points where appropriate, and keep medical jargon to a minimum. Remember to emphasize that while these drugs are generally effective for {**user_info[0]**}, individual experiences may vary."

Conclusion and Future Perspectives

Our analysis revealed several opportunities for improvement and future development:

Model Consistency and Enhancement

- While Gemini provided consistent thematic responses, the quality varied between iterations. This variability could be addressed through refined prompt engineering techniques.
- Implementing more advanced language models like GPT for summary generation and embeddings in BERTopic could enhance topic modeling quality and improve drug recommendations based on patient preferences.

System Robustness

- Integration of Retrieval-Augmented Generation (RAG) could make the approach more robust and reliable.
- The current system demonstrates effective drug recommendation capabilities, but could benefit from additional layers of verification and contextualization.

User Interface Improvements

- While the current list-based recommendation system is functional, a conversational interface (chatbot) could provide a more user-friendly experience.

- A chatbot implementation would be particularly beneficial when patients need to choose between multiple drug options (e.g., 5-6 medications), offering interactive guidance through the decision-making process.

These enhancements would contribute to a more comprehensive and user-friendly drug recommendation system, better serving both healthcare providers and patients in making informed medication choices.

THANK YOU