Continuous Glucose Monitoring - Social Media Analysis

Data Science for Product Manager
Team Members: Mukunda Aithal, Cindy Huang, Luyu Huang, Kelly Wang, Victor Wang

Background

Continuous glucose monitor (CGM) device is a wearable medical technology that helps people to track blood sugar level. Traditionally, people take fingerstick tests to measure their glucose level and the introduction of CGM since 1999 has revolutionized how people manage diabetes. In the span of 20 years, CGM has become an increasingly common medical technology that both hospitals and people at home use to gain insights into their diabetic conditions. However, there are still more than half of the diabetics, for both Type 1 and Type diabetes, who are not adopting the technology and that represents opportunity for commercial organizations that are in the industry and healthcare policy makers.

The 2020 Covid-19 pandemic has a huge impact on the CGM market as people are staying home more, the increased demand for remote monitoring of patients resulted in a 18% increase in the market from 2019 to 2020. We will be looking at social media data collected between 2021 and 2022, which will help us to gain insights on how the CGM industry looks after its recent expansion.

Problem Statement

We constructed two questions that we want to answer with our analysis. First, we want to understand how customers are viewing CGM in general. Then, we want to identify opportunities for two specific brands, Dexcom and FreeStyle Libre, that would help differentiate their products and increase their market share. The analysis results will provide useful insights to CGM companies, especially Dexcom and Abbott, and public health programs such as Medicare to make informed decisions regarding resource allocation.

Project Approach

We will first conduct an Exploratory Data Analysis (EDA) to understand the dataset we are looking at and prepare the data for further analysis. Knowing the information of the dataset such as post source, will help us to be cognizant of the potential biases that exist in the data as we review the analysis results. After EDA, we will attempt to answer our initial question “what do customers think of CGM” by extracting out the main topics people discussed on social media, with different clustering techniques. Finally, we will delve into specific product series from two major players in the industry, Dexcom and Abbott, and find what people like and dislike about the product.

Analysis and Insights by Section

Exploratory Data Analysis

In the EDA section, we conducted analysis for the following steps:

1. Inspect columns
2. Check missing values
3. Summary statistics
4. Check outliers

We have the following high-level observations:

1. Low to no data availability across the majority of columns: Among the 63 features, 24 features have only missing values, 18 features have 95%-100% values missing.
2. A main reason why many columns display low data availability is that the dataset contains posts from multiple sources with different metrics and information collected. About 92% of the posts were sourced from forums. About 89% were from Reddit.
3. All of the numerical columns (mostly related to post popularity) observe long-tailed distributions. In our context the outlier posts may require the most attention of the business stakeholders, as these are posts that may be representative or have bigger reach. We do not drop these outliers and instead seek to understand the themes of the sample texts. Sampled texts show that some themes of the trending posts include insurance and affordability, lifestyle influencing, and app product issues.
4. Social media discussion volumes have an upward trend over time and observe spikes in the summer months.

General Inspection

Our dataset contains 37.844 post-level data ranging from March 2021 to September 2022. In the dataset, we have the information regarding the following features (the features with >90% missing values are marked in *italic*)

- **Post content**: 'Sound Bite Text', 'Title', 'Richness', *'@Mention Media Tags'*
- **Post sentiment**: 'Sentiment', 'Positive Objects', 'Negative Objects', *'Ratings and Scores'*
- **Post popularity**: 'Followers/Daily Unique Visitors/Subscribers', 'Total Engagements', 'Post Comments', 'Post Likes', *'Post Shares', 'Post Views'*, *'Post Dislikes'*
- **Post metadata**: 'Post ID', 'Source Type', 'Post Type', 'Is Paid', 'Media Type', 'URL', 'Media Link', 'Domain', 'Published Date (GMT-04:00) New York', 'Reddit Score', 'Source Name', 'Rating', *'Tags'*, *'Product Name', 'Product Hierarchy'*
- **Repost information**: 'Quoted Post', 'Quoted Author Name', 'Quoted Author Handle'
- **Author account and demographics information**: 'Author Gender', 'Author URL', 'Author Name', 'Author Handle', 'Author ID', 'Author Location - Country 1', 'Author Location - State/Province 1', 'Author Location - City 1', 'Author Location - Country 2', 'Author Location - State/Province 2', 'Author Location - City 2', 'Author Reddit Karma', 'Professions', 'Interests', *'Author Location - Other'*
- ***LexisNexis information**: 'LexisNexis Source Publisher', 'LexisNexis Source Category', 'LexisNexis Source Genre', 'LexisNexis Source Quality', 'LexisNexis Company - High',*

'LexisNexis Company - Any', 'LexisNexis Person - High', 'LexisNexis Person - Any', 'LexisNexis Institution - High', 'LexisNexis Institution - Any', 'LexisNexis Subject Group 1', 'LexisNexis Subject 1', 'LexisNexis Subject Group 2', 'LexisNexis Subject 2', 'LexisNexis Other Subjects'

Check Missing Values

We observed that missing value is a prevalent issue with features in this dataset.

- Among the 63 features, 24 features have only missing values, 18 features have 95%-100% values missing. All LexisNexis related columns have only missing values.
- The features with low data availability tend to be source specific and can be informative. Below shows a summary table of the available features by source.

Source	**Author**	**Popularity Metrics**	**Repost**
Blogs	Gender, Name, Handle, Location (Country level)	N/A	N/A
Comments	ID	Followers/Daily Unique Visitors/Subscribers	N/A
Facebook	Location	Engagement, Comment, Like	N/A
Forums	Gender, Name, Handle, Reddit Karma	Followers/Daily Unique Visitors/Subscribers	N/A
Instagram	Location (Country level)	Engagement, Comment, Like	N/A
Twitter	Gender, Name, Handle, Location, Interests	Followers/Daily Unique Visitors/Subscribers	Available

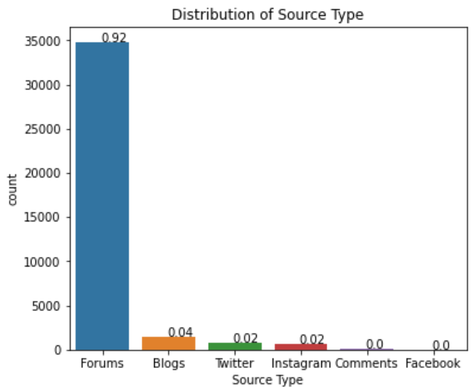
Summary Statistics

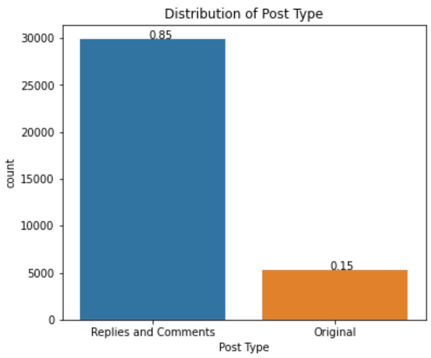
Among the categorical features, we observe some data imbalance from multiple dimensions.

- Most of the posts come from forums, specifically reddit. The discussions on forums occur the most in diabetes-related forums. Facebook contributes the fewest posts to this dataset.
- We can reasonably suspect that our dataset contains mostly Type 1 diabetic CGM users. According to our research, reddit’s age group is between 18 and 29 (Statista),

and Type 1 diabetes are mainly in youth and Type 2 diabetes are generally developed in people over age 45. This speculation also aligns with the known CGM user demographic, CGM has a higher penetration in Type 1 diabetic patients than Type 2 diabetic patients.

- Most of the posts are replies and comments, which may be because most forum posts are replies and comments.
- None of the posts is marked as paid, so the “Is Paid” feature is not very informative.
- Most of the posts are of neutral sentiment. There are more positive posts than negative ones, which may be a good sign that the products are perceived more as positive than negative.

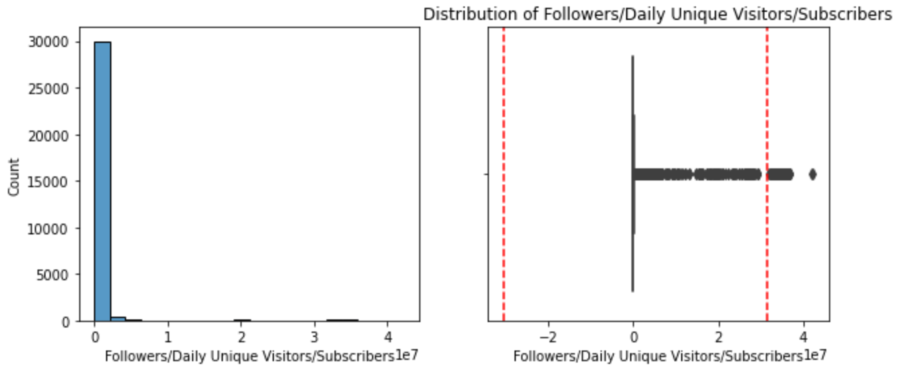




All the numerical columns (mostly related to post popularity) observe long-tailed distributions.

- We will be interested in what the few posts with very high popularity say, as they may represent the opinions of a large number of social media users.
- Based on the sampled texts, we see that some of the trending posts focus on insurance and affordability, lifestyle influencing, product issues with non-smartphone mobile devices.

Example analysis: Followers/Daily Unique Visitors/Subscribers (cut-off = 10 standard deviations away from the mean)



Post ID	Title	Sound Bite Text	Source Type
BRDR	How do	It depends on what insurance I have at that	Forums

DT2-t1_impvxki	Americans manage to pay for their medical bills? What are some tips for making it easier?	time. Last year we had United healthcare and I had to pay my $6,000 deductible before they would cover anything and then just for my CGM and insulin pump supplies it was about $4,000 every 3 months. We have Blue Cross now and they pay a little more but it's hard to get the supplies in a timely manner. Just went without some supplies for a month. But then again United healthcare would pay for my eye injections so I don't go blind but Blue Cross doesn't want to pay for that.	
BRDR DT2-t1_hbn6fu0	What can kill you but you are doing it everyday?	Being a diabetic in the US is ridiculous. I have a really good insurance plan. So while it doesn't cost me a whole lot for the insulin I get highly screwed on the supplies for my pump and cgm. It costs me about 650$ a month to be diabetic and run the system that can maintain my sugar levels. It's a joke really.	Forums

Of the only date feature, we see the following trends:

- There is a general upward trend of the number of posts in each month. This shows that the topic is gaining traction and discussions.
- We see a particularly increased number of posts during the summer months (July and August). This may be potentially due to people’s attention to lifestyle and diet, as they may go out and consume ice cream more.

Check Outliers

As mentioned in the “Summary Statistics” section, in our context the outlier posts may require the most attention of the business stakeholders, as these are posts that may be representative or have bigger reach. We do not remove these outliers and instead check sample texts of the outliers in the “Summary Statistics” section to understand their themes.

General CGM Analysis

After understanding the data through exploratory data analysis, we decided to see if performing segmentation on the text data would provide any actionable insights. Multiple methods were considered to cluster the textual data but, in the end, it was decided to try K-Means clustering and Latent Dirichlet Allocation (LDA). K-Means clustering was chosen as a baseline model since it is relatively simple, making strong assumptions about the clusters being non-overlapping and spherical, and fast. LDA was chosen as it is normally used for topic modeling in text documents and is based on probabilistic vectors of words, which indicate their relevance to the text corpus. It also allows for one data point to belong to multiple topics, which is more relevant for the data we are working with.

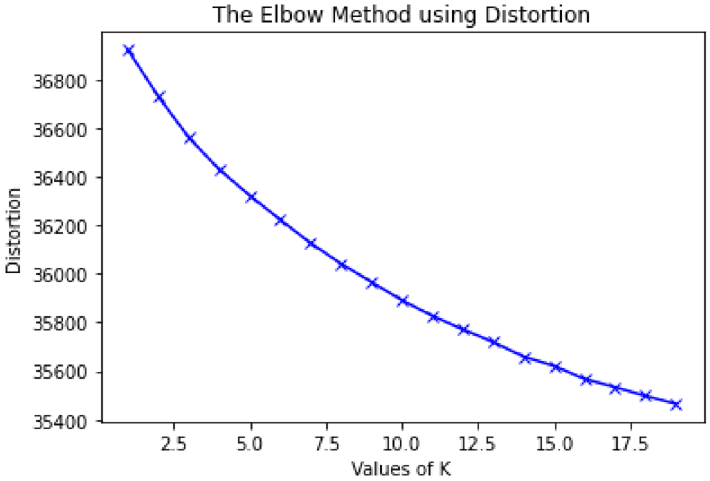
K-Means Clustering

Data Preparation

Preceding model learning, the sound bite text and titles were concatenated together since we believed the titles were just as important when clustering. Next, these strings were tokenized and any URLs were removed. Finally, the text data was lemmatized and vectorized using TF-IDF (Term Frequency Inverse Document Frequency) to transform the text into a meaningful binary vector representation.

Cluster Selection

In order to choose the number of clusters (value of variable k), the elbow method and the silhouette method were checked. Unfortunately, these methods did not provide a definitive answer. The elbow graph below shows no real visible and distinct elbow where the distortion suddenly plateaus. The argument could be made that the curve starts to flatten around the value of 8 but it was deemed that this would be too many clusters to interpret meaningfully.



Similarly, the silhouette score values for every value of k between 2 and 10 did not provide further insight. The list of scores in the table below show that the scores kept rising until 10, indicating that data points were still not as well separated as they could be and many points were still on decision boundaries.

K (# of clusters)	2	3	4	5	6	7	8	9	10
Silhouette Coefficient	0.00421	0.00472	0.00441	0.00494	0.00544	0.00544	0.00565	0.00589	0.00579

Therefore, since the LDA analysis was run in parallel, the number of clusters identified during the LDA topic selection process was used for K-Means clustering. The LDA process will be discussed later, but for now, the number of clusters was determined to be four (k = 4).

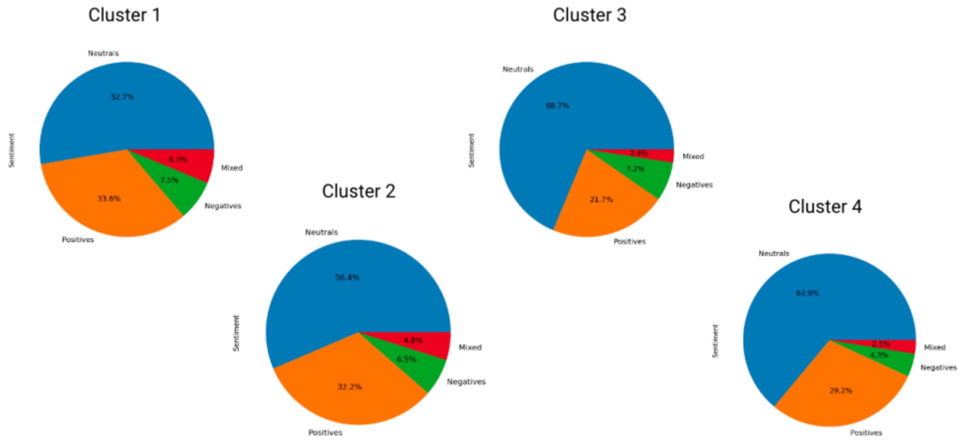
Clustering Results and Analysis

After K-Means clustering was performed, the top terms per cluster were identified to obtain an intuition of the most influential words for each cluster. The table below shows the top ten influential terms per cluster, the count of data assigned to each cluster, and our interpretation of the topic each cluster represents. The underlined terms in each cluster highlight that these terms were unique to that cluster, and therefore more importance was given to these terms when trying to understand the underlying topic for each cluster.

Cluster Number	Count	Top 10 Terms	General Topic
1	4581	libre, freestyle, sensor, dexcom, use, day, app, cgm, like, phone	CGM monitors
2	5197	pump, medtronic, dexcom, tandem, cgm, year, insulin, omnipod, tslim, slim	Insulin pumps
3	21337	dexcom, cgm, low, like, time, day, sensor, insulin, year, use,	Unsure
4	6729	blood, sugar, glucose, monitor, continuous, cgm, low, insulin, eat, high	Personal habits?

Clusters 1, 2, and 4 seem to be more distinct when looking at their top defining terms. Cluster 1 has mostly CGM specific terms while Cluster 2 has several insulin pump terms and brands. On the other hand, Cluster 4 does not have any CGM brand terms and has more terms related to biology and lifestyle. Unfortunately, the majority of data landed in Cluster 3 which did not seem to have any defining or unique terms.

Next, we wanted to compare the clusters to other categorical features in the data to see if clustering had unknowingly created some distinctions. Specifically, we compared the sentiments to see if any of the clusters had significantly different sentiments. We also compared the forum sources (subreddit sources) to check if certain groups of people aligned with one or more cluster topics. The graphs below show the percentages of sentiment score for each cluster. Unfortunately, this did not provide much insight since all of the cluster seemed to have similar ratios of sentiment scores. The only notable difference was that Cluster 1 and 2 had slightly more positive sentiments compared to Cluster 3 and 4.



Checking the top ten subreddits (based on count) for each cluster was unfortunately not too insightful as well. Unsurprisingly, every cluster contained the majority of their data in subreddits directly related to diabetes. One notable observation is that Cluster 2 does not have type 2 diabtes related subreddits. This may be due to the fact that Cluster 2's topic centers around insulin pumps and most type 2 diabetics do not need insulin pumps. Another interesting subreddit that was listed in the top ten for almost all clusters r/GestationalDiabetes.

Gestational diabetes is diabetes diagnosed for the first time during pregnancy (gestation). This causes high blood sugar that can affect pregnancy and the baby's health but it can be controlled with diet and possibly medication. Generally, gestational diabetes only lasts during the pregnancy and blood sugar returns to its usual level soon after delivery. However, having gestational diabetes indicates a higher risk of getting type 2 diabetes later. [1]

The mention of this subreddit so often may indicate that official/accessible knowledge around CGM usage for gestational diabetes is limited.

Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
r/diabetes 1443 r/diabetes_t1 1016 r/Type1Diabetes 416 r/Freestylelibre 225 r/diabetes_t2 167 r/GestationalDiabetes 105 r/dexcom 88 r/Biohackers 37	r/diabetes_t1 2573 r/Type1Diabetes 1140 r/diabetes 1037 r/dexcom 42 r/AskReddit 24 r/TandemDiabetes 22 r/BumpersWhoBolus 10 r/MadeMeSmile 6	r/diabetes_t1 8529 r/diabetes 5541 r/Type1Diabetes 3718 r/dexcom 504 r/diabetes_t2 117 r/AskReddit 99 r/BumpersWhoBolus 36 r/GestationalDiabetes 35	r/diabetes 1170 r/diabetes_t1 1166 r/Type1Diabetes 613 r/diabetes_t2 200 r/GestationalDiabetes 119 r/keto 114 r/PCOS 79 r/AskReddit 75

r/type2diabetes 32 r/AskReddit 27	r/Medtronic670G 6 r/Omnipod 5	r/senseonics 34 r/Freestylelibre 28	r/AmItheAsshole 58 r/Biohackers 43

Conclusion – K-Means Clustering

Overall, K-Means clustering provides some actionable information, but it is clear to see that the text data can belong to multiple topics. Therefore, LDA would be a better choice for topic modeling rather than K-means clustering.

LDA Topic Modeling

Hyperparameter Tuning: Number of Topics

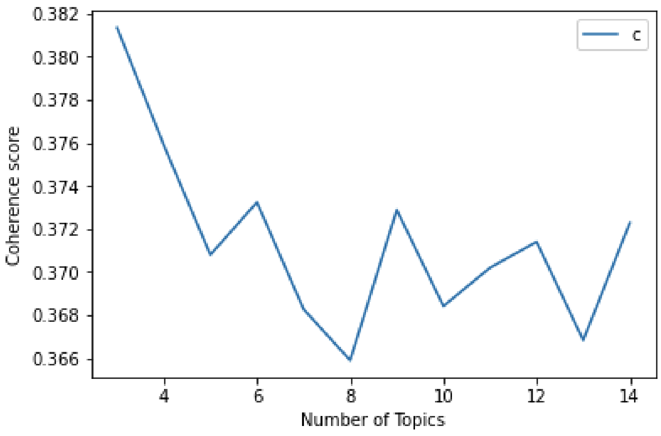
Before jumping in to train our LDA model, we will conduct hyperparameter tuning in order to get the best possible outcome. First, we will test out the number of topics from *3 to 15* using the same lemmatized dataset the K-mean model was fed with and select the one with the highest coherence score. In our case, we believe that 4 topics would generate the best result as it has the highest coherence.

Hyperparameter Tuning: Alpha and Beta

Secondly, we will work on running the best value for the Alpha and Eta variables. In LDA models, alpha represents document-topic density. A higher value of alpha represents that the documents are composed of more topics, whereas a lower alpha represents fewer topics contained in the documents. Beta('Eta' is used in gensim python package) represents topic-word density; A higher value of Beta indicates that a large portion of the corpus we created contributes to topics, and we will get a lower value of Beta if topics are composed of a relatively lower number of words from the corpus. Therefore, intuitively the ranges of Alpha and Beta are from *0(no inclusive) to 1*. Given the limited computational resources we have, the steps we use for tuning is *0.2.*

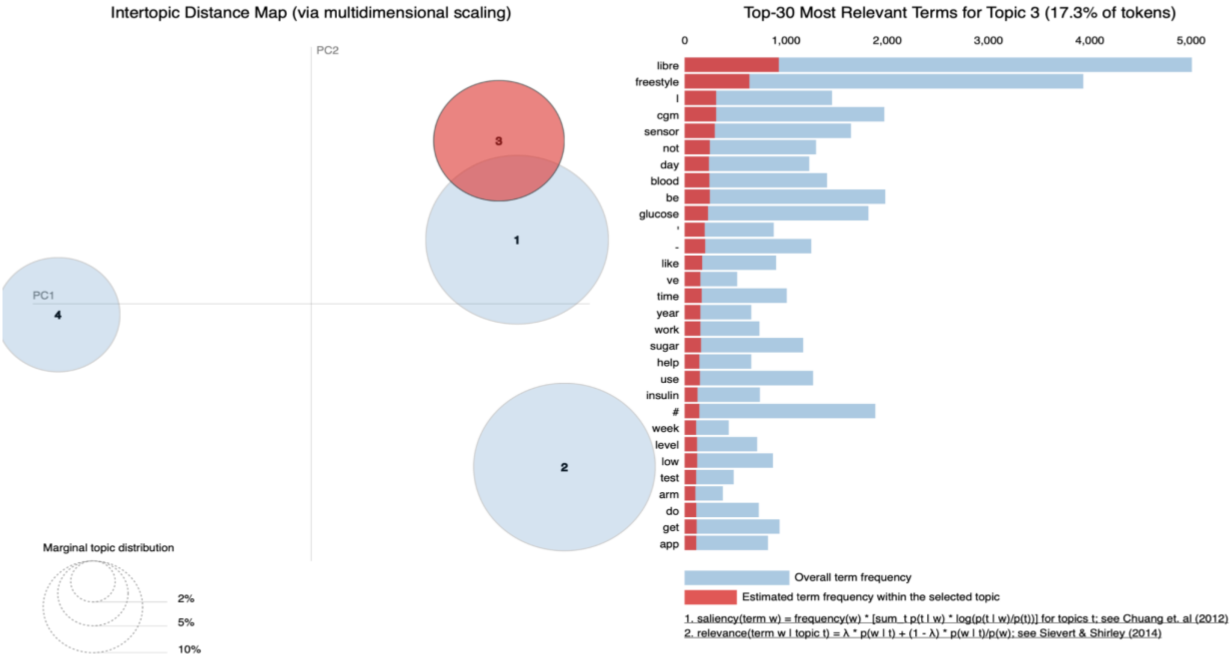
As shown in the table on the left side below, we believe the best hyperparameters to be applied on our model are: *[#Topics: 4, Alpha: 0.01, Beta: 0.21]*. We realized, based on the table on the right side, 4 topics would not generate the highest coherence score. However, we believe 4 still be the most suitable value to implement if we take the interpretability of the result into consideration: results from fewer (less than 4) clusters could be too generalized for us to extract insightful information, whereas too many topics create artificial boundaries within real data clusters.

Hyperparameters	Value
# of cluster:	4
Alpha:	0.01
Beta:	0.21

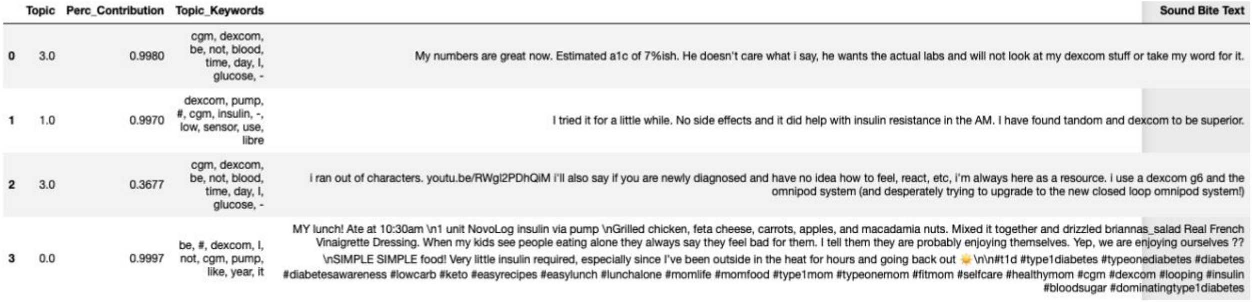


Clustering Results and Analysis

As the result of LDA clustering is shown below, the first thing we find is that there is a situation of overlapping happening between clustering 1 and clustering 3. This result consolidates the assumptions and hypotheses about overlapping data we made in the K-means model and proves that the LDA model can deliver better performance on our dataset. The bar chart on the right shows an example of Top-30 most relevant terms for each cluster/topic. Although the most relevant terms for each cluster vary in some degrees compared to the result of the K-means model, the majority of the top words, such as “freestyle” and “blood”, are the same. Hence, we will map the clustering result to the original dataset to explore more useful information



Thanks to the powerful python package *gensim*, we are able to rank all the records based on their contribution within each cluster. The contribution score describes the percentage of portion in each text that contributes to their corresponding topics. The higher the contribution score is, the more relatant the text is to its topic. We then read through the top 10 texts that have the highest contribution scores to get a better understanding of what each topic is talking about.



	Topic	Perc_Contribution	Topic_Keywords	Sound Bite Text
0	3.0	0.9980	cgm, dexcom, be, not, blood, time, day, I, glucose, -	My numbers are great now. Estimated a1c of 7%ish. He doesn't care what i say, he wants the actual labs and will not look at my dexcom stuff or take my word for it.
1	1.0	0.9970	dexcom, pump, #, cgm, insulin, -, low, sensor, use, libre	I tried it for a little while. No side effects and it did help with insulin resistance in the AM. I have found tandom and dexcom to be superior.
2	3.0	0.3677	cgm, dexcom, be, not, blood, time, day, I, glucose, -	i ran out of characters. youtu.be/RWgl2PDhQiM i'll also say if you are newly diagnosed and have no idea how to feel, react, etc, i'm always here as a resource. i use a dexcom g6 and the omnipod system (and desperately trying to upgrade to the new closed loop omnipod system!)
3	0.0	0.9997	be, #, dexcom, I, not, cgm, pump, like, year, it	MY lunch! Ate at 10:30am \n1 unit NovoLog insulin via pump \nGrilled chicken, feta cheese, carrots, apples, and macadamia nuts. Mixed it together and drizzled briannas_salad Real French Vinaigrette Dressing. When my kids see people eating alone they always say they feel bad for them. I tell them they are probably enjoying themselves. Yep, we are enjoying ourselves ?? \nSIMPLE SIMPLE food! Very little insulin required, especially since I've been outside in the heat for hours and going back out ☀️ \n\n#t1d #type1diabetes #typeonediabetes #diabetes #diabetesawareness #lowcarb #keto #easyrecipes #easylunch #lunchalone #momlife #momfood #type1mom #typeonemom #fitmom #selfcare #healthymom #cgm #dexcom #looping #insulin #bloodsugar #dominatingtype1diabetes

Topic	Examples of Texts with High Contribution Score	Defined Topic
1	*"…I was in school and my Dexcom CGM was beeping like crazy…"* *"...good, my Blood Glucose responds more to bread-ish things…"*	The daily experience of people have diabetes
2	*"The ability to track and map the ups, downs and patterns…"* *"Ever since a late Google update to Android 10 on my old Pixel 2…"*	Features built into the product and associated hardware devices
3	*"And what pump do you suggest?..."* *"Mine reads low sometimes, and my doctor says to trust my symptoms…"*	Users Questions and knowledge Gap
4	*"I loved our MM2/Libre combo and if we hadn't had to switch to…."* *"Dexcom G6 works great, far better than Medtronic CGM…"*	Users' preference on different brands of CGMs

Reflection and Improvement on LDA

A few things have been addressed during the modeling process and could be implemented to improve the performances of our LDA model.

1. **Balance Class Size**: Gather more samples from the group that has less samples to minimize the effect of imbalance on our model.
2. **Data processing techniques**: the data filtering and lemmatization process remains relatively general. For filters, we can design a more accurate filter that exclusively works on tackling specific questions such as product features or users' knowledge gap. At the same time, the ways that we tokenize and lemmantize the words will also affect the result of our model. Hence it is a worth to try out different techniques such as bigram/trigram lemmatization

3. **Different NLP models**: We can use the LDA model as the baseline model and test out other NLP models, for example, Bernoulli Naive Bayes and Support Vector Machine.

Conclusions – Topic Modeling

Based on the clustered text and distilled topic content, we obtain more key takeaways as a conclusion of our general analysis:

Some Users are concerned with the accuracy of certain CMGs as they realize that CGM readings do not match with their objective feelings and symptoms. Therefore, a question like “Should users put more importance into the CGM readings or their own symptoms?” are raised, disregard the brands.

There are groups who are interested in how CGMs can interact with other hardware/devices, such as insulin pumps. So we suggest finding out if there is a certain brand of CGM that integrates with a specific insulin pump or other hardware.

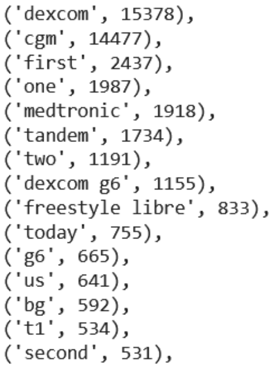
There are mixed opinions about alert/notification systems built into CGMs - some find it annoying, while others find it helpful. This finding could help companies to improve their products in terms of user experience. For example, allowing users to customize the alert system in some degree without affecting the main function of features

People who are considering switching from their current CGM brand are usually comparing Dexcom and Freestyle Libre. Echoing our previous findings, we believe that Dexcom and Freestyle libre are the two main competitors in the CGMs market. Hence, we will explore the characteristics of the products of both companies and compare them side-by-side in the following sections.

CGM Product Related Analysis

Before we delve into the two major players of the industry, Dexcom and Abbott, we wanted to see what other CGM products people discussed on social media. After performing a named-entity recognition analysis, we found that among the top 15 most talked about named entities, Dexcom, FreeStyle Libre, and Medtronic are the three most popular CGM products on social media posts, which makes sense since CGM is an oligopoly market. In addition, Dexcom CGM related words including “dexcom” and “g6” were mentioned ten times more than the other brands.

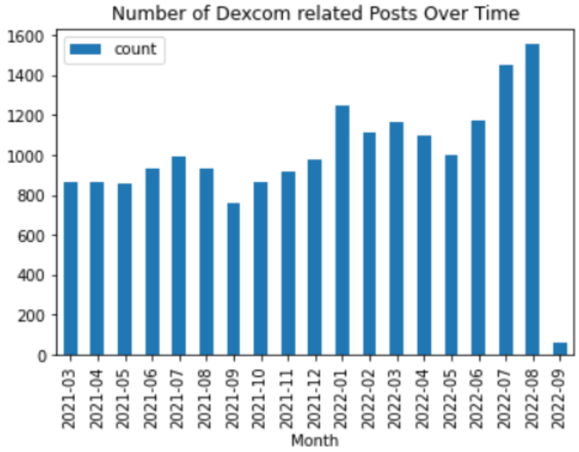
It is worth noting that Tandem insulin pump was brought up quite frequently by the customers, it is among the top 10 most talked about named entities in the dataset. The only CGM product that can be integrated with Tandem insulin pump is Dexcom G6.

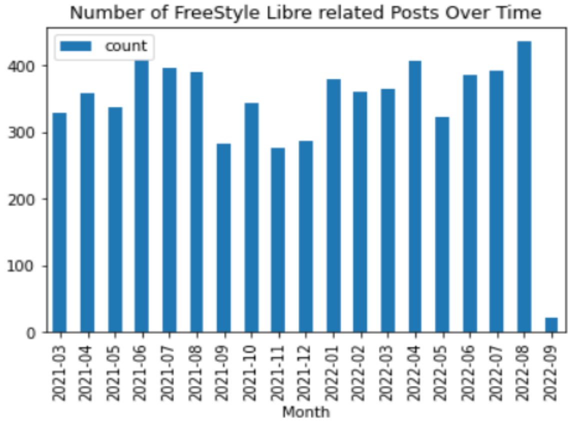


```
('dexcom', 15378),
('cgm', 14477),
('first', 2437),
('one', 1987),
('medtronic', 1918),
('tandem', 1734),
('two', 1191),
('dexcom g6', 1155),
('freestyle libre', 833),
('today', 755),
('g6', 665),
('us', 641),
('bg', 592),
('t1', 534),
('second', 531),
```

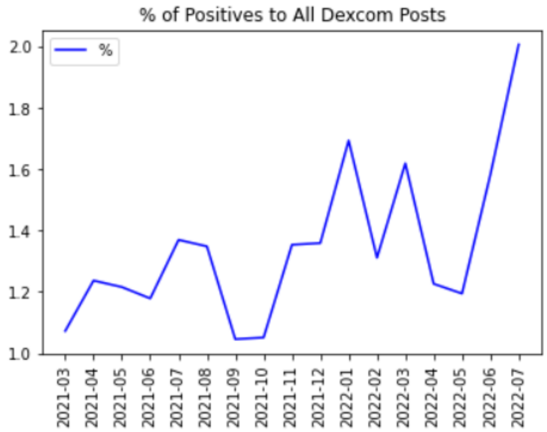
CGM Product Analysis - Dexcom and FreeStyle Libre

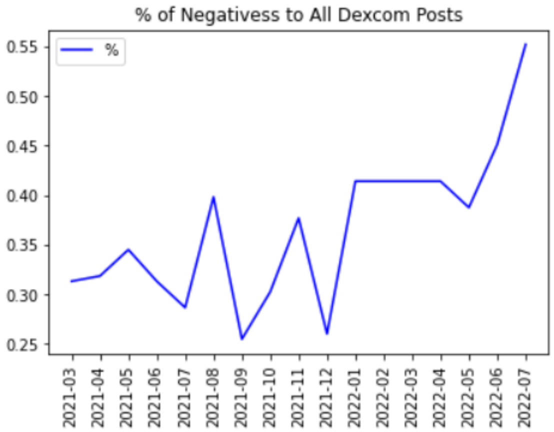
We wanted to break down the overall increasing post counts by brands to better understand product popularity on social media. Unsurprisingly, Dexcom was the main driver behind the overall trend in CGM posts. Dexcom has significantly higher year-over-year increase in post counts and average monthly post count compared to FreeStyle Libre.

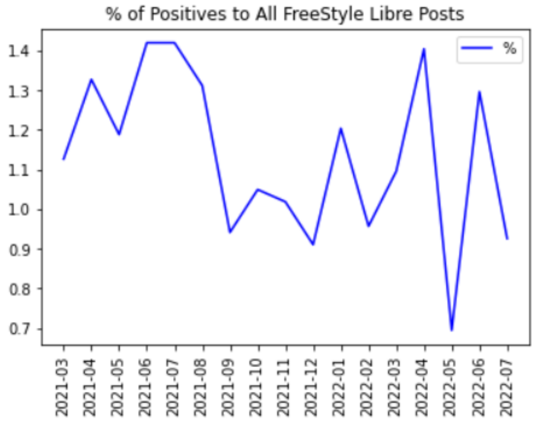


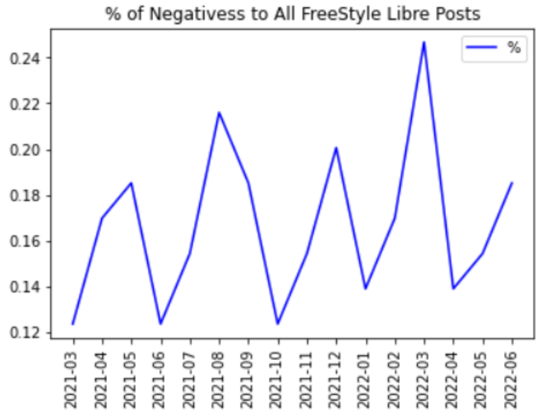


To better understand how customers' opinions shifted over time on each brand, we also broke down the proportion of positive and negative posts according to each brand. The data shows that while Dexcom's overall post count increased over time, more people are expressing positive and negative opinions about the product. However, FreeStyle Libre's positive posts seemed to have experienced a slight decrease over time, which is a signal worth paying attention to for the Abbott team.



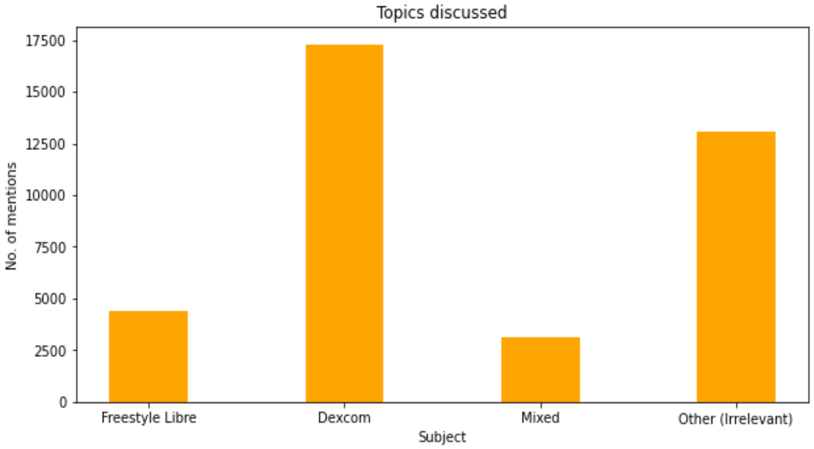


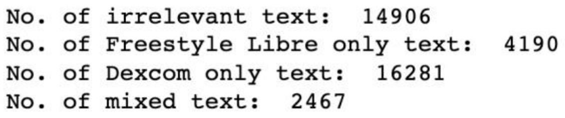




To understand what do customers like and dislike about each product series, we segment data by product topics (Dexcom and Freestyle Libre), and here are our observations:

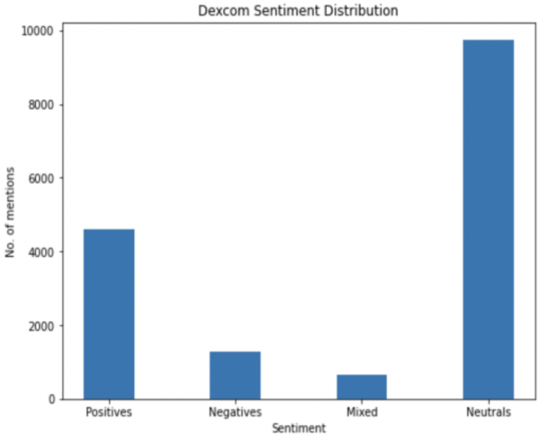
1. Imbalanced dataset: Dexcom topics are mentioned 10 times more frequently than Freestyle Libre

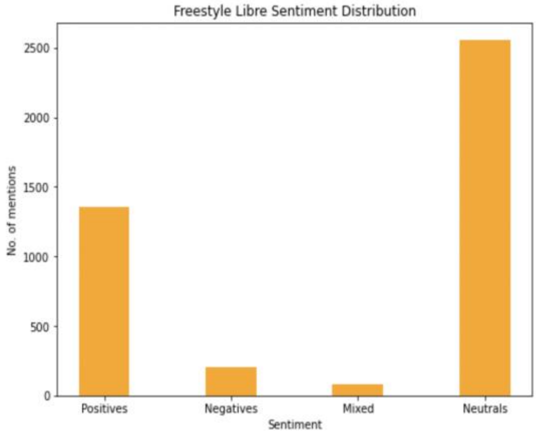




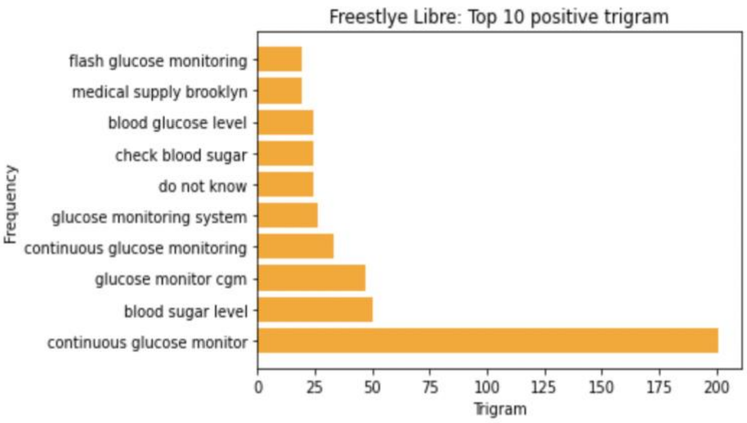
```
No. of irrelevant text:  14906
No. of Freestyle Libre only text:  4190
No. of Dexcom only text:  16281
No. of mixed text:  2467
```

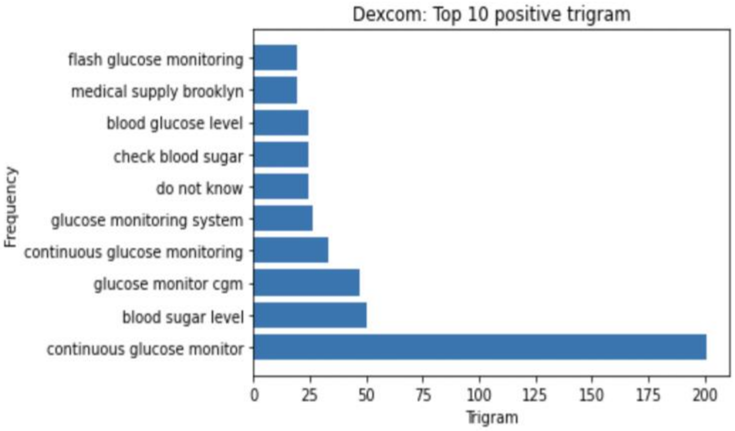
2. Neutral and Positive are the main sentiments expressed about either product brand



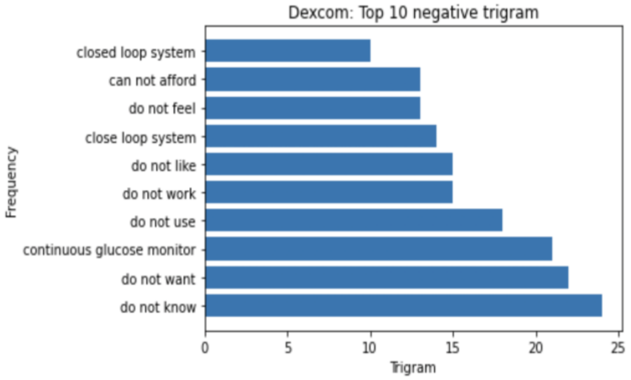


3. Under dexcom product top positive sentiments, “continuous glucose monitoring” is mentioned, suggesting Dexcom’s product feature of continuously tracking glucose levels in real-time and updating result on smartphone app/receiver device is receiving positive feedback from users

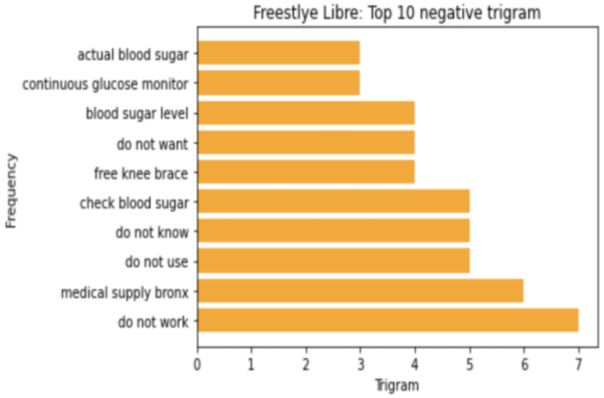




4. One of the top phrases mentioned in Dexcom negative sentiments is “can not afford”, showing customers are potentially dissatisfied with high cost of dexcom cgm products.



5. Under top negative sentiments trigram, there are phrases like “free knee brace” and “medical supply bronx” , which are noise in the data not relevant to our product topic. Going forward, we can further improve our analysis by consulting subject matter experts (SMEs) to determine relevant words and filter out irrelevant phrases and words before running our analysis.



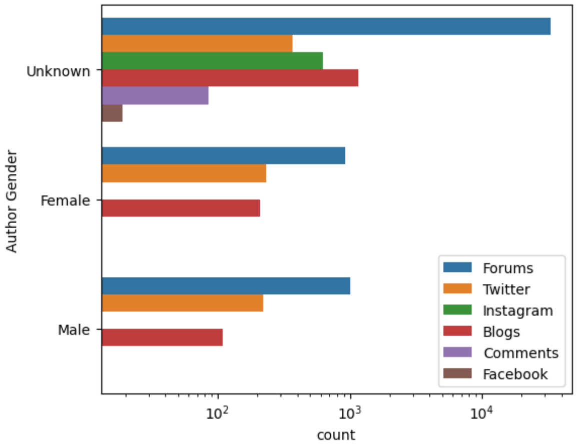
CGM Customer Related Analysis

In this section we further zoomed into understanding the post authors. As mentioned in the EDA section, many source-specific features have low data availability. Acknowledging the constraints of our dataset, we conducted consumer analysis using the available data. We looked into the following areas:

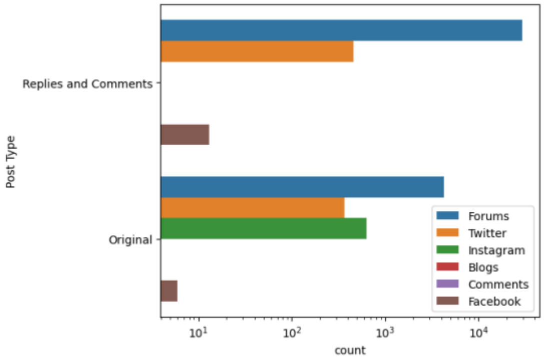
1. General demographics of post authors in the dataset
2. Segments of the “super authors” in the dataset who made the most posts or are quoted the most

General descriptions of the post authors by source

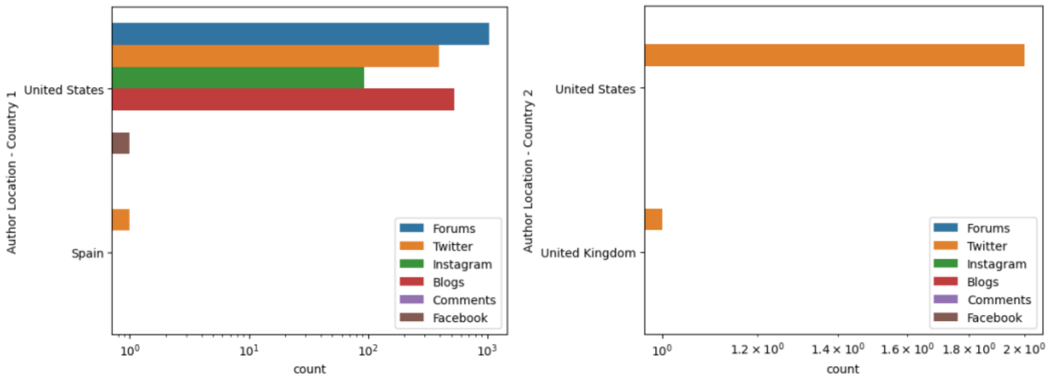
- There is no author gender information for all posts from Instagram, Comments, or Facebook, and for most of the posts from the remaining sources. We observe more posts from female authors than male authors on Blogs.



- Most of the posts in this dataset are replies and comments instead of original posts, except for Instagram where all posts are original. This may be a good sign, as the topic is spurring discussions in addition to one-way announcements/statements.

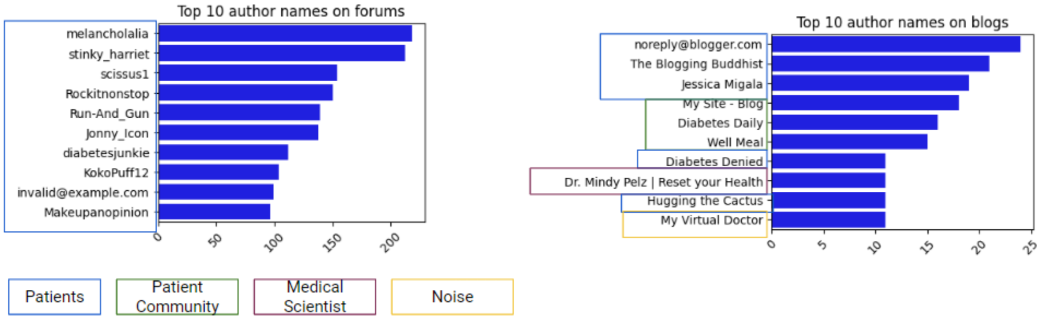


- The majority of the posts are posted by users in the United States. Among Twitter post authors, the top 5 source states are: California, Texas, New York, Florida, and Washington. Those states may be where most of the CGM consumers and potential adopters come from.



Super Authors: authors with a large number of posts/reposts

- The top 10 author names on forums have more than 90 posts each. The top 10 author names on blogs have more than 10 posts each. The difference in magnitude makes sense, as composing blogs requires much more effort than posting on forums.
- Most of the super authors are patients or patient communities sharing device usage experience, monitoring and lifestyle tips, as well as pricing.



Recommendation

We will start by addressing our key takeaways for the general CGM market, followed by brand specific recommendations to Dexcom and FreeStyle Libre. Overall, people agree that CGM is a better alternative compared with finger stick tests. After years of development since CGM's invention, people today are not too concerned about CGM's reading accuracy. However, price and user experience still remain top of mind for many customers. Therefore, CGM companies should continue to focus on improving their user experience and increase insurance coverage to ensure successful market expansion and increased patients' access to the technology.

Our product-specific analysis has shown that the main concern customers have for Dexcom is its product price. According to healthline.com, Dexcom G6 costs $1,500 more than FreeStyle Libre annually pre-insurance. Therefore, having more medical insurance coverage on the product is particularly important to Dexcom if it wants to sustain its profit margin and customer retention. As Dexcom plans to roll out its future products across the globe, having more local insurances to cover the product is one of the most critical steps to market penetration.

For FreeStyle Libre, its weak presence on social media compared to Dexcom implies it is not winning the leadership position in the market, and other players such as Meditronic may surpass it. Therefore, it is recommended that Abbott team engage with customers on social media, especially reddit, to raise brand awareness for the product. Based on our topic modeling analysis, a closed-loop system with an insulin pump seemed to be a valuable feature to customers. FreeStyle Libre should aim to integrate its future products with insulin pumps such as Tandem to stay competitive in the market. On the other hand, while the biggest difference between FreeStyle Libre and Dexcom is in their reading system, where Dexcom automatically records all the readings and FreeStyle Libre users have to manually scan the device, conflicting customer opinions on the real-time alert system Dexcom provides suggests FreeStyle Libre does not need to rush in launching a similar feature.