Beyond the Reviews: Understanding Wellness Intent and Supplement Experience through NLP

Mitali Acharya

Matriculation number: 30008255

Data Science for Society and Business (DSSB), Constructor University, Bremen

MDSSB-MET-02: Text Analysis and Natural Language Processing, Spring 2025

Dr. Adalbert Wilhelm, Dr. Matthias Meckel

31 May 2025

1. Executive Summary

This project, ***"Beyond the Reviews: Understanding Wellness Intent and Supplement Experience through NLP,"*** analyses over 31,000 Reddit posts from supplement-related communities using Natural Language Processing. Multivitamin supplements have become a regular part of daily health routines. People increasingly depend on online discussions and reviews to make their choices about taking the right supplement. But online information is often conflicting and confusing. Through techniques such as Topic Modeling, Sentiment analysis, and supplement brand extraction, this project aims to uncover the most frequently discussed multivitamin supplements, common health goals, and the differences between what users ask and what they talk about.

Key Findings:

- **Sleep, fatigue, anxiety, gut health, and focus** emerged as the most frequently discussed health concerns across Reddit comments.
- **Thorne** is the most consistently mentioned brand across all conversations.
- **Sentiment analysis** showed mostly neutral tone across discussions, with negative sentiment tied to unmet expectations or unclear results.
- A clear contrast emerged between what users ask (titles) and what they share (comments), offering valuable insight into both **intent and lived experience** in supplement use. People are mainly asking about supplement choices, side effects, mental health experiences, and seeking advice on supplement combinations or fitness-related outcomes. In the comments section, the focus is more on sharing personal outcomes related to energy, sleep, anxiety, and gut health. Thorne is frequently mentioned in comments, which suggests that curiosity doesn't always translate into long-term trust or usage, a key behavioural contrast.

2. Introduction and motivation

In today's fast-paced world, people experience burnout, exhaustion, and high levels of stress, especially students, workers, and those who live away from home. People find it harder to stay healthy. People often eat at odd times; they do not get enough rest. Many people also lack access to regular health care. As a result, dietary multivitamin supplements have become a common part of daily health routines to boost energy, fill nutritional gaps, or improve overall well-being (NIH, 2022). Picking the right supplement can be confusing. The internet is filled with many conflicting personal reviews, different claims, and different emotional stories, which makes it difficult to know what actually helps (Mehta et al., 2021).

This is a space where Natural Language Processing (NLP) offers new dimensions. Social media platforms and online forums contain many health-related lived experiences from users. Recent studies show that NLP methods, for example, Topic modelling and sentiment analysis, can be applied to social media to identify user sentiment, experience patterns, and health-related concerns (Palanisamy & Jha, 2023; Okon et al., 2020). The methods find user sentiments, patterns, and health concerns by turning unstructured discussions into clear insights. This helps people make better and more personal choices.

This project began from a personal struggle. Last year, when I moved to Germany, I felt the need to take multivitamin supplements due to sudden lifestyle and weather changes. I tried to look for multivitamins, thinking they would help. But picking the right one was hard. The problem was that each supplement had very different reviews. Some people mentioned that it changed their lives. Others wrote that it did nothing. I tried some well-rated ones, but didn't feel much difference in my health. **This experience made me think: What helps, and for whom?** Do these comments show patterns that we miss? Can we look past ratings and ads to understand what people feel about multivitamin supplements? This led to my research question: **How can NLP techniques extract wellness intent and lived supplement experiences from Reddit discussions to support personalized supplement recommendations?**

The purpose of this exploratory work is to determine whether or not unstructured health conversations can be turned into meaningful insights. User content, especially Reddit comments, reveals more than just product feedback. It includes personal health goals, problems, as well as human emotional nuances that we often don't notice. To answer my research question, I applied different Natural Language Processing techniques, including sentiment analysis and topic modeling.

3. Sample, Data, and Corpus

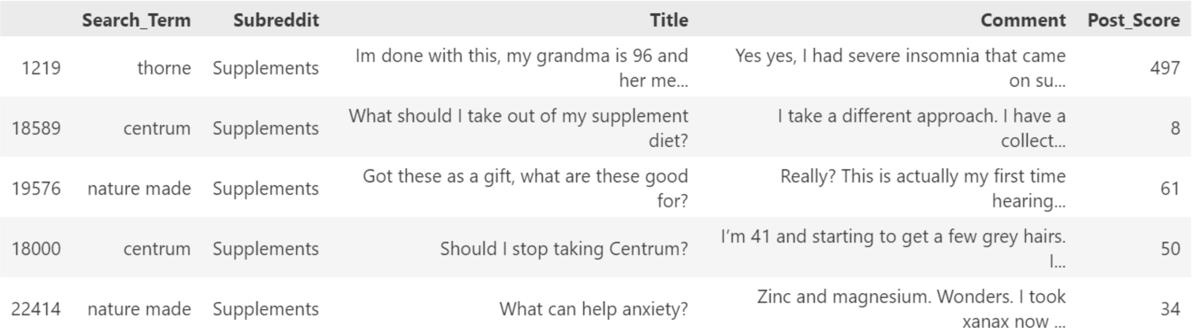
I extracted the dataset for this project from Reddit. This public forum has many discussions about health, personal experiences, and supplement recommendations. Instead of curated product reviews, this data includes open, user-written stories making it well-suited for large-scale analysis.

I scraped the data from multiple health-related subreddits, including r/Supplements, r/Nutrition, and r/Vitamins. Using a keyword-based approach, approximately **36,130 comments** were collected from the above subreddits. Each Reddit post and its associated comments, and other metadata, were scraped using the PRAW (Python Reddit API Wrapper) library.

Table 1: *Overview of Reddit Dataset Variables and Descriptions*

Variable Name	Description	Values / Type
Search_Term	Keyword used for searching (e.g., magnesium, multivitamin)	Categorical (string)
Subreddit	Source subreddit	Categorical
Title	Title of the original Reddit post	String
Post_ID	Unique ID of the post	String
Author	Author of the post	String
URL	Direct link to the Reddit post	URL string
Post_Score	Upvotes on the post	Integer
Num_Comments	Number of comments on the post	Integer
Created_UTC	Post creation time (UTC)	Float
Comment	Body text of the user comment	Text
Comment_Author	Author of the comment	String
Comment_Score	Upvotes on the comment	Float

Figure 1: *Sample rows from the Reddit supplement dataset, showing user posts and associated metadata.*



	Search_Term	Subreddit	Title	Comment	Post_Score
1219	thorne	Supplements	Im done with this, my grandma is 96 and her me...	Yes yes, I had severe insomnia that came on su...	497
18589	centrum	Supplements	What should I take out of my supplement diet?	I take a different approach. I have a collect...	8
19576	nature made	Supplements	Got these as a gift, what are these good for?	Really? This is actually my first time hearing...	61
18000	centrum	Supplements	Should I stop taking Centrum?	I'm 41 and starting to get a few grey hairs. I...	50
22414	nature made	Supplements	What can help anxiety?	Zinc and magnesium. Wonders. I took xanax now ...	34

In comparison to structured product reviews, these Reddit comments vary in length, tone, and clarity, ranging from brief remarks to in-depth advice. This difference makes the data messy but more real, as it gives a valuable insight into wellness concerns and human lived experiences.

4. Descriptive Statistics

4.1 Distribution of Comments Across Supplement-Related Subreddits

Figure 2: *Number of Comments per Subreddit in the Reddit Supplement Dataset*

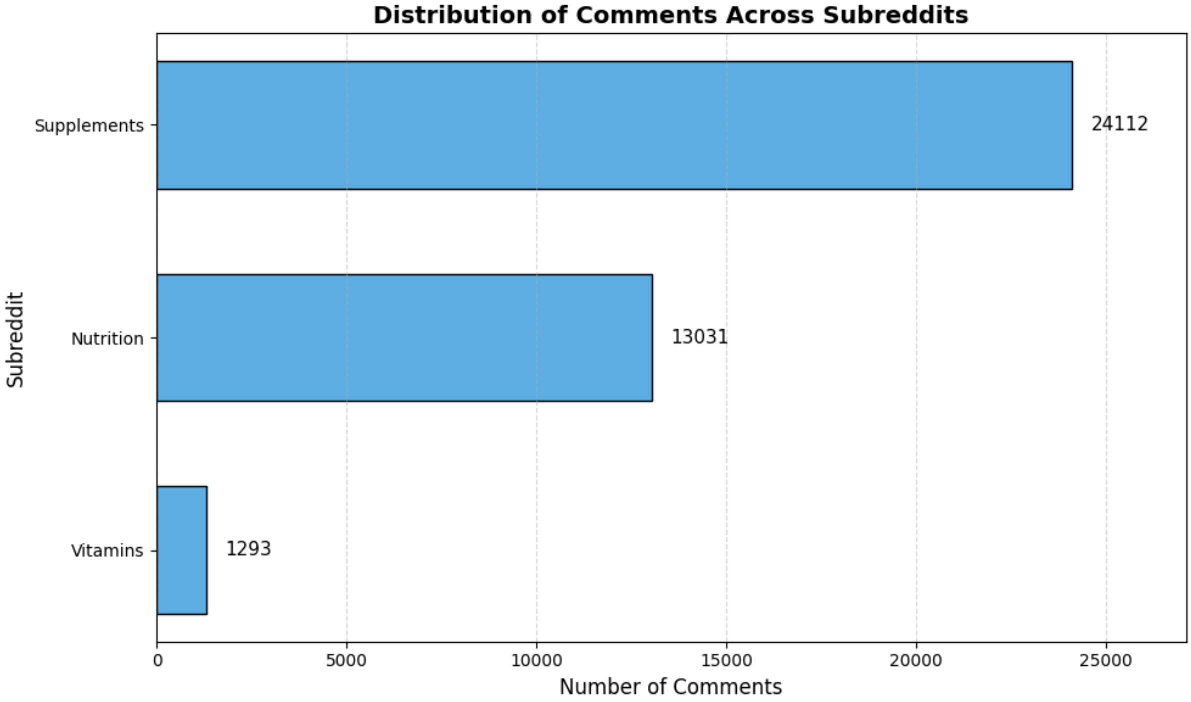


Figure 2 shows the distribution of comments across subreddits, highlighting where users were most active. Most comments came from r/Supplements, the primary subreddit for wellness and supplement discussions.

5. Main Analysis, Methodology:

5.1 Main Analysis:

This project examines Reddit conversations about dietary supplements using natural language processing to learn user tone, sentiments towards supplements, and their supplement choices. Motivated by previous research using NLP to extract user experiences from online platforms (Rosamma KS, 2024; Danish & Amjad, 2023; Zhang, 2022), I applied sentiment analysis, brand/entity extraction, and topic modeling to extract structured insights from unstructured data.

The methodology has three main components:

1. **Supplement Mention Extraction**

I used a whitelist-based approach to extract the most frequently mentioned supplements from comment texts. This step helped me identify the most popular multivitamin supplement.

2. **Sentiment Analysis**

I applied both VADER (rule-based) and the RoBERTa (transformer-based) models to classify user sentiment. This approach offers more nuanced emotional classification than the earlier work's simpler TextBlob method (Rosamma KS, 2024).

3. **Topic Modeling**

I applied Topic Modeling to uncover common health concerns and supplement-related themes, Models like LDA often produce less interpretable results on unstructured Reddit data, while CorEx enables more meaningful topic discovery through anchor words and mutual information (Rizvi et al., 2019). Inspired by Rizvi's work on supplement-related questions, I used CorEx to extract both wellness themes from comments and free-form discussion topics from titles. This helped identify both user intent (e.g., sleep, energy) and actual supplement experiences.

5.2 Methods:

5.2.1 Text cleaning

I started with basic cleaning of raw Reddit data to prepare the data for analysis. This step included converting text to lowercase and removing links, special characters, HTML tags, and extra spaces. More importantly, I also addressed issues with missing data and duplicate comments.

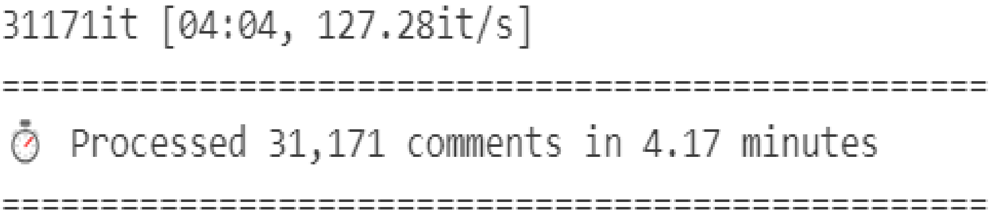
I dropped 2,331 comments because they were empty or null. I found 4,934 duplicate comments, so I dropped them. After this process, I had 31,105 distinct comments, which I used for further analysis.

5.2.2 Text Preprocessing

The next step is Text Preprocessing. I tokenized and lemmatized the text using **spaCy**, a faster and efficient Natural Language Processing library than NLTK, which is used more in academic projects. Preprocessing involved breaking each comment into individual tokens (words), removing punctuation and stopwords, and reducing words to their base form through lemmatization (e.g., "taking" → "take").

I built a custom batch function using **spaCy.pipe()** to process the text to improve processing speed and memory efficiency. That way, the text was processed in batches. This method was faster and more scalable than processing each comment individually. These techniques are widely adopted in NLP projects and have been successfully demonstrated in prior work combining spaCy with practical text analysis tasks (Olabiyi, Olaoye, & Daniel, 2024). As shown in the result below, over 31,000 comments were processed in under five minutes. This task usually took ten to fifteen times longer with basic NLTK without batching.

Figure 3: *Processing Speed and Time for Comment Preprocessing using spaCy*



```
31171it [04:04, 127.28it/s]
==================================================
⏱ Processed 31,171 comments in 4.17 minutes
==================================================
```

I also created a **whitelist of supplement names and phrases** (e.g., “vitamin d”, “omega 3”, “ashwagandha”) and ensured these were **preserved during tokenization** to prevent them from being mistakenly split or altered.

Multiple Text Versions for Comparative Analysis

During processing, I created three different text versions to support later analyses. This helped me find how various levels of cleaning influenced the analysis outcomes.

1. tokens-dirty

This version keeps only tokenized words with punctuation removed, but **stopwords** are retained. It was useful for understanding natural sentence structure and informal phrasing.

2. tokens

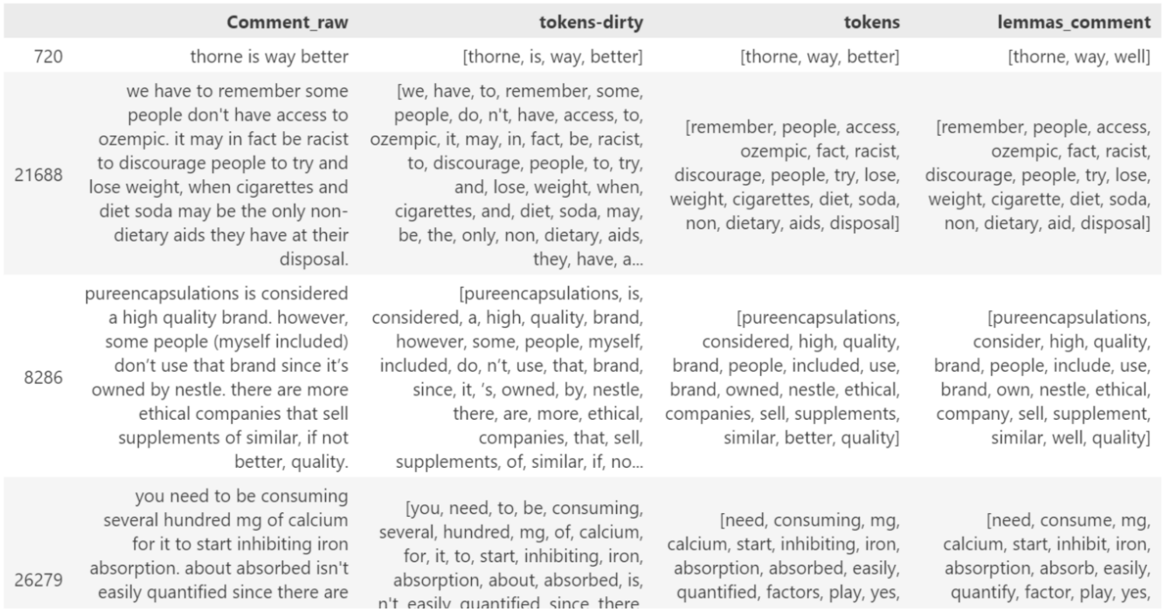
This version contains **tokenized words with both stopwords and punctuation removed**.

3. lemmas

This version includes lemmatized, cleaned tokens, with both stopwords and punctuation removed. Lemmas are given as **primary input for topic modeling and all further text-based analyses**, as they provide a clean yet semantically meaningful representation of each comment.

For illustration, a few comments are extracted and shown in the figure below. Their tokenized and lemmatized versions are also shown. The comments illustrate different types of supplement-related discussions. Some discuss brand trust. Some discuss how chemicals work together. This demonstrates how spaCy’s pipeline cleaned the language without losing nuance.

Figure 4: *Sample Reddit Comments with Tokenized and Lemmatized Outputs*



	Comment_raw	tokens-dirty	tokens	lemmas_comment
720	thorne is way better	[thorne, is, way, better]	[thorne, way, better]	[thorne, way, well]
21688	we have to remember some people don't have access to ozempic. it may in fact be racist to discourage people to try and lose weight, when cigarettes and diet soda may be the only non-dietary aids they have at their disposal.	[we, have, to, remember, some, people, do, n't, have, access, to, ozempic, it, may, in, fact, be, racist, to, discourage, people, to, try, and, lose, weight, when, cigarettes, and, diet, soda, may, be, the, only, non, dietary, aids, they, have, a...	[remember, people, access, ozempic, fact, racist, discourage, people, try, lose, weight, cigarettes, diet, soda, non, dietary, aids, disposal]	[remember, people, access, ozempic, fact, racist, discourage, people, try, lose, weight, cigarette, diet, soda, non, dietary, aid, disposal]
8286	pureencapsulations is considered a high quality brand. however, some people (myself included) don't use that brand since it's owned by nestle. there are more ethical companies that sell supplements of similar, if not better, quality.	[pureencapsulations, is, considered, a, high, quality, brand, however, some, people, myself, included, do, n't, use, that, brand, since, it, 's, owned, by, nestle, there, are, more, ethical, companies, that, sell, supplements, of, similar, if, no...	[pureencapsulations, considered, high, quality, brand, people, included, use, brand, owned, nestle, ethical, companies, sell, supplements, similar, better, quality]	[pureencapsulations, consider, high, quality, brand, people, include, use, brand, own, nestle, ethical, company, sell, supplement, similar, well, quality]
26279	you need to be consuming several hundred mg of calcium for it to start inhibiting iron absorption. about absorbed isn't easily quantified since there are	[you, need, to, be, consuming, several, hundred, mg, of, calcium, for, it, to, start, inhibiting, iron, absorption, about, absorbed, is, n't, easily, quantified, since, there	[need, consuming, mg, calcium, start, inhibiting, iron, absorption, absorbed, easily, quantified, factors, play, yes,	[need, consume, mg, calcium, start, inhibit, iron, absorption, absorb, easily, quantify, factor, play, yes,

5.2.3 Sentiment Analysis

I began the sentiment analysis with **VADER (Valence Aware Dictionary for Sentiment Reasoning)**, a rule-based model designed for short, informal social media text. Studies have shown that VADER outperforms traditional models and sometimes even individual human critics in sentiment classification, achieving an **F1 accuracy of 96%** (Hutto & Gilbert, 2014).

Unlike basic keyword-based sentiment tools, VADER combines a well-constructed lexical dictionary with five grammatical and syntactical rules that account for how sentiment is expressed in real-world online language (Hutto & Gilbert, 2014). These rules allow the model to understand elements like intensifiers ("very good"), negations ("not great"), punctuation ("!!!"), capitalization ("AMAZING"), and contrastive conjunctions ("but"), all of which influence the emotional tone of a sentence. For example, it correctly interprets that "This really works!!!" is far more positive than "This works,".

I extracted comments from the dataset and tested VADER's performance to assess its ability to identify emotional tone.

Example 1: "This supplement is absolutely amazing!!!" → Positive

In the above example, VADER sees the word *'amazing'*, which it recognizes as positive. Then it sees *'absolutely'*, an intensifier, and the exclamation marks, all of these boost the sentiment score. So, VADER confidently labels it as strongly positive.

Example 2: "I thought it would help, but honestly it didn't do anything" → Neutral

This one gets trickier for VADER. VADER doesn't find any negative keywords, so it calls it neutral, even though, to a human, the disappointment is obvious.

That's the limitation of VADER. Since VADER follows fixed lexical rules and can miss emotional tone or nuance when explicit strong positive or negative words are missing from the text.

RoBERTa: Transformer-Based Sentiment Model

To improve the results of sentiment analysis, I used **RoBERTa**, a transformer-based model fine-tuned on real-world social media text.

RoBERTa builds on the **transformer** architecture introduced in "Attention is All You Need" (Vaswani et al., 2017), using self-attention to understand the context of words in a sentence. Instead of reading text step-by-step like older models, transformers process the whole input at once, helping them catch subtle meaning and emotional cues. RoBERTa was trained on a massive 160GB of data from books, Wikipedia, and web content, and performed extremely well on standard benchmarks like SQuAD (F1: 94.6) and SST-2 (F1: 96.4) (Liu et al., 2019). Since the original RoBERTa wasn't trained specifically for sentiment tasks, I used a fine-tuned version (cardiffnlp/twitter-roberta-base-sentiment) that is designed for analysing real-world social media content more accurately.

The following example demonstrates RoBERTa's ability to detect implied sentiment:

Example 1: "I thought it would help, but honestly.. it didn't do anything"

Unlike VADER, RoBERTa doesn't treat words in isolation, and it understands how a word's meaning changes depending on which words are around it. Hence, it is perfectly able to capture the hidden disappointment in the sentence, even though there aren't any negative words.

While RoBERTa proved more reliable for longer comments and subtle tone shifts, it required significant computational resources. To address this, I migrated the pipeline to Google Colab with GPU support and implemented batch inference for scalability.

5.2.4 Topic Modelling

Online discussions about supplements are often unstructured, making it difficult to find clear patterns or themes. Topic modelling helps identify clear themes from such messy data. I explored the deeper themes in supplement conversations by performing Topic Modeling. Topic modelling is an approach used to identify patterns and topics. The goal was to uncover what people discuss and what they ask.

Why CorEx instead of LDA?

While Latent Dirichlet Allocation (LDA) is a standard technique in topic modeling, it often produces vague or overlapping topics when applied to noisy social media data. I chose Correlation Explanation (CorEx) instead, a model that **maximizes mutual information to discover coherent word clusters**. Unlike LDA's generative approach, CorEx allows the use of anchor words, making it particularly effective for short, casual text like Reddit comments.

This choice is supported by Rizvi et al. (2019), who applied CorEx to over 16,000 Yahoo! Answers posts and found it produced more accurate and interpretable topics. Their results closely matched human-labeled categories such as stomach issues, energy, and sleep, themes also present in my dataset.

How It Was Applied on Data:

I applied topic modeling separately to two text fields:

- Post Titles, which often capture user intent in concise form.
- Comment Bodies, which contain richer detail, longer narratives, and reactions

By modeling both, I aimed to capture both high-level user search intent (from titles) and contextual discussion themes (from comments). Using CorEx from the corextopic library, I experimented with different topic counts (e.g., 10, 15, 20) and reviewed the top keywords manually for coherence.

Finding Supplement Associations per Topic

To connect specific health concerns with relevant supplements, I extracted supplement mentions using a phrase-matching technique, grouped comments by their assigned CorEx topic, and calculated the frequency of each supplement within those groups.

To do this, I:

1. Used a phrase-matching technique to extract supplement mentions from each comment
2. Grouped comments by assigned CorEx topic
3. Counted supplement frequency within each topic

I wanted to know which supplements are associated with a topic. I counted how often a supplement appeared in each topic group. The count showed a topic-to-supplement connection. This helped me find the supplement people mentioned most in each theme. For example, magnesium came up most often when people talked about sleep and muscle cramps. Ashwagandha is linked most often with conversations about worry and stress.

Why This Matters

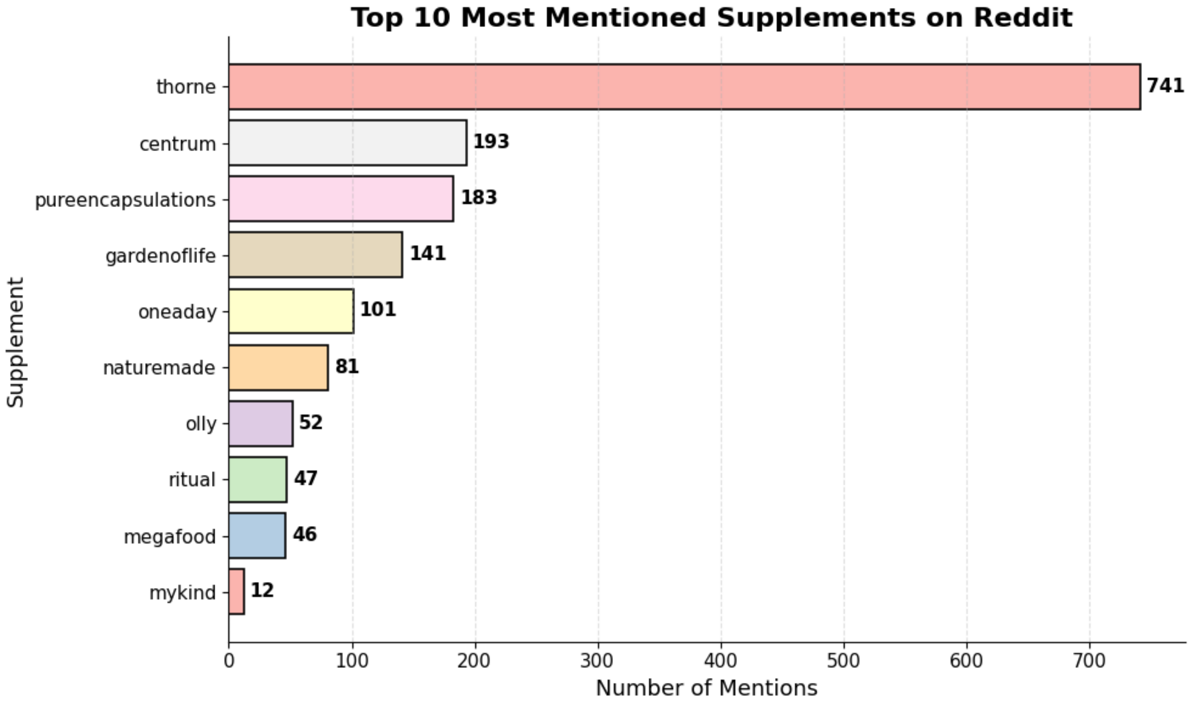
This approach not only revealed the most common wellness goals, such as sleep, digestion, and energy, but also highlighted which supplements users associate with those needs. By aligning user intent with specific products, this analysis provides a foundation for goal-based supplement recommendations. These insights can inform personalized recommendation systems, improve health information platforms, and guide future research on consumer supplement behaviour.

6. Results

6.1 Most Mentioned Supplements

I started the analysis by identifying the top 10 most mentioned supplements on Reddit. I applied a phrase-matching method to the pre-processed, lemmatized comments to identify the supplements most frequently discussed by users.

Figure 5: *Top 10 most mentioned supplements on Reddit*

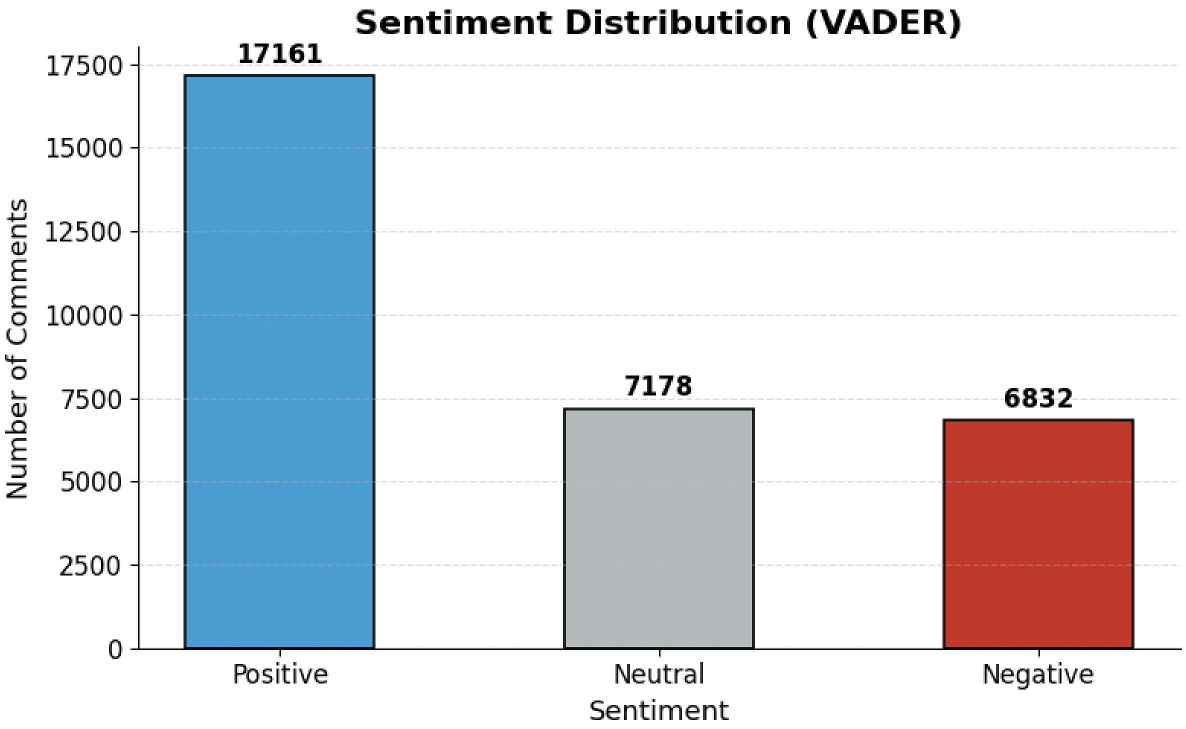


Thorne stands out as the most discussed brand by a huge margin - it receives 741 mentions. This number is almost four times more than the brand in second place, which shows Throne is very popular, and perhaps many people on Reddit trust the brand. Centrum, followed by Pure Encapsulations, comes next. Centrum receives 193 mentions, and Pure Encapsulations receives 183. Centrum is easy to find in most stores. Pure Encapsulations has a name for good ingredients. These factors may explain their frequent mentions. The list contains both well-known brands and niche ones, which indicates that users have varied user preferences and awareness.

6.2 Sentiment Analysis Results

6.2.1 VADER Sentiment Distribution

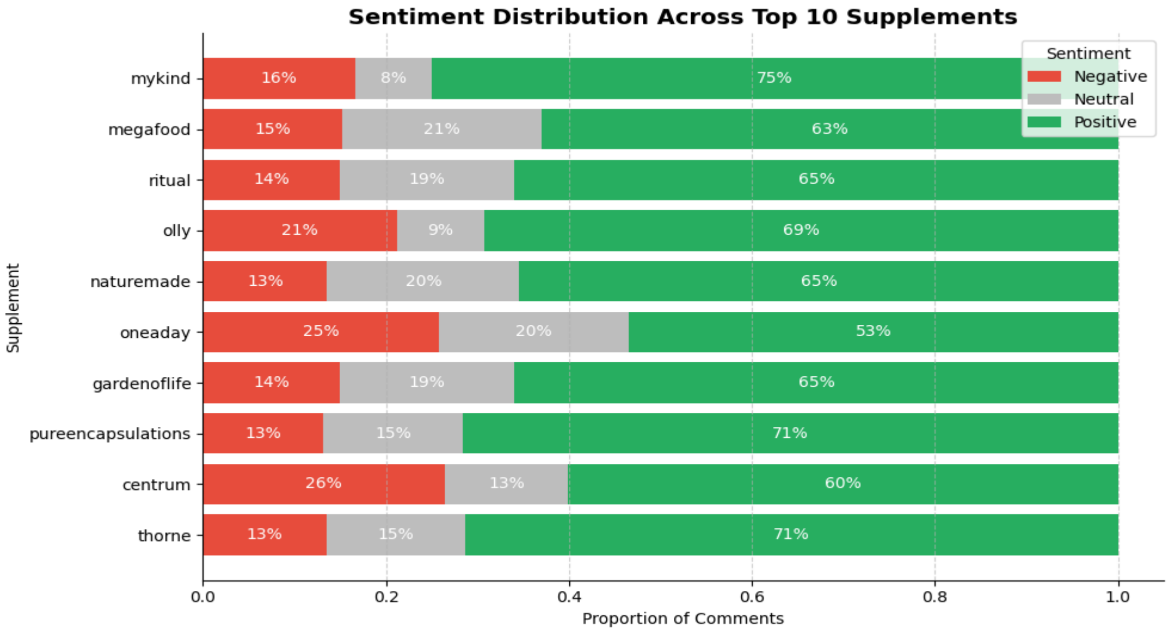
Figure 6: *Overall Sentiment Distribution of Supplement Discussions (VADER Model)*



VADER identifies more than half of the comments as positive: **17,161** in total. It classified 7,178 comments as neutral and 6,832 as negative. This overwhelmingly positive rating suggests that VADER emphasizes certain positive keywords due to its purely rule-based design.

Sentiment by Supplement (VADER)

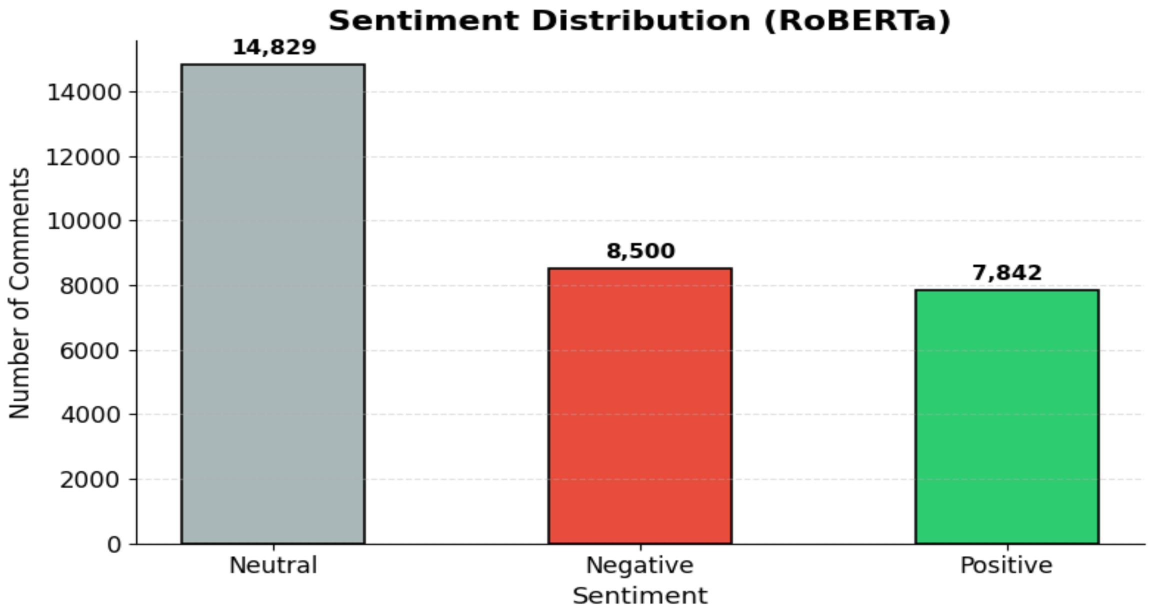
Figure 7: *Sentiment Distribution Across the Top 10 Most Mentioned Supplements (VADER)*



The VADER sentiment distribution across the top 10 major supplements shows a tilt towards positivity. Brands such as **MyKind, Thorne, and Pure Encapsulations** get more than 70% positive sentiment. Even widely popular brands like **Centrum** and **One A Day**, despite having higher negative proportions (26% and 25%), still receive a majority of favourable responses. This suggests that VADER tends to amplify sentiment polarity, and it made me realize that a context-aware tool should be used for a deeper sentiment analysis.

6.2.2 RoBERTa Sentiment Distribution

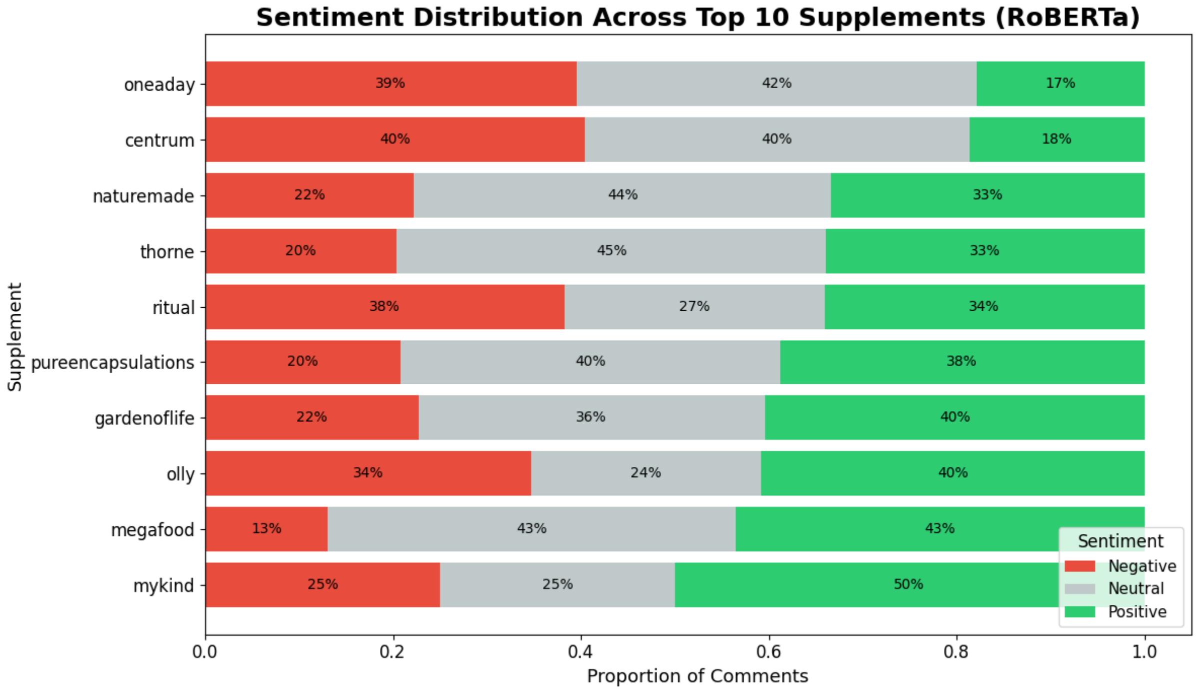
Figure 8: *Overall Sentiment Distribution Using RoBERTa*



RoBERTa, a transformer-based model trained on real-world text, classified most comments as **neutral** (14,829), followed by negative (8,500) and positive (7,842). Unlike VADER, RoBERTa doesn't rely solely on emotional keywords. Instead, it uses contextual embeddings and attention mechanisms to detect hidden dissatisfaction or conflicted opinions, even in polite or indirect language. Hence, I was able to get more intuitive and real results.

Sentiment by Supplement (RoBERTa)

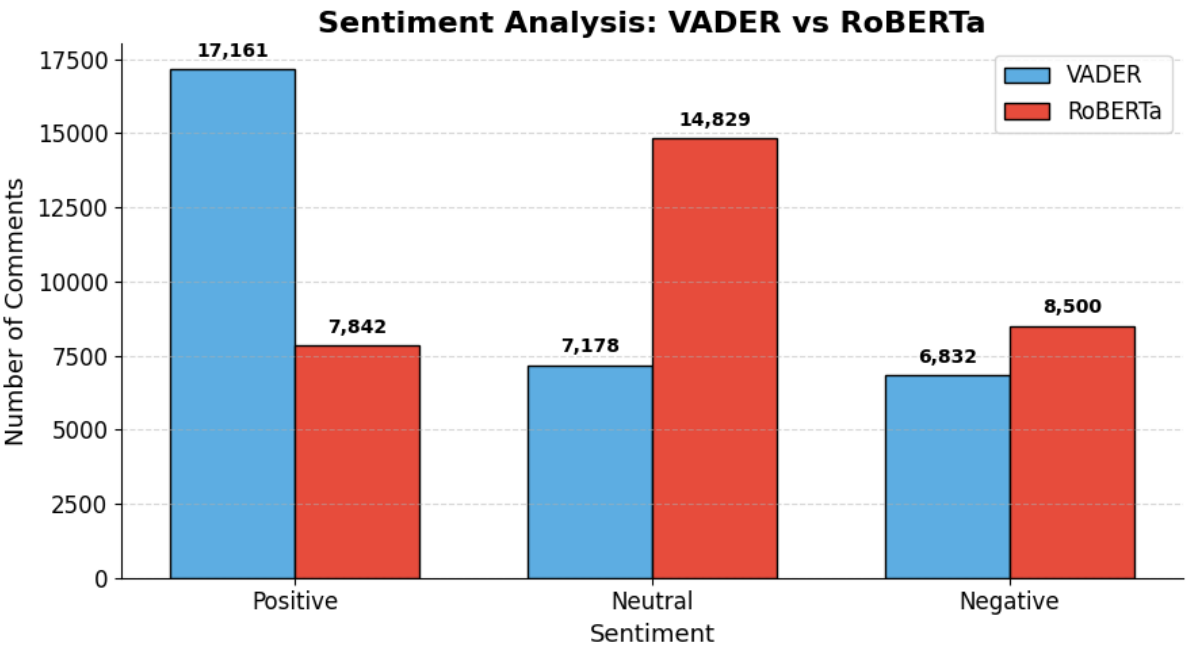
Figure 9: *Supplement-Wise Sentiment Proportions Using RoBERTa*



This RoBERTa-based sentiment analysis shows a more nuanced view of supplement discussions. Supplements like MyKind and MegaFood had higher proportions of positive sentiment (50% and 43%, respectively). In contrast, Centrum and One A Day showed more neutral or negative sentiment, with Centrum having 40% neutral and One A Day 39% negative. Compared to VADER, this distribution suggests RoBERTa captures mixed feelings and subtle dissatisfaction more effectively.

Sentiment Analysis Comparison: VADER vs RoBERTa

Figure 10: *Comparison of Sentiment Classification Between VADER and RoBERTa*

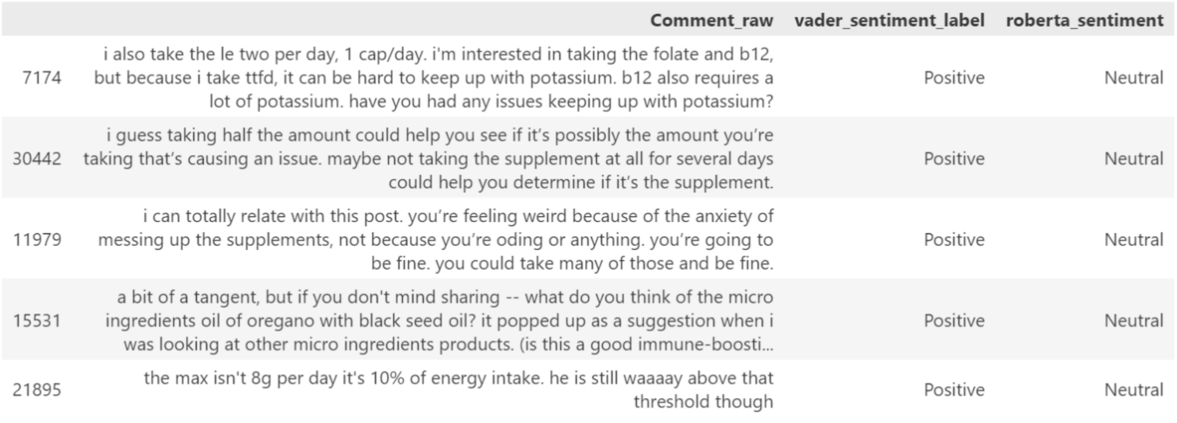


This further highlights the contrasting sentiment interpretations of the two models. VADER classified a majority of comments as positive (17,161), while RoBERTa identified significantly more neutral (14,829) and negative (8,500) comments.

When Models Disagree: Understanding Hidden Tone

I extracted comments from the dataset, I noticed that VADER often labeled many comments as "Positive" simply because they contained polite or technical language, even when the actual tone was neutral, confused, or hesitant. RoBERTa, on the other hand, as expected, proved more capable of detecting subtle emotional cues and implied uncertainty.

Figure 11: *Sample Comments with Divergent Sentiment Labels (VADER vs. RoBERTa)*



	Comment_raw	vader_sentiment_label	roberta_sentiment
7174	i also take the le two per day, 1 cap/day. i'm interested in taking the folate and b12, but because i take ttfd, it can be hard to keep up with potassium. b12 also requires a lot of potassium. have you had any issues keeping up with potassium?	Positive	Neutral
30442	i guess taking half the amount could help you see if it's possibly the amount you're taking that's causing an issue. maybe not taking the supplement at all for several days could help you determine if it's the supplement.	Positive	Neutral
11979	i can totally relate with this post. you're feeling weird because of the anxiety of messing up the supplements, not because you're oding or anything. you're going to be fine. you could take many of those and be fine.	Positive	Neutral
15531	a bit of a tangent, but if you don't mind sharing -- what do you think of the micro ingredients oil of oregano with black seed oil? it popped up as a suggestion when i was looking at other micro ingredients products. (is this a good immune-boosti...	Positive	Neutral
21895	the max isn't 8g per day it's 10% of energy intake. he is still waaaay above that threshold though	Positive	Neutral

These cases made it clear why relying only on keyword-based sentiment tools can lead to misleading conclusions, especially in a domain like health, where emotions are often implied or subtle. Using RoBERTa allowed for a more honest understanding of how people feel when they talk about supplements.

6.3 Topic Modeling

6.3.1 Extracting Themes from User Comments

I used the Correlation Explanation (CorEx) topic modeling algorithm to extract meaningful and interpretable themes from Reddit comments. I defined **7 anchored topics** based on recurring health-related concerns found in user discussions. I configured the model to extract **10 topics** in total: 7 anchored and 3 additional discovery topics inferred from the data. I set the anchor strength to 3 to balance the influence from the anchors and the data.

The resulting topics effectively captured distinct user concerns. Below is a summary of the top terms from each topic:

Table 2 *Anchored Topics from Reddit Comments with Anchor Terms, Labels, and Top Words*

Topic #	Anchors Used	Topic Name (Self-Given)	Top Words
1	fatigue, tired, energy, exhausted	Fatigue / Energy	energy, fat, weight, calorie, body, diet, need, carb, tired, muscle
2	sleep, melatonin, insomnia, rest	Sleep	sleep, rest, day, mg, magnesium, insomnia, glycinate, time, week, melatonin
3	anxiety, stress, calm, relax	Anxiety / Stress Relief	anxiety, stress, vitamin, level, cause, help, calm, blood, effect, dose
4	focus, adhd, clarity, attention	Cognitive / Focus	focus, people, health, study, adhd, risk, say, nutrition, attention, benefit
5	gut, digestion, bloating, probiotic	Gut Health	gut, probiotic, digestion, food, plant, process, source, base, nutrient, animal
6	acne, skin, glow, clear	Skin / Beauty	skin, clear, thing, year, acne, think, bad, actually, change, right
7	immunity, cold, sick, flu	Immunity	oil, fish, sick, cold, come, big, probably, small, olive, make
8	None	Nutrition / Food	eat, fruit, protein, sugar, meat, chicken, veggie, bean, healthy, vegetable
9	None	Product Feedback	supplement, product, brand, company, research, Thorne, know, look, quality, test
10	None	Purchase Decision	like, good, high, lot, want, add, way, use, cheap, easy

6.3.2 Extracting Themes from Post Titles

To better understand **what users are asking or seeking** in supplement-related discussions, I applied topic modeling on the **titles alone**, without any predefined anchors. Unlike the comment-based model, this approach was entirely **discovery-driven**, allowing latent themes to emerge naturally from the data. The goal here was to surface **common question themes or informational needs** expressed by users. Below is a summary of the 8 identified title topics, along with their most representative words and an interpretation based on manual review.

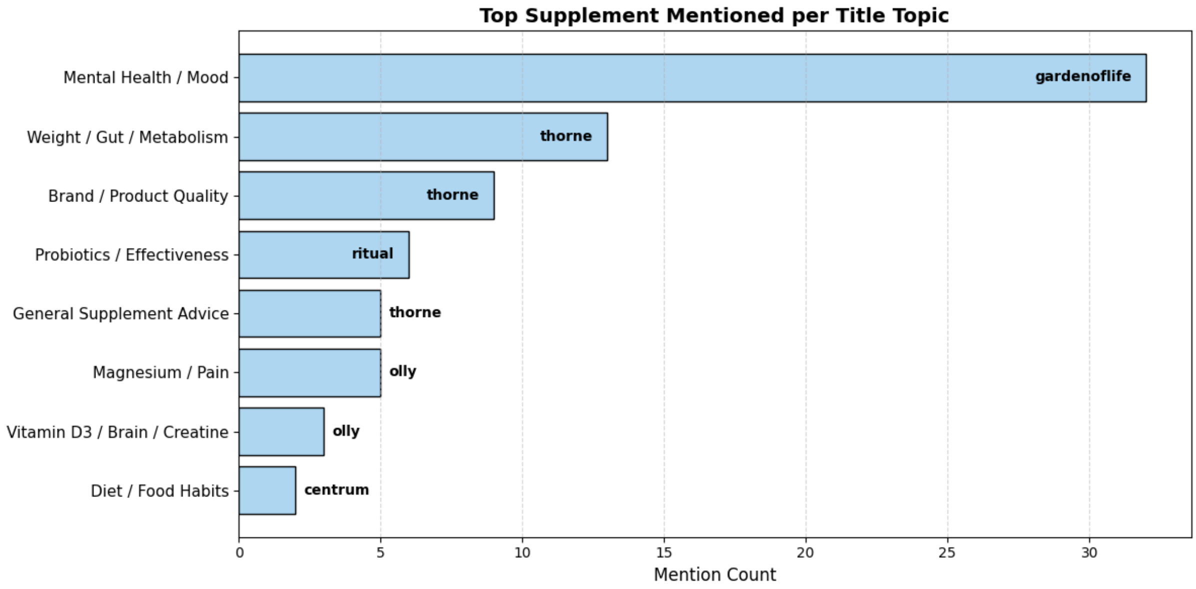
Table 3: *Topics from Reddit Titles, Labels, and Top Words*

Topic #	Top Words	Topic Name (Self-Given)
1	depression, try, anxiety, male, reaction, week, old, stress, goodbye, caffeine	Mental Health & Personal Experiences
2	pure, brand, encapsulation, extract, calcium, purchase, doctor, bottle, flavor, like	Product Quality & Purchase Decisions
3	d3, iu, powder, k2, heart, blood, protein, creatine, brain, pressure	Vitamin D, Heart Health & Fitness
4	weight, gain, loss, naturemade, start, berberine, constipation, focus, lose, quality	Weight Management & Constipation
5	eat, day, meal, food, calorie, fat, body, vegetable, process, water	Diet, Meal Timing & Nutrition
6	supplement, multivitamin, help, vs, man, advice, gardenoflife, vitamin, need, centrum	Supplement Recommendations & Comparisons
7	magnesium, oil, glycinate, fish, cause, stop, pain, disease, stomach, citrate	Magnesium & Pain Relief
8	look, probiotic, think, work, pill, source, right, ingredient, change, ritual	Evaluation of Supplements / Ingredients

These topics provide a **high-level overview of user concerns** when they initiate discussions, ranging from mental health and fitness goals to specific nutrient evaluations and product-related questions.

6.3.3 Brand Mentions Across Title Topics

Figure 12: *Most Mentioned Supplement per Topic Identified in Post Titles*

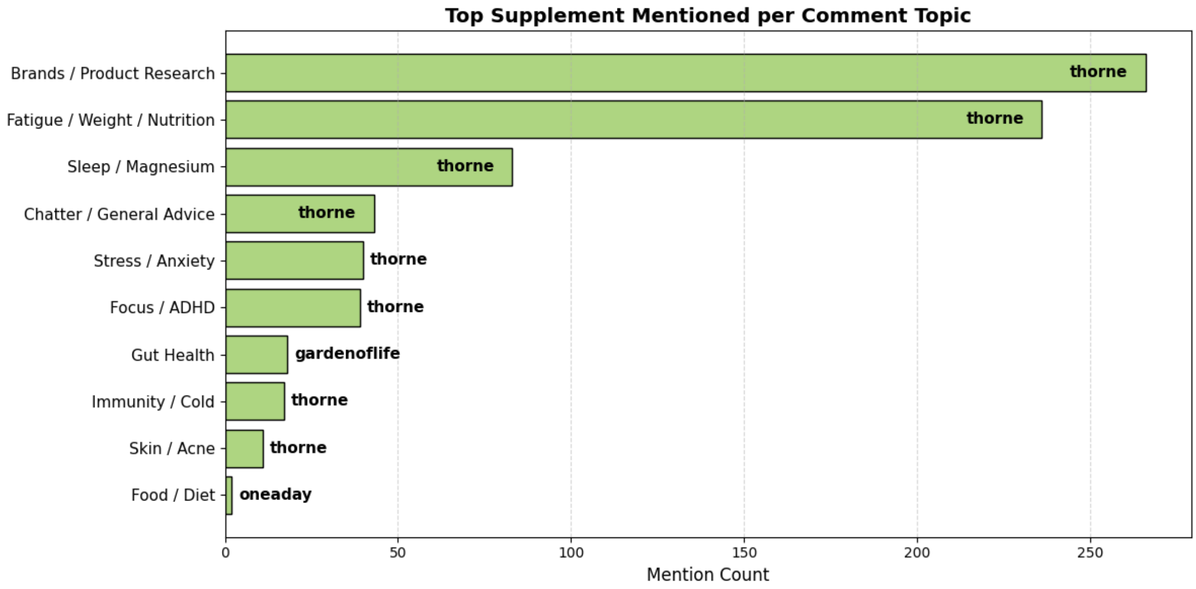


I analyzed brand mentions from Reddit post titles grouped by topic to understand which supplement brands users are actively asking about. I used a predefined list of known supplement brands and identified which brand appeared most frequently in each topic.

The results show that Garden of Life is frequently mentioned in discussions related to mental health. Thorne appears often in several topics, such as metabolism and product quality. Although the overall counts are low, these patterns reflect brand-specific curiosity and user intent at the point of initiating discussion.

6.3.4 Brand Mentions Across Comment Topics

Figure 13: *Most Mentioned Supplement per Topic Identified in Post Titles*



I also analyzed supplement brands frequently mentioned within Reddit comments. I grouped comments by their assigned topic labels. The analysis shows that **Thorne** is by far the most talked about brand - it comes up a lot in discussions about Product Research, Fatigue / Nutrition along with Sleep / Magnesium, which means people prefer Thorne and talk about it often, both for general questions and for particular health problems. Other brands, such as Garden of Life and One A Day, appeared in selected topic categories. Garden of Life was frequently mentioned in gut health discussions. One A Day was associated with food and diet conversations, though it was mentioned far less frequently. Overall, comment-level analysis highlights how a few trusted brands consistently recur in both problem-solving and product comparison conversations.

What Users Ask vs. What They Recommend

Thorne emerged as the most consistently mentioned supplement across nearly all comment topics, suggesting strong user trust and brand familiarity in real conversations. While **Garden of Life** was frequently mentioned in titles (indicating curiosity), it appeared far less in comments, highlighting a gap between interest and actual engagement. This contrast suggests users may explore multiple brands initially but return to a few trusted names when sharing experiences or advice. Overall, comment-based mentions offer a more accurate reflection of **what users rely on and talk about**, not just what they search for.

7. Conclusion

This project aimed to use Natural Language Processing to understand people's experiences with multivitamin supplements based on real Reddit discussions. The analysis reveals a key insight: most of the users are uncertain and feel conflicted while picking up the supplements, and most people struggle to find out what is best for them.

Sentiment analysis proved useful for identifying emotional responses toward both general supplements and specific brands. While VADER captured general sentiment patterns, RoBERTa detected more subtle dissatisfaction often missed by rule-based models.

Topic modeling revealed that users frequently discuss issues like sleep, anxiety, digestion, and focus, indicating that supplement choices are closely tied to individual wellness goals. Overall, the study demonstrated that unstructured online conversations can be transformed into structured insights, offering an understanding of user needs and experiences in the supplement space.

Limitations:

Despite its insights, this study has a few limitations, which must be acknowledged:

1. **Lack of Labeled Data**: The Reddit dataset is messy and unlabeled, which makes it difficult to apply analysis methods to extract insights and evaluate accuracy.
2. **Scraping Restrictions**: Additional data like ratings or verified reviews from platforms such as Amazon or iHerb could have enriched the analysis, but web scraping limitations prevented access to these sources.
3. **No Ground Truth for Outcomes**: User claims about supplement effectiveness are subjective and unverifiable, which limits the ability to draw concrete conclusions about health benefits.
4. **Bias in User Base**: Reddit users tend to be younger, more tech-savvy, and predominantly from Western countries. As a result, the findings may not generalize well to broader or more diverse populations.

Outlook:

The techniques applied in this project, such as topic modeling, sentiment analysis, and entity extraction, can be extended to clinically relevant platforms such as PatientsLikeMe or WebMD forums. These platforms offer more structured, health-focused discussions, where similar NLP pipelines could extract patient-reported outcomes, treatment satisfaction, and symptom trajectories. Such applications hold potential for supporting public health research, digital therapeutics, and personalized health recommendations based on real-world user narratives.

References:

Danish, M., & Amjad, M. (2023). *Empowering recommendations with NLP: Exploiting textual reviews for enhanced rating-based systems.* International Journal on Recent and Innovation Trends in Computing and Communication, 11(10s). https://doi.org/10.17762/ijritcc.v11i10s.7596

Hutto, C. J., & Gilbert, E. (2014). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.* Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM-14), 216–225.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach.* arXiv preprint arXiv:1907.11692.

Mehta, A., Singh, R., & Sharma, P. (2021). *Review Mining and Sentiment Analysis on Dietary Supplements.* International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 7(2), 33–39.

NIH. (2022). *Dietary Supplements: What You Need to Know. National Institutes of Health.* Retrieved from https://ods.od.nih.gov/factsheets/WYNTK-Consumer/

Okon, E. E., Ibharalu, F. T., & Olaleye, A. (2020). *Sentiment and topic modelling analysis of tweets on the COVID-19 pandemic using NLP techniques.* International Journal of Advanced Computer Science and Applications, 11(12), 336–345.

Palanisamy, S., & Jha, R. (2023). *Social media analytics for wellness and health product marketing: An NLP-based approach.* Journal of Intelligent & Fuzzy Systems, 44(5), 6231–6241.

Rizvi, M. A., Sarraf, S., Rho, Y., & Nguyen, D. (2019). *CorEx for analyzing consumer health questions: Improving topic coherence using anchor words.* IEEE International Conference on Healthcare Informatics (ICHI), 1–8. https://doi.org/10.1109/ICHI.2019.8904619

Rosamma, K. S. (2024*). Analyzing online conversations on Reddit: A study of stress and anxiety through topic modeling and sentiment analysis.* Cureus, 16(2). https://doi.org/10.7759/cureus.53788

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). *Attention is all you need*. Advances in Neural Information Processing Systems, 30, 5998–6008.

Appendix

Code file and datasets

The complete implementation of this project work, including web scraping, data preprocessing, sentiment analysis, topic modelling, along with the cleaned and processed datasets, can be accessed at my GitHub repository: https://github.com/mitalin92/TANLP_supplement_analysis