# Medical Requests Classification

## Diabetes Vs. Asthma

By Mekdes Wassie

# Problem Statement

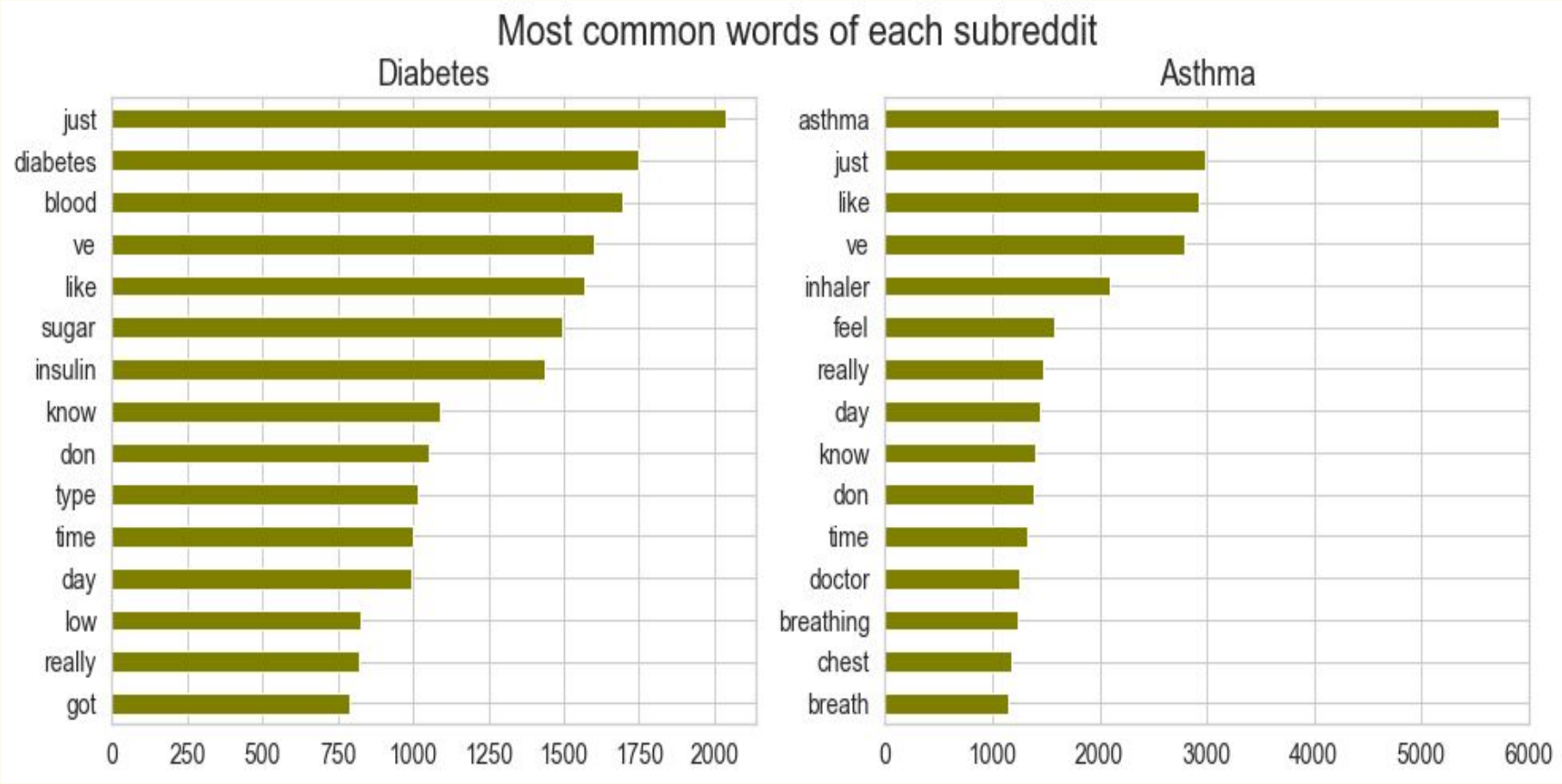
- A research institute has plans to conduct a research on medical apps for self-diagnosis of diabetes.
- They requested a solution for sorting patient requests based on their diagnosis.

# Data Collection and EDA

- Data source: Diabetes and Asthma subreddits
- Data collection: 5000 posts from each subreddit
- EDA:
  - compared word counts and character length
  - most common words

# character & word count of posts

| Subreddit | Character Length | Word Count |
|---|---|---|
| Diabetes | 562 | 105 |
| Asthma | 651 | 121 |


Most common words of each subreddit
Diabetes
just
diabetes
blood
ve
like
sugar
insulin
know
don
type
time
day
low
really
got
0
250
500
750
1000
1250
1500
1750
2000
Asthma
asthma
just
like
ve
inhaler
feel
really
day
know
don
time
doctor
breathing
chest
breath
0
1000
2000
3000
4000
5000
6000

# Baseline Accuracy

46.9%

# Modeling

| Estimator | Transformer | Best score (%) | Accuracy on train (%) | Accuracy on test (%) |
|---|---|---|---|---|
| Knn | TF-IDF Vectorizer | 55.7 | 98.6 | 57.3 |
| Multinomial Naive Bayes | Countvectorizer | 95.6 | 96.4 | 95.5 |
| Multinomial Naive Bayes | TF-IDF Vectorizer | 95.4 | 96.5 | 95.2 |
| Random forest | Countvectorizer | 94.8 | 98.6 | 95.3 |

# Evaluation Metrics

- False negative Vs. False positive
- Sensitivity
- Accuracy

# Evaluation Metrics Cont.

| Models | Sensitivity | Accuracy |
|---|---|---|
| Knn | 15.5 | 55.7 |
| Multinomial Naive Bayes - CV | 97.2 | 95.6 |
| Multinomial Naive Bayes - TF-IDF | 95.3 | 95.4 |
| Random forest | 97.6 | 94.8 |

# Best Model

**Multinomial NB with CountVectorizer**

- Accuracy of 95.6%
- Sensitivity of 97.2%

# Best Model Cont.

- CV parameters:
  - Max_df: 92%
  - Max_features: 5000
  - Min_df: 4
  - Ngram_range: 1, 1

# Best Model Cont.

| Feature Importance | Features |
|---|---|
| 0.116981 | asthma |
| 0.042034 | inhaler |
| 0.041308 | diabetes |
| 0.028943 | insulin |
| 0.021768 | sugar |
| 0.020404 | breathing |
| 0.018593 | diabetic |
| 0.017025 | blood |
| 0.015591 | breath |
| 0.014694 | chest |

# Conclusion

- The research institute may use the NB model to sort patient requests made on the apps.
- Accuracy of 95.6% and Sensitivity of 97.2%
- Improve the model:
  - stop words