Diabetes Social Media Sentiment Analysis

By Jonathan W, Sai R, and Nehal C

Data Science for Product Managers, Final Project #1

Agenda

- Background
- Problem
- EDA
- Modeling
- Results
- Next Steps

Background

What Problem is being solved?

Help the CGM industry understand product gaps and sentiments using Social Media Text Analysis

Use Cases

1. What are the differences in sentiment between Libre and Dexcom?
2. Can we predict user sentiment based on CGM social media posts?

Who Benefits?

1.) Dexcom and Libre → 2.) Diabetic Patients

EDA

1) Manual Reading of Samples
2) Grouping
3) Vader Sentiment Analysis
4) Noun and Adjective tagging
5) Word Cloud

Overview

Overview | Alerts 124 | Reproduction

Dataset statistics

Number of variables	63
Number of observations	37844
Missing cells	1662490
Missing cells (%)	69.7%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	18.2 MiB
Average record size in memory	504.0 B

Variable types

Categorical	30
Unsupported	26
Boolean	1
Numeric	6

Out[3]:

	Post ID	Sound Bite Text	Title	Source Type	Sentiment	Positive Objects	Negative Objects	Source Name
0	BRDRDT2-t1_imq98sr	My numbers are great now. Estimated a1c of 7%i...	Have you been denied a second/third pump? Feel...	Forums	Neutrals	number	NaN	r/diabetes_t1
1	BRDRDT2-t1_impbcf4	I tried it for a little while. No side effects...	Metformin	Forums	Positives	NaN	NaN	r/diabetes_t1
2	1565738759353602048	i ran out of characters. youtu.be/RWgl2PDhQiM ...	NaN	Twitter	Positives	dexcom g6, omnipod system	NaN	NaN
3	17944607459251789	MY lunch! Ate at 10:30am \n1 unit NovoLog insu...	NaN	Instagram	Neutrals	NaN	NaN	NaN
4	BRDRDT2-t1_imq8h9m	This is also because like a soak in a hot tub ...	No bath salts, bath oils, soaks?	Forums	Neutrals	NaN	NaN	r/diabetes

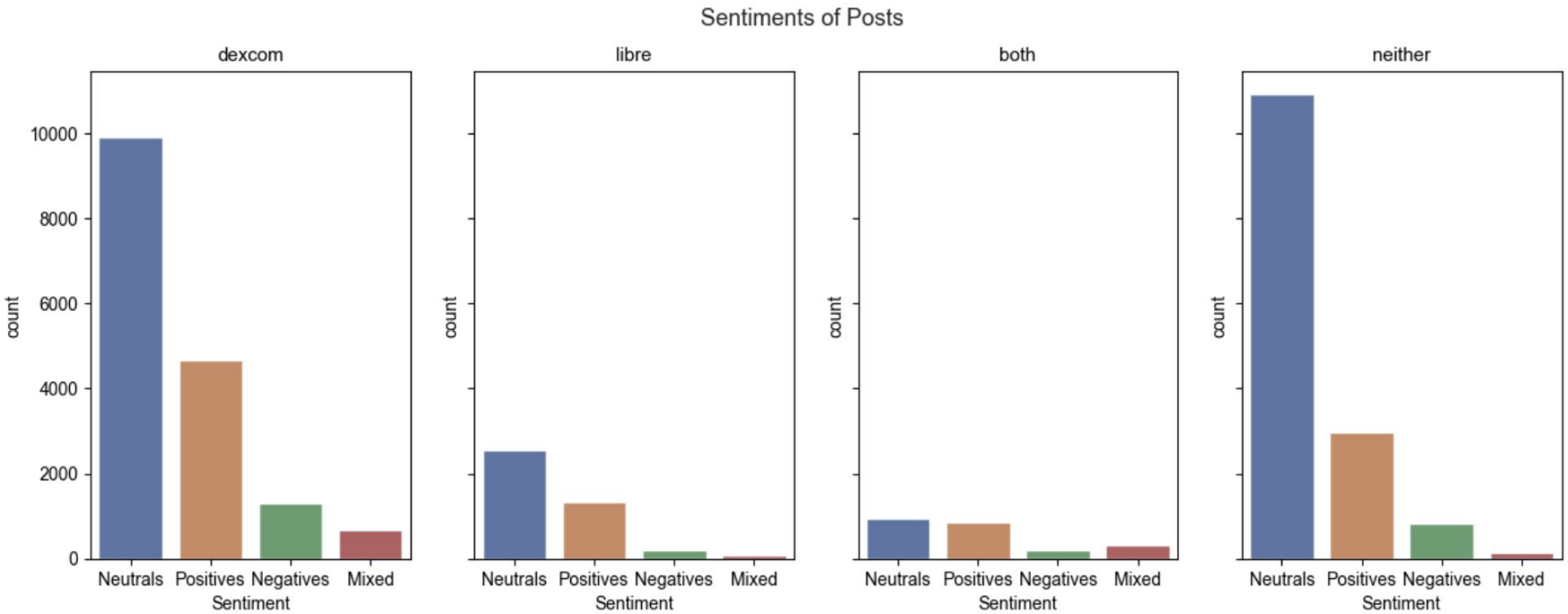
Initial Insights

Approach – Manually read **30 random ‘Sound Bite Text’** samples per team member. Look for **‘Original’** Post Type and filter on **‘Source Type’** for **‘r/dexcom’** and **‘r/Freestylelibre’.**

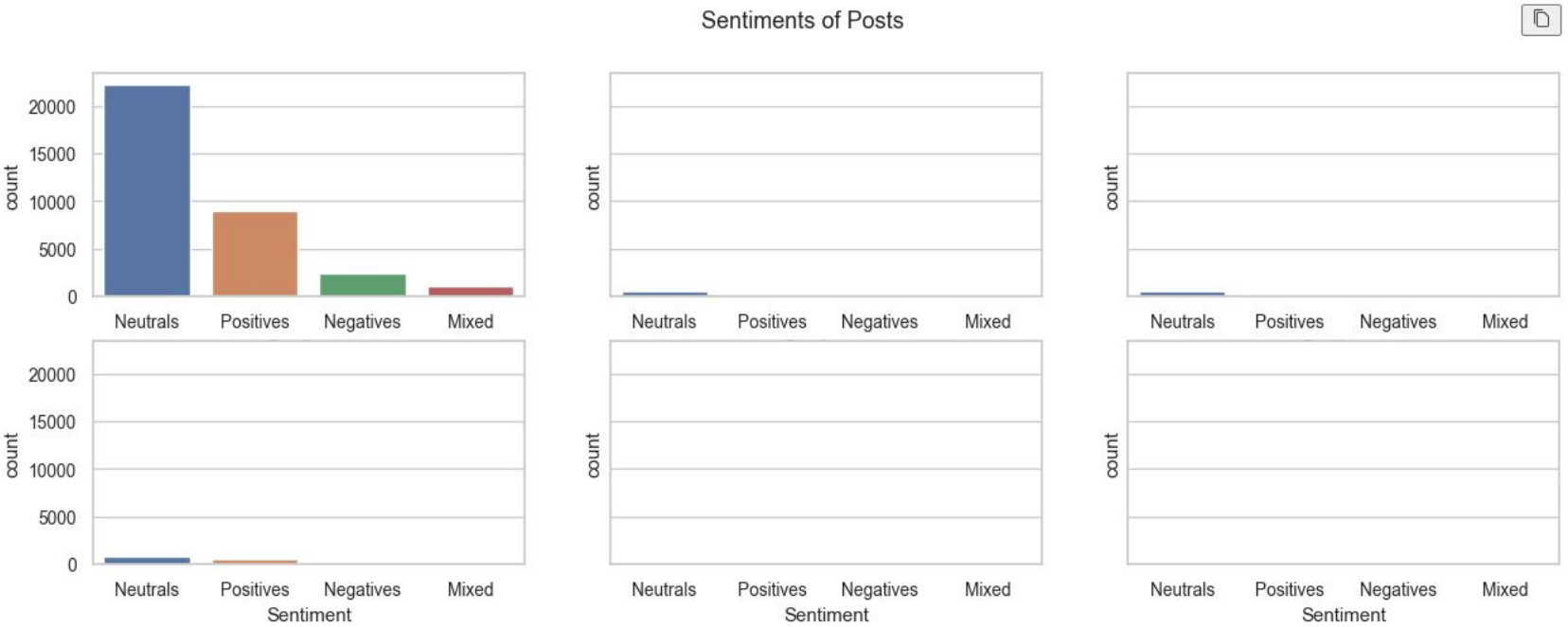
High Level Findings

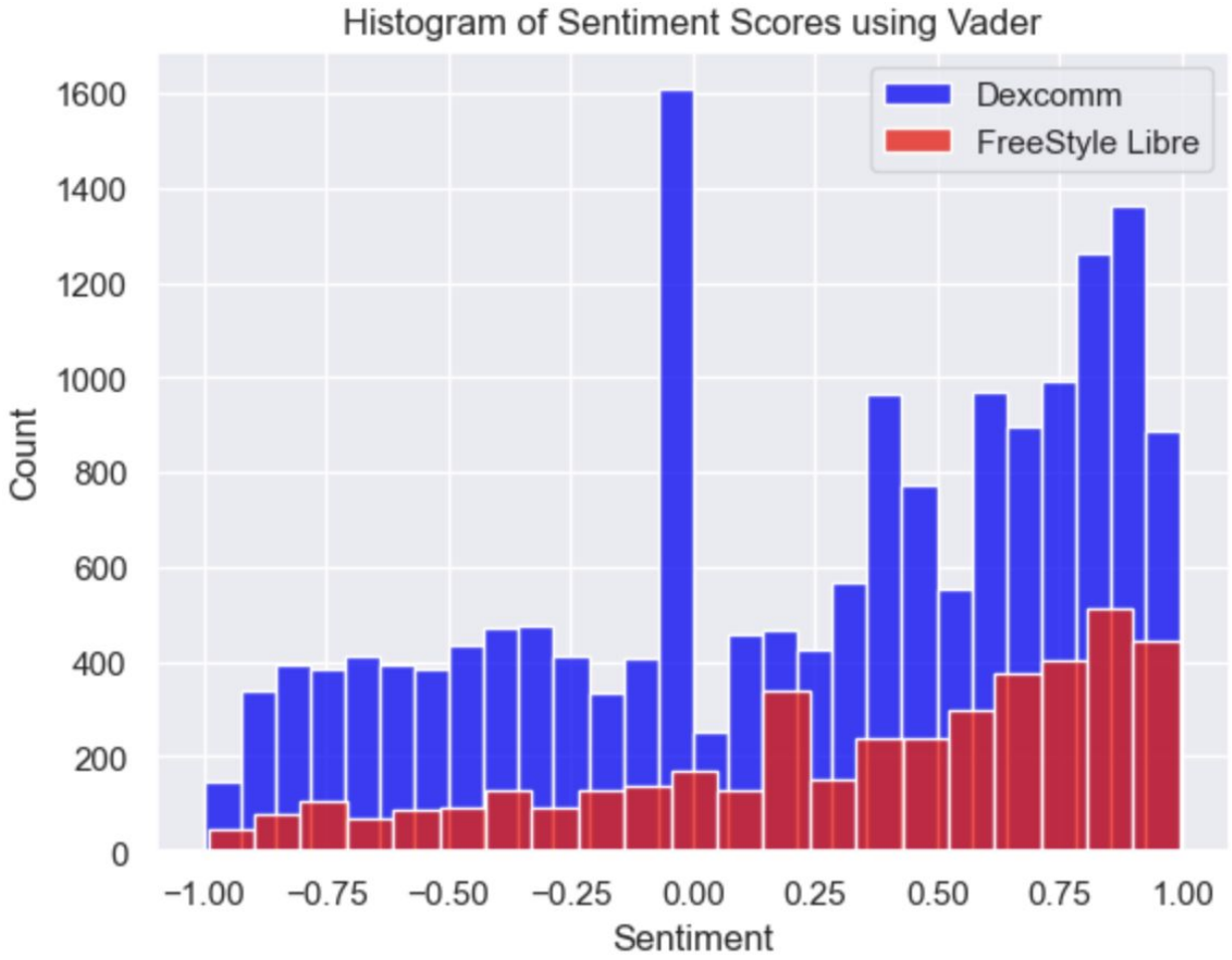
1) Mostly positive experiences compared to finger pricking alternative (life saving, more convenient, etc).
2) Advice and general discussion threads over CGM experiences.
3) Knowledge gaps: Questioning accuracy of some CGM monitors, setup issues, service coverage
4) Patient Expectations: better manage your glucose levels every day, have fewer low blood glucose emergencies, need fewer finger sticks

Dataset Sentiments Summary



Source type Grouping




Histogram of Sentiment Scores using Vader
Dexcomm
FreeStyle Libre
Count
1600
1400
1200
1000
800
600
400
200
0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Sentiment

Vader Sentiment Scores

```
pd.Series(scores_dexcom).describe()
```

```
count    17426.000000
mean         0.226902
std          0.552349
min         -0.996500
25%         -0.200300
50%          0.340000
75%          0.726900
max          0.997600
dtype: float64
```

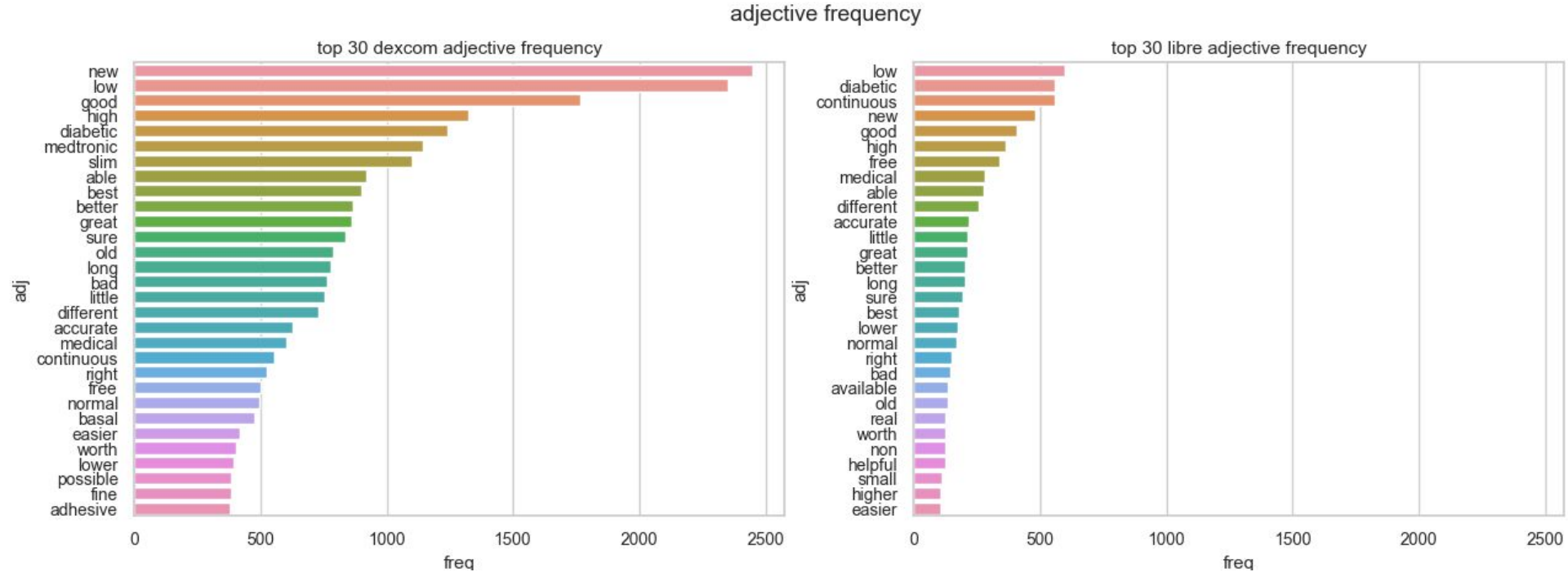
```
pd.Series(scores_libre).describe()
```

```
count    4269.000000
mean        0.343184
std         0.518926
min        -0.991300
25%         0.000000
50%         0.493900
75%         0.784500
max         0.995600
dtype: float64
```

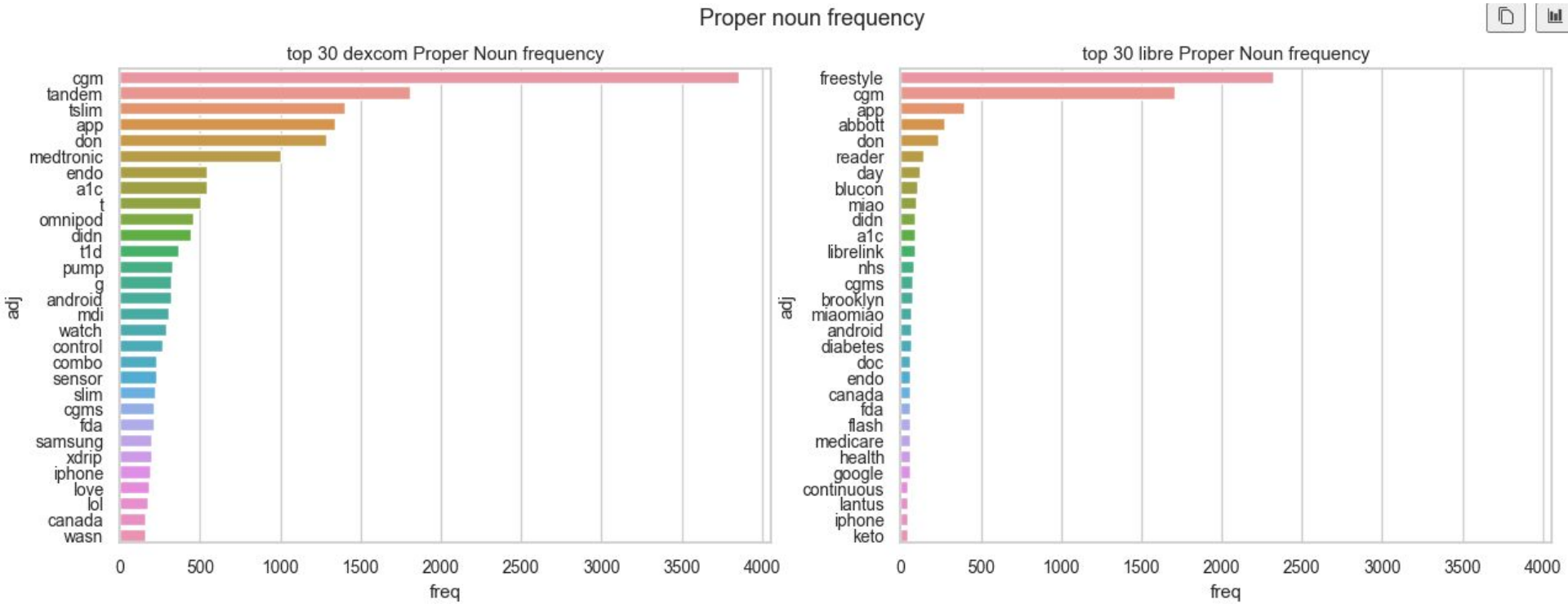
Histogram of Top Adjectives for Positive CGM Sentiments
Frequency (log scale)
2×10^3
10^3
1980
1628
1206
1139
1133
1024
835
731
low
continuous
good
high
diabetic
new
better
able
best

Histogram of Top Adjectives for Negative CGM Sentiments
Frequency (log scale)
6 × 10^2
4 × 10^2
3 × 10^2
2 × 10^2
586
318
266
246
217
172
159
157
151
low
new
high
good
diabetic
better
bad
continuous
able
old

Top Adjectives



Top Proper Nouns



Adjectives

Proper Nouns

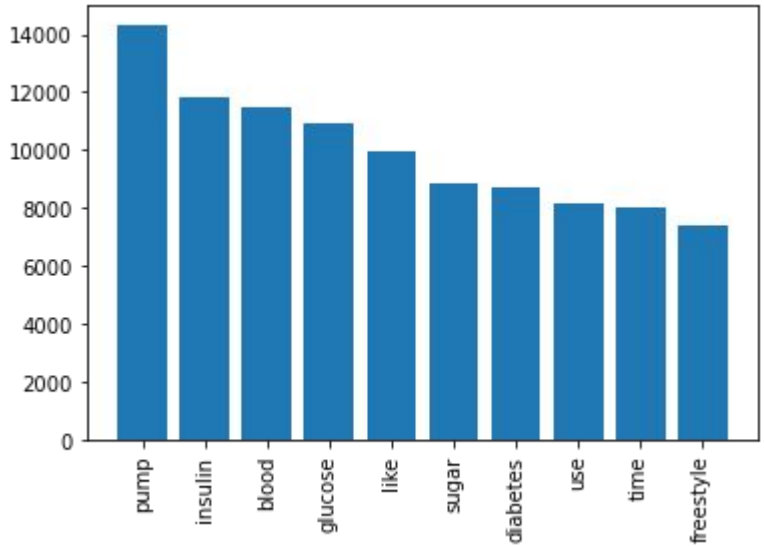
Word Clouds

Data Preparation

General goal: Given a body of text on the topic of Diabetes CGM, can we correctly predict user sentiment?

- Combined text columns
- Filtering criteria
  - Stop words, alphanumeric, ‘http’ in string, etc.
- Tokenization and Lemmatization

Top 10 words after processing



Additional Preparation (Train/Test Set)

- Cleaning and lemmatization
- **TF-IDF vectorization**
  - (1,3) ngram range
- 698 Features
- Target variable mapping

Target variable mapping

Sentiment	target
Mixed	0
Negatives	1
Neutrals	2
Positives	3

```
tfidf_vectorizer= TfidfVectorizer(min_df=300, stop_words="english", max_df=0.8,
                                  analyzer="word", token_pattern=r"(?u)\b[A-Za-z']+\b", ngram_range=(1,3))
```

Modeling

- **3 Classifier Models**
  - Multinomial Naive Bayes
  - Random Forest Classifier
  - Gradient Boosting Classifier

  - Train/Test Split, 80% – 20%
  - Grid Search CV

Results Evaluation

Multinomial Naive Bayes

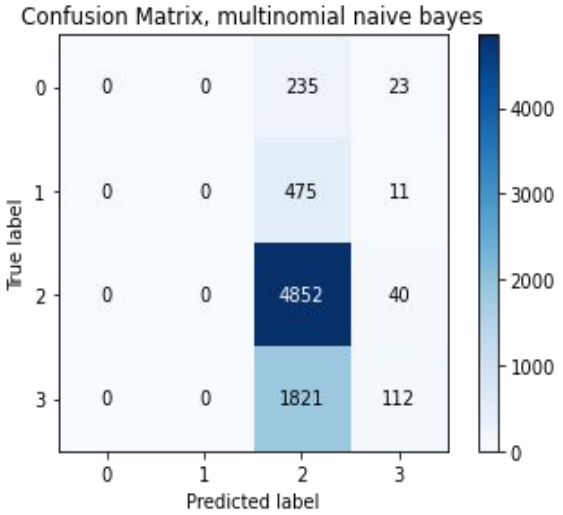
- Accuracy: 0.65
- F1 Score: 0.78

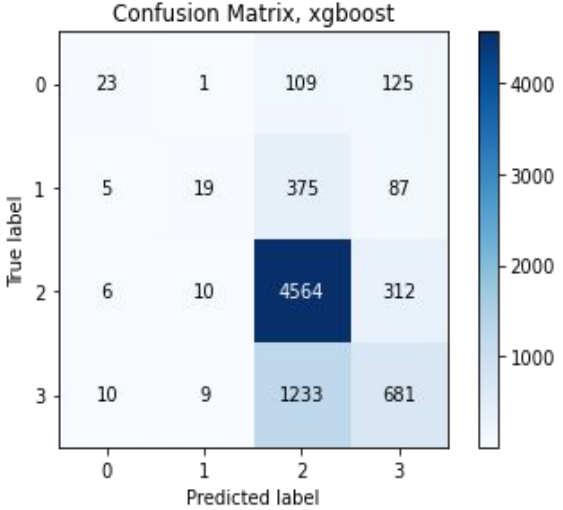
Random Forest Classifier

- Accuracy: 0.66
- F1 Score: 0.75

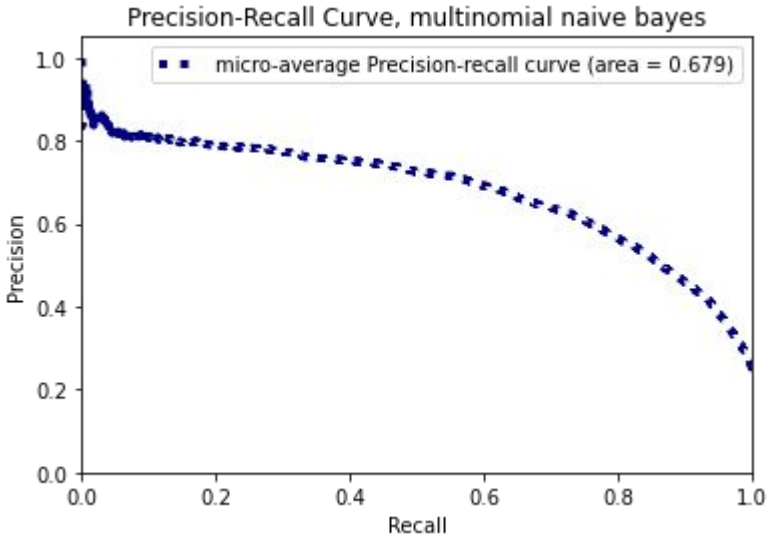
XGBoost Classifier

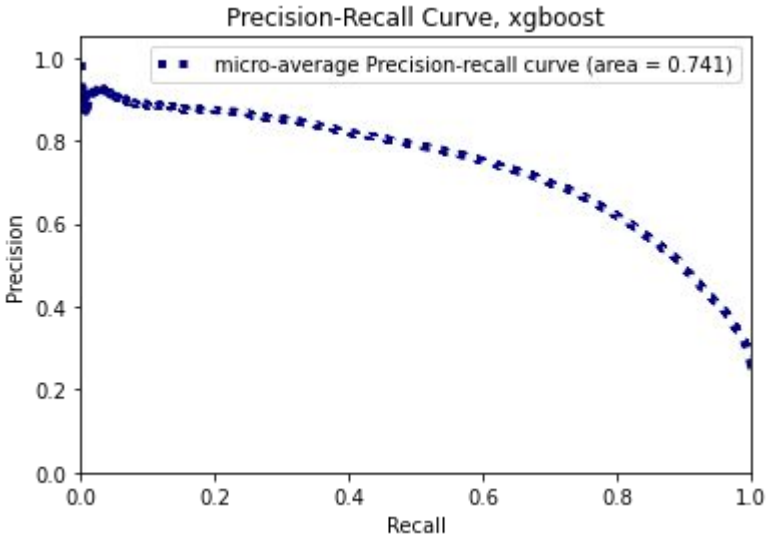
- Accuracy: 0.66
- F1 Score: 0.73



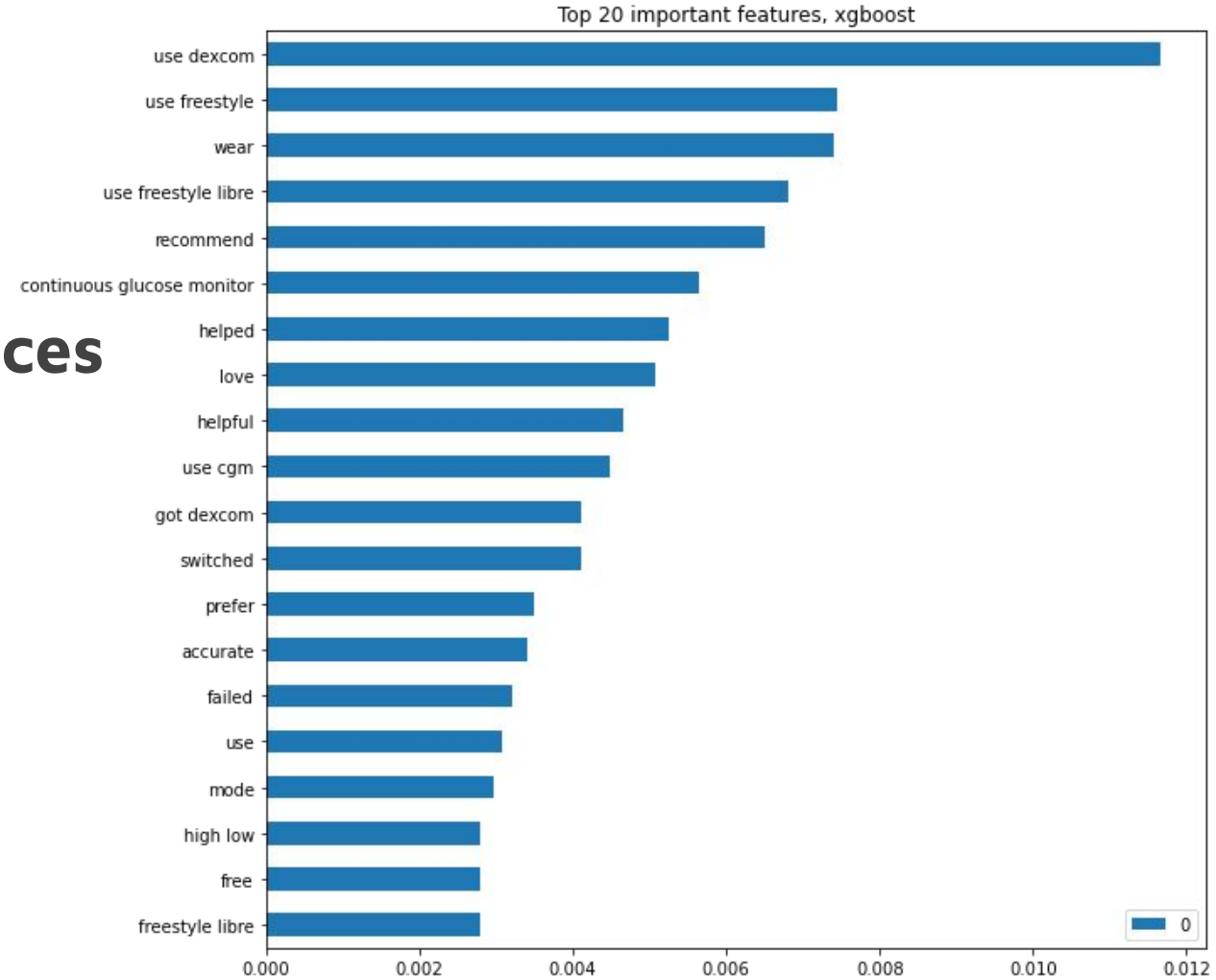


Precision-Recall Curves





Feature Importances XGBoost



Recommendations and Future Steps

Data Science

- Data needs more information – current sentiment column is limited.
- Propose one or two **numerical columns**, with sentiment ranges from *-5 to 5* in each.
- **Class imbalance**, differentiate models by Dexcom and Libre

Product

- In addition to sentiment analysis, **survey data** to clearly differentiate between products
- Include business analytics for Dexcom and Libre using tools such as **CLTV**, **churn from products**, and revenue incrementality.

Questions