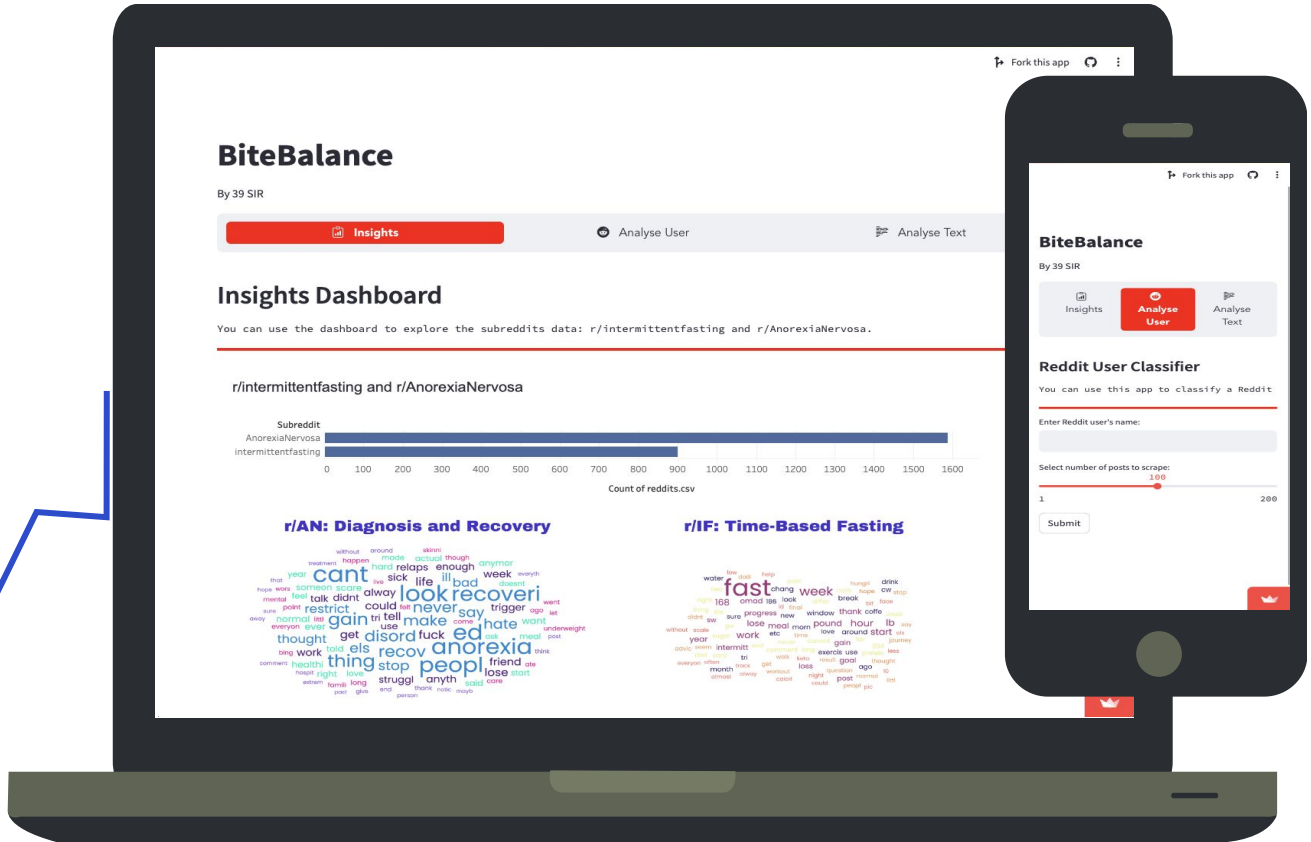
Scan QR code to try the app!



BiteBalance by 39 SIR Health

For GA-DSI-39

By 39-SIR (Alice, Eugenia, Farhan, Ivan, Sheila)

29 Sep 2023

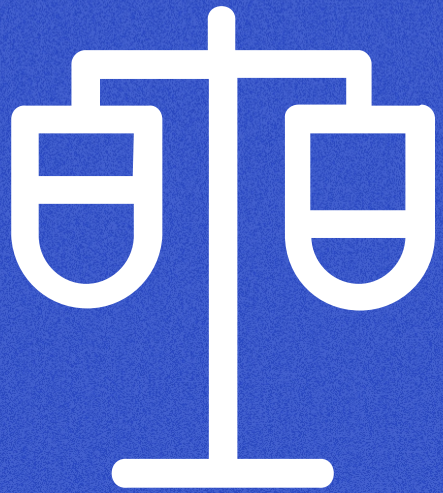
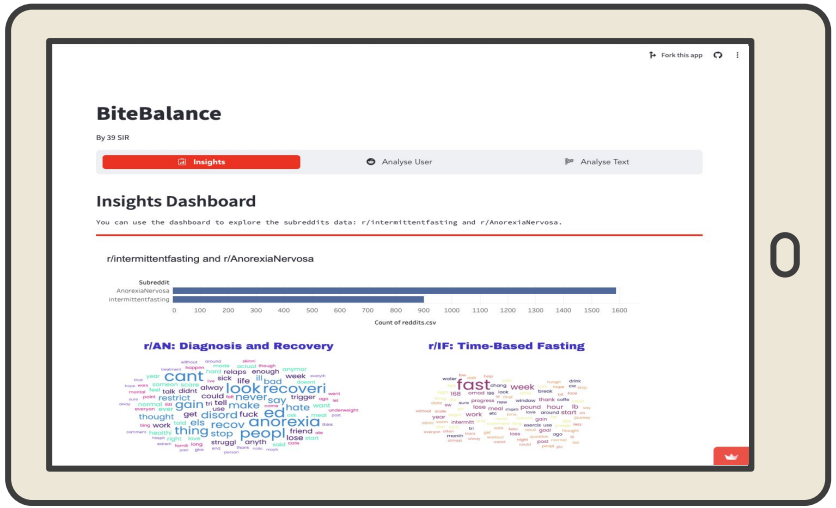


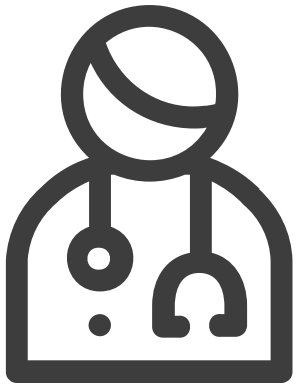
Table of Contents

Problem Statement

Dear **40 SIB Wellness** and **Corporate Partners**,

Welcome to the launch of **BiteBalance** by **39 SIR Health**

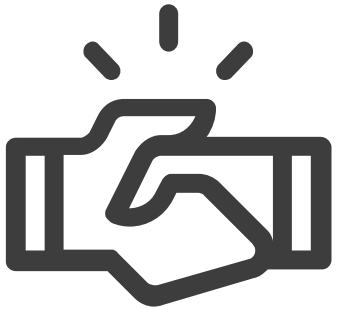




BiteBalance for You

Text-based classifier that predicts between **intermittent fasting** and **anorexia nervosa**

For clients
Investigation of early signs of disordered eating



For partners
Lead generation and referrals for health and wellness support

Background

Intermittent Fasting (IF)

A popular health trend:

- Controlled fasting and eating
- For weight management and wellness

Anorexia Nervosa (AN)

A severe mental health disorder:

- Extreme dietary restriction
- Fear of gaining weight
- Distorted body image

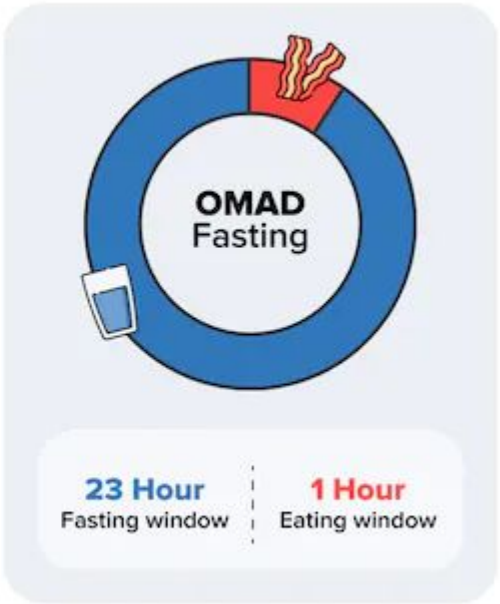
Sources:
https://www.hopkinsmedicine.org/health/wellness-and-prevention/intermittent-fasting-what-is-it-and-how-does-it-work
https://www.aware.org.sg/information/eating-disorders

Extreme Intermittent Fasting?

12:12

16:8

OMAD:
One Meal A Day



Source: https://www.carbmanager.com/article/yoherxeaaceazayu/how-to-choose-an-intermittent-fasting-schedule

Key Question

How can we **differentiate** between users from **r/intermittentfasting** and **r/AnorexiaNervosa** based on their Reddit posts?

Our PRAW-rocess

01 Web Scraping

3960 posts from **r/intermittentfasting** and **r/AnorexiaNervosa**

02 Data Cleaning & EDA

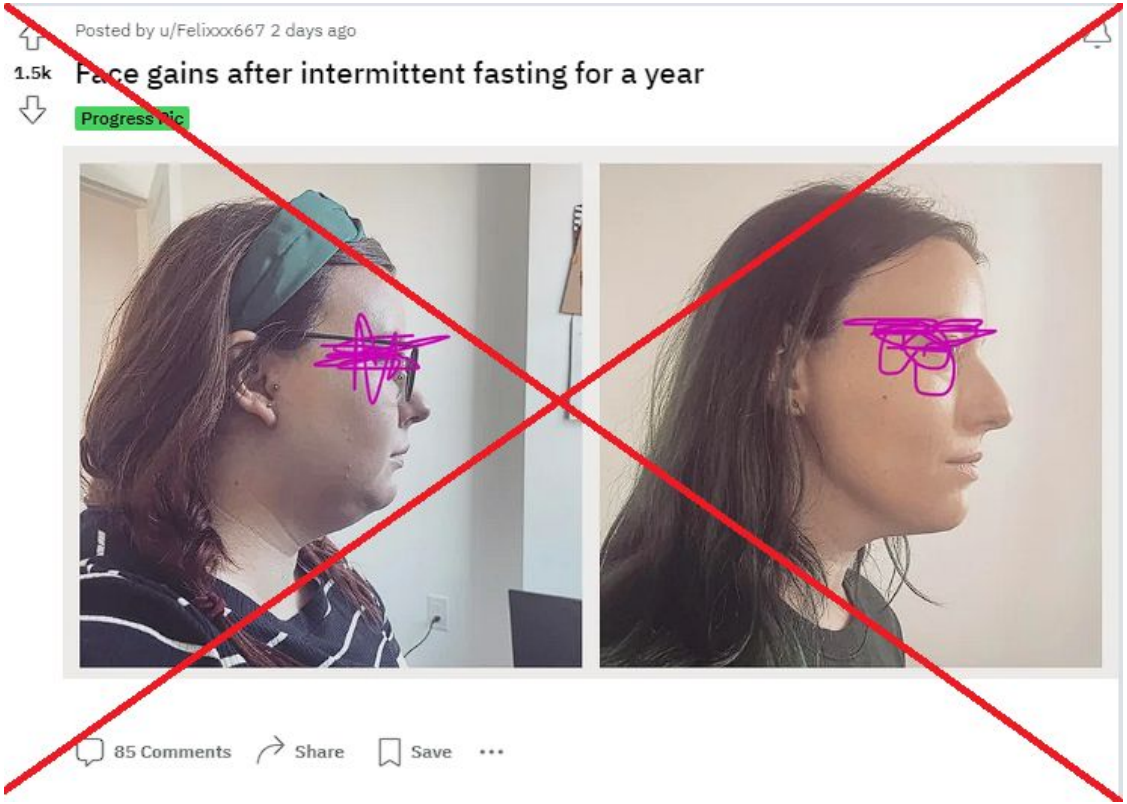
Removal of duplicates, and combining of title and text

03 Model and Evaluation

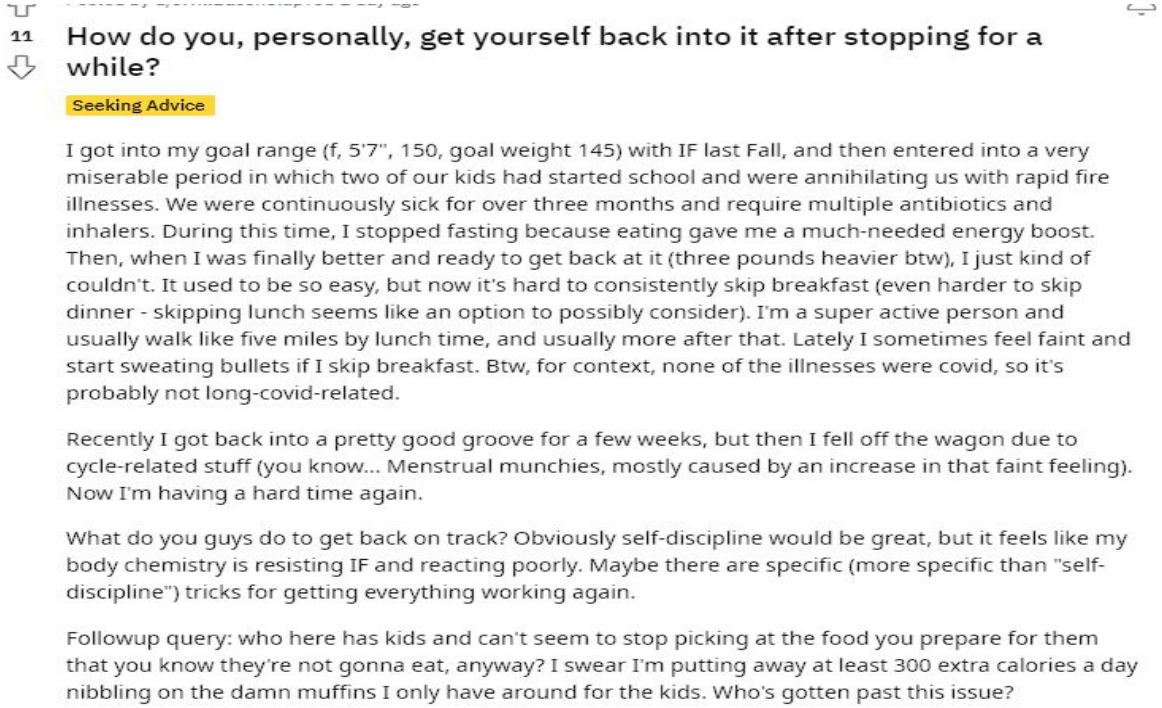
3 classifiers and 3 vectorizers

04 App Development

Reddit Habits: Text in Title Only



Reddit Habits: Text in Title and Body



11

How do you, personally, get yourself back into it after stopping for a while?

Seeking Advice

I got into my goal range (f, 5'7", 150, goal weight 145) with IF last Fall, and then entered into a very miserable period in which two of our kids had started school and were annihilating us with rapid fire illnesses. We were continuously sick for over three months and require multiple antibiotics and inhalers. During this time, I stopped fasting because eating gave me a much-needed energy boost. Then, when I was finally better and ready to get back at it (three pounds heavier btw), I just kind of couldn't. It used to be so easy, but now it's hard to consistently skip breakfast (even harder to skip dinner - skipping lunch seems like an option to possibly consider). I'm a super active person and usually walk like five miles by lunch time, and usually more after that. Lately I sometimes feel faint and start sweating bullets if I skip breakfast. Btw, for context, none of the illnesses were covid, so it's probably not long-covid-related.

Recently I got back into a pretty good groove for a few weeks, but then I fell off the wagon due to cycle-related stuff (you know... Menstrual munchies, mostly caused by an increase in that faint feeling). Now I'm having a hard time again.

What do you guys do to get back on track? Obviously self-discipline would be great, but it feels like my body chemistry is resisting IF and reacting poorly. Maybe there are specific (more specific than "self-discipline") tricks for getting everything working again.

Followup query: who here has kids and can't seem to stop picking at the food you prepare for them that you know they're not gonna eat, anyway? I swear I'm putting away at least 300 extra calories a day nibbling on the damn muffins I only have around for the kids. Who's gotten past this issue?

Data Cleaning

DataFrame

- Removed Irrelevant Posts
- Deduplicated Data
- Standardized Column Headers
- Handled Missing Data
- Merged 'Title' and 'Post' Columns

Text

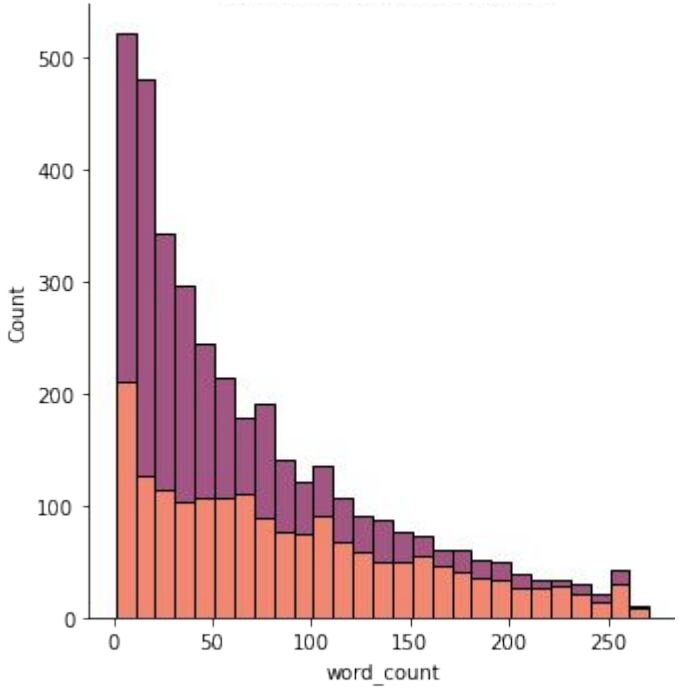
- Removed Punctuation
- Removed Stopwords
- Lowercase
- Drop Hyperlinks
- Stemming
- Lemmatization

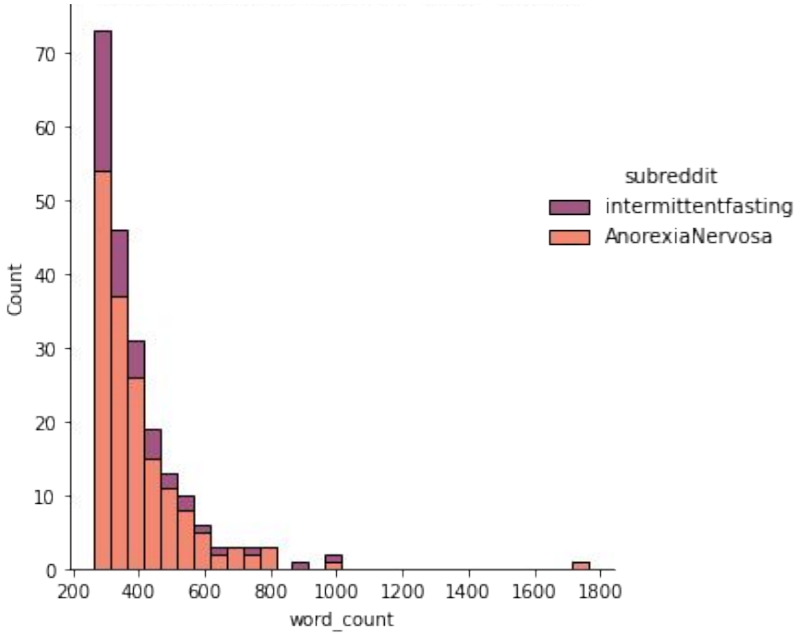
Data Cleaning: Stemming produces greater variety than lemmatizing

IF (S)	IF (L)	AN (S)	AN(L)
fast	fasting	cant	cant
week	fast	look	ed
hour	week	ed	anorexia
work	hour	anorexia	people
lb	lb	peopl	recovery
start	168	recoveri	thing
168	pound	thing	never
lose	year	gain	look
pound	meal	recov	else
year	work	disord	hate

EDA: Longer text posts in r/AN compared to r/IF

Distribution of Text Length (Outliers Removed vs Upper Outliers)





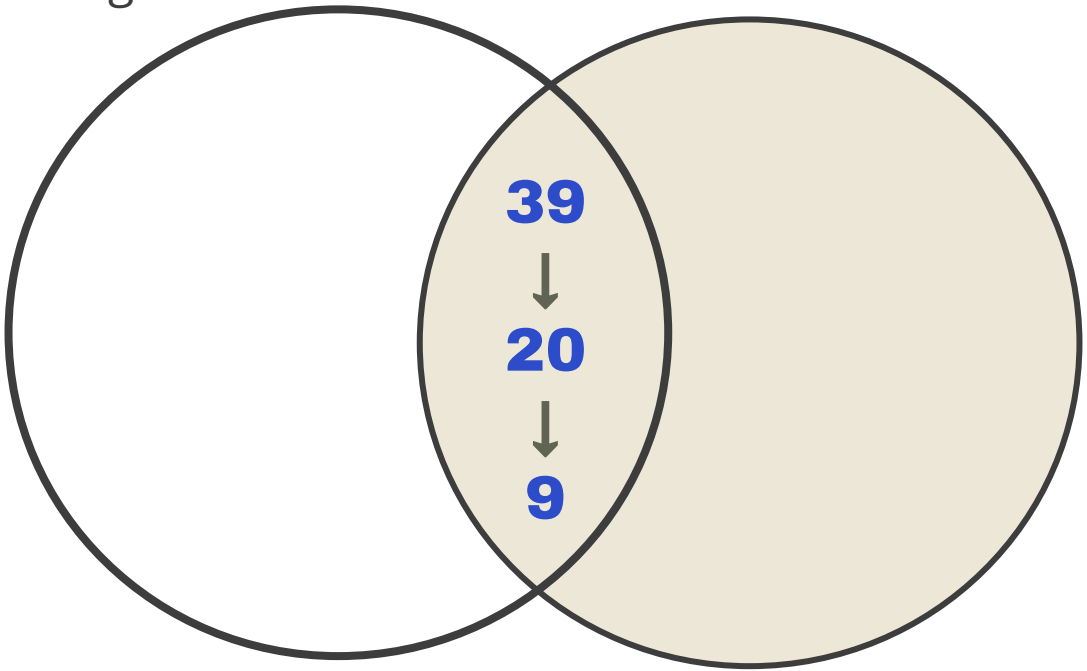
r/IF: Time-Based Fasting

r/AN: Diagnosis and Recovery

Creation of Stopwords

Top 100 Words
r/intermittentfasting

Top 100 Words
r/AnorexiaNervosa



r/IF: Time-Based Fasting

Intermittent Fasting	Anorexia
intermitt fast	sick enough
break fast	mental health
hour fast	ice cream
fast hour	anorexia nervosa
black coffe	extrem hunger
fast window	best friend
fast water	hate hate
type fast	high school
daili fast	diagnos anorexia
hour etc	struggl anorexia

r/AN: Discussion of Mental Health

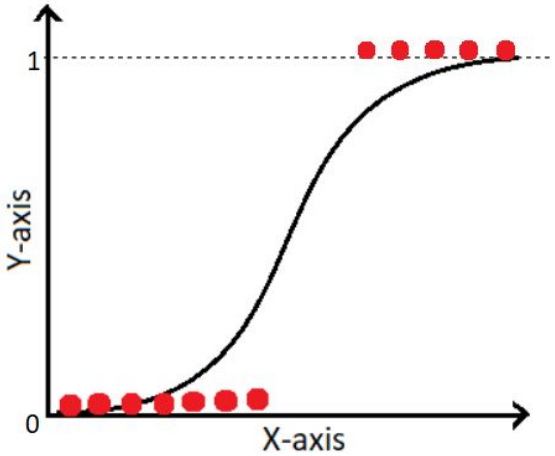
Intermittent Fasting	Anorexia
intermitt fast	sick enough
break fast	mental health
hour fast	ice cream
fast hour	anorexia nervosa
black coffe	extrem hunger
fast window	best friend
fast water	hate hate
type fast	high school
daili fast	diagnos anorexia
hour etc	struggl anorexia

r/IF, r/AN: Food Keywords

Intermittent Fasting	Anorexia
intermitt fast	sick enough
break fast	mental health
hour fast	ice cream
fast hour	anorexia nervosa
black coffe	extrem hunger
fast window	best friend
fast water	hate hate
type fast	high school
daili fast	diagnos anorexia
hour etc	struggl anorexia

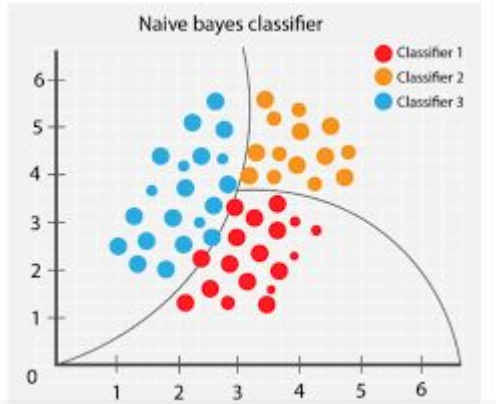
Classification Models

Logistic Regression



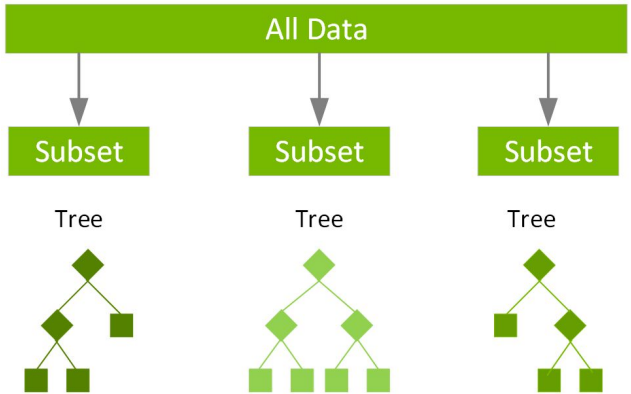
Graph

Naive Bayes



Boundaries

XGBoost

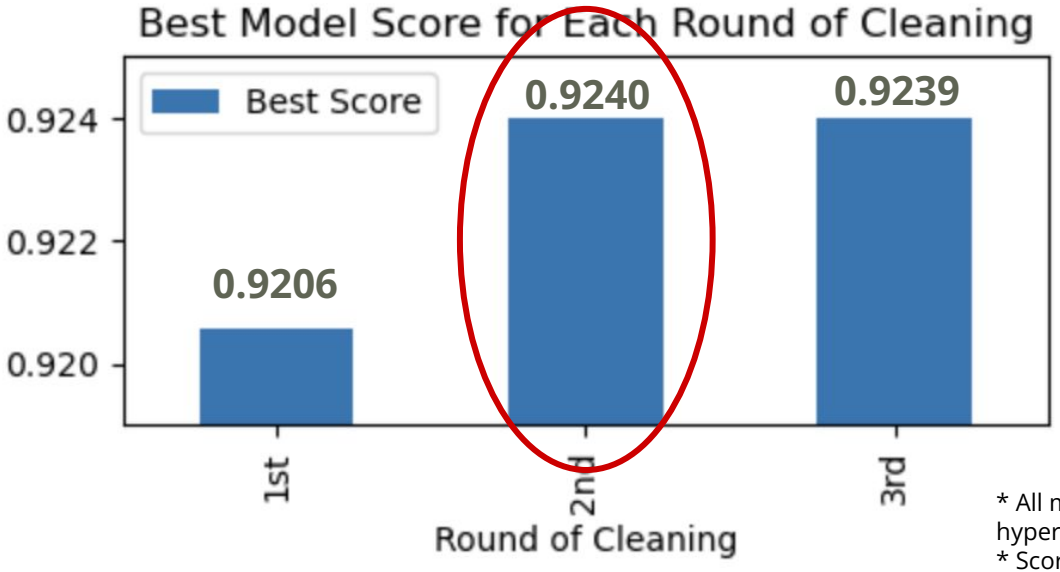


Decision Trees


Getting the best score
1. Feature Engineering
2. Compare Against Other Models

Baseline Model: Logistic Regression

Model	Vectorizer	Accuracy (train)	Accuracy (test)	Precision (test)	Recall (test)	F1 (test)	ROC (test)	Execution time (ms)
Logistic Regression	Count Vectorizer	0.96	0.92	0.94	0.9	0.92	0.92	10.33



* All models have been optimised by tuning their hyperparameters before being run.
* Scores closer to 1 the better

Total of 9 Models Deployed

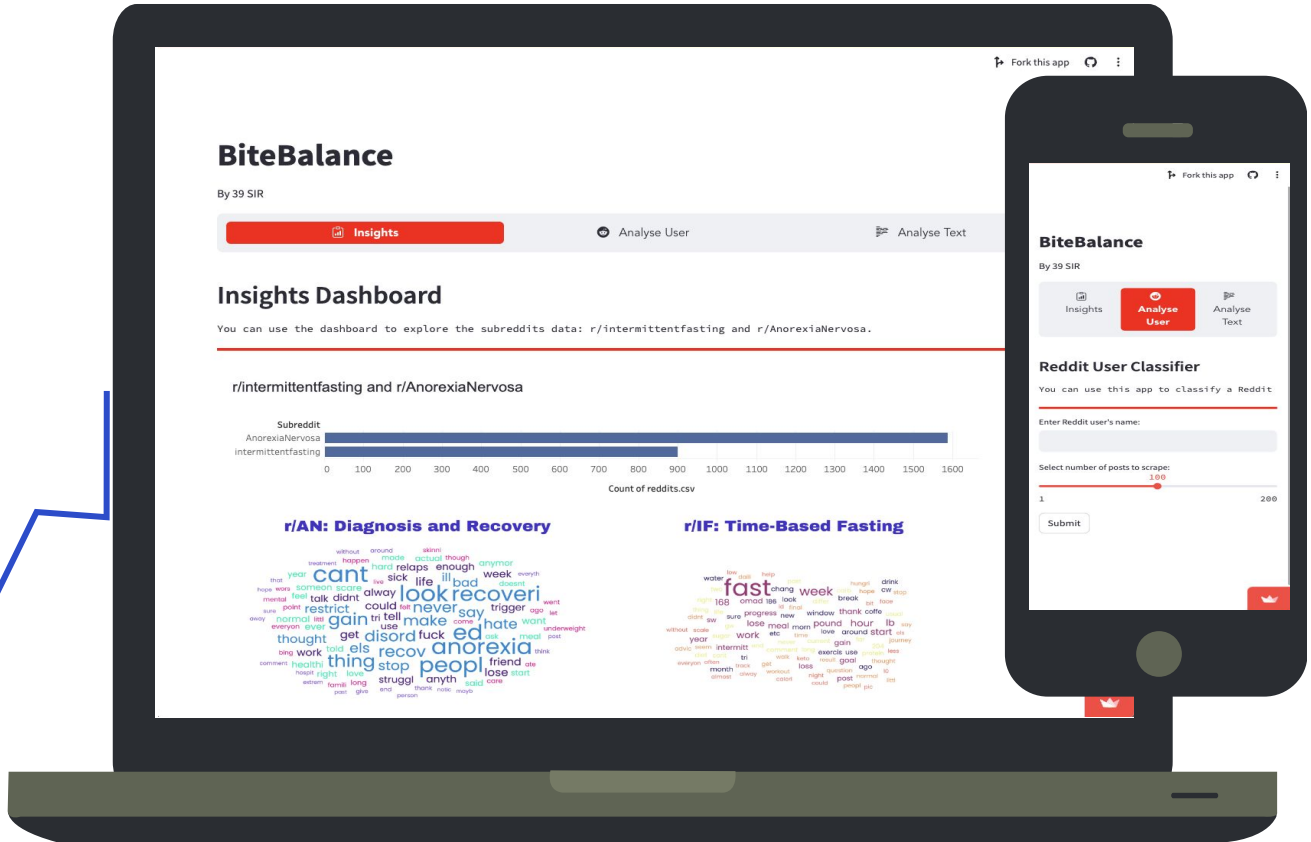
Logistic Regression CountVectorizer	**Multinomial Naive Bayes** CountVectorizer	**XGBoost** CountVectorizer
Logistic Regression N-gram (2,2)	**Multinomial Naive Bayes** N-gram (2,2)	**XGBoost** N-gram (2,2)
Logistic Regression TF-IDF	**Multinomial Naive Bayes** TF-IDF	**XGBoost** TF-IDF

Deployed: Multinomial CV

* All models have been optimised by tuning their hyperparameters before being run.
* Scores closer to 1 the better

Model	Vectorizer	Accuracy (train)	Accuracy (test)	Precision (test)	Recall (test)	F1 (test)	ROC (test)	Execution time (ms)
Logistic Regression	Count Vectorizer	0.96	0.92	0.94	0.9	0.92	0.92	10.33
XGBoost	N-gram (2,2)	0.85	0.79	0.88	0.64	0.74	0.78	810.11
Logistic Regression	N-gram (2,2)	0.94	0.81	0.93	0.66	0.77	0.80	49.11
Multinomial NB	N-gram (2,2)	0.98	0.90	0.90	0.89	0.89	0.9	1.88
Logistic Regression	TF-IDF	0.99	0.93	0.93	0.91	0.92	0.93	9.71
Multinomial NB	TF-IDF	0.96	0.93	0.91	0.94	0.93	0.93	1.03
XGBoost	TF-IDF	0.98	0.93	0.91	0.94	0.93	0.93	794.21
XGBoost	Count Vectorizer	0.97	0.94	0.92	0.95	0.94	0.94	165.05
Multinomial NB	**Count Vectorizer**	**0.96**	**0.95**	**0.95**	**0.94**	**0.94**	**0.95**	**0.96**

App Demo


Fork this app
BiteBalance
By 39 SIR
Insights
Analyse User
Analyse Text
Insights Dashboard
You can use the dashboard to explore the subreddits data: r/intermittentfasting and r/AnorexiaNervosa.
r/intermittentfasting and r/AnorexiaNervosa
Subreddit
AnorexiaNervosa
intermittentfasting
0 100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 1600
Count of reddits.csv
r/AN: Diagnosis and Recovery
r/IF: Time-Based Fasting
Fork this app
BiteBalance
By 39 SIR
Insights
Analyse User
Analyse Text
Reddit User Classifier
You can use this app to classify a Reddit
Enter Reddit user's name:
Select number of posts to scrape:
100
1
200
Submit

Conclusion

Deployment of **Multinomial NB** model with **0.95** accuracy

r/IF Posts

- **Shorter** text posts
- Content on **time-based** fasting
- Food keywords - e.g. **black coffee**

r/AN Posts

- **Longer** text posts
- Discussion of **mental health** concerns
- Food keywords - e.g. **ice cream**

BiteBalance Recommender: Products and Services

Intermittent Fasting

- **Gym Equipment** by Colony.co
- **Protein Supplements** by Protein la Muffins
- **Grooming Classes** by 79 @ Anson

Anorexia Nervosa

- **Stress Management** by 40 SIB Wellness

Next Steps

Share

Share with your clients the differences between intermittent fasting and symptoms of anorexia

Investigate

Keep an eye out for clients practicing intermittent fasting concerned about mental health

Review

39 SIR will share initial findings after 30 days to determine market size

BiteBalance v2

Cloud-Based Data Storage

Protection of client data as user-base grows

Social Media Integration

Scrape platforms that have lesser moderation than Reddit

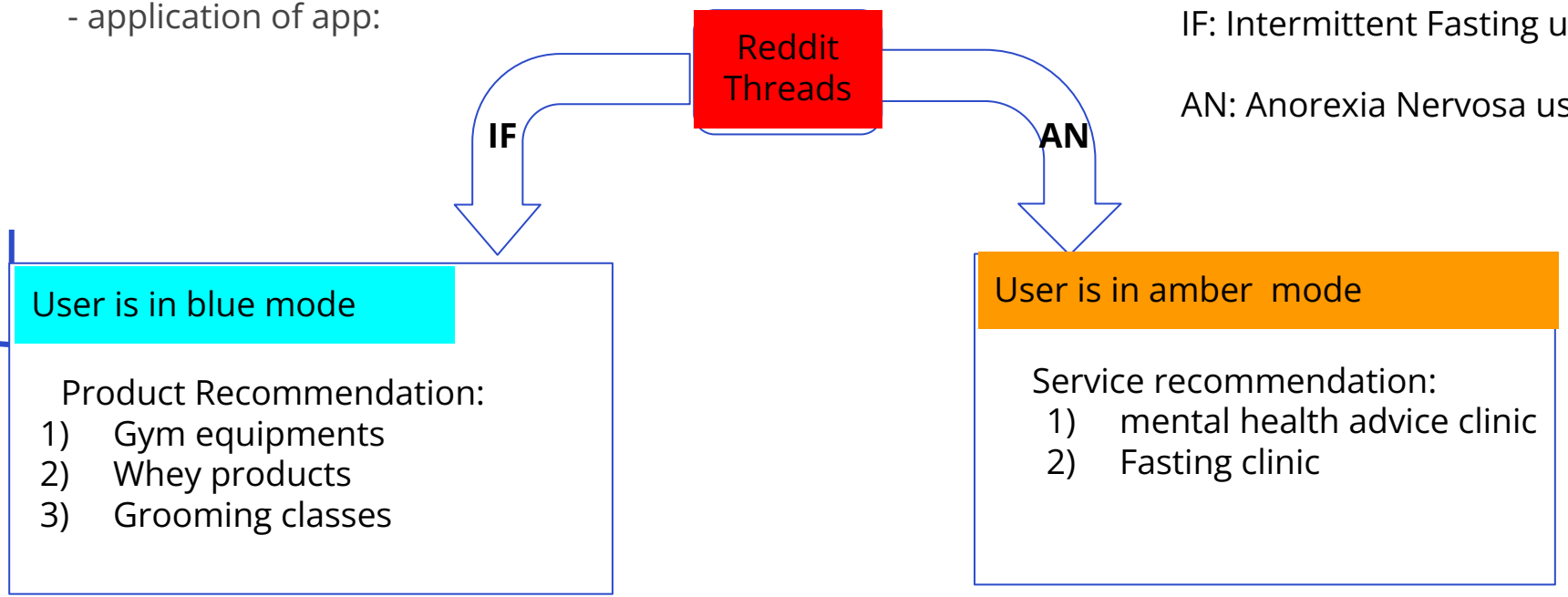
Image Recognition

Training of the image-only posts from Reddit on our model

Thank you.

Conclusion and Recommendations

- application of app:



IF: Intermittent Fasting user

AN: Anorexia Nervosa user

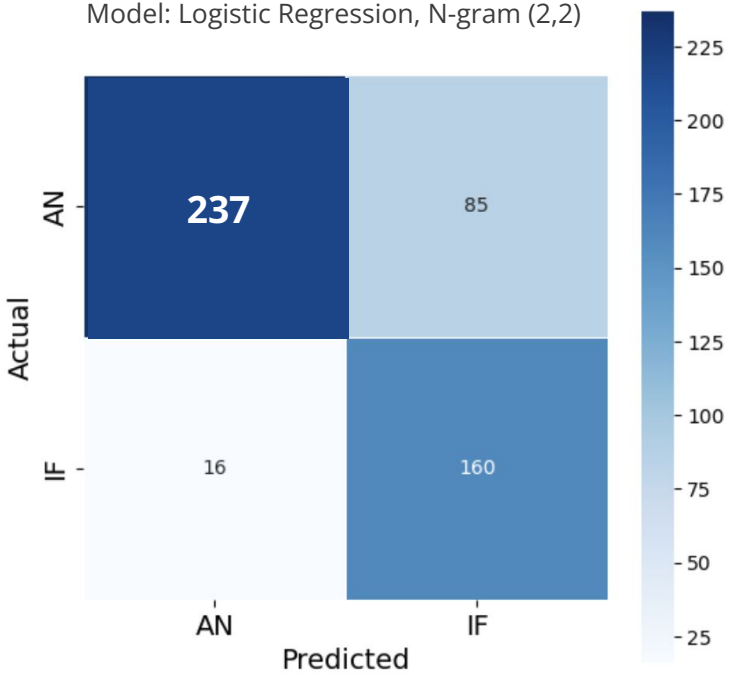
How we increase model scores

Feature engineering
Looked at other models

Performance increased by 0.15

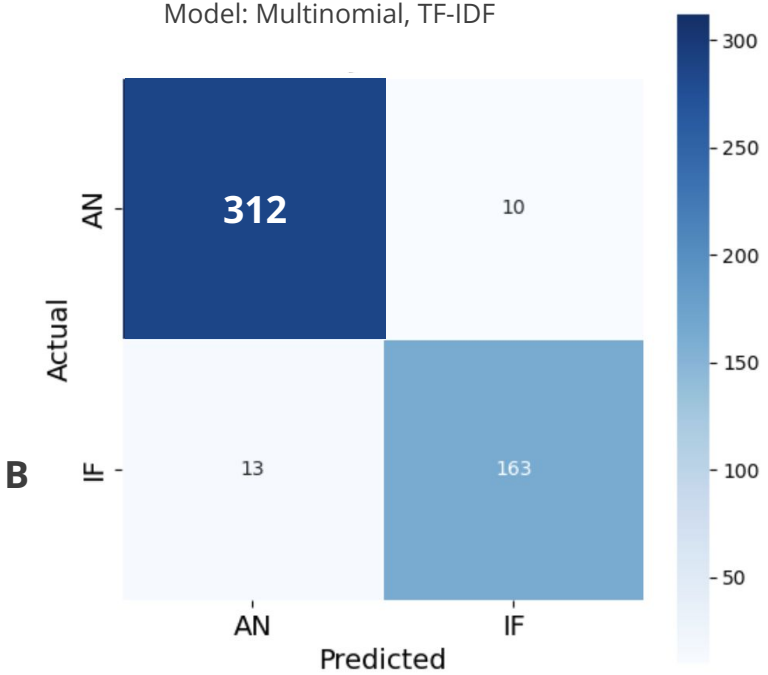
Baseline Model:
Accuracy of **0.80**

Model: Logistic Regression, N-gram (2,2)



Deployed Model:
Accuracy of **0.95**

Model: Multinomial, TF-IDF

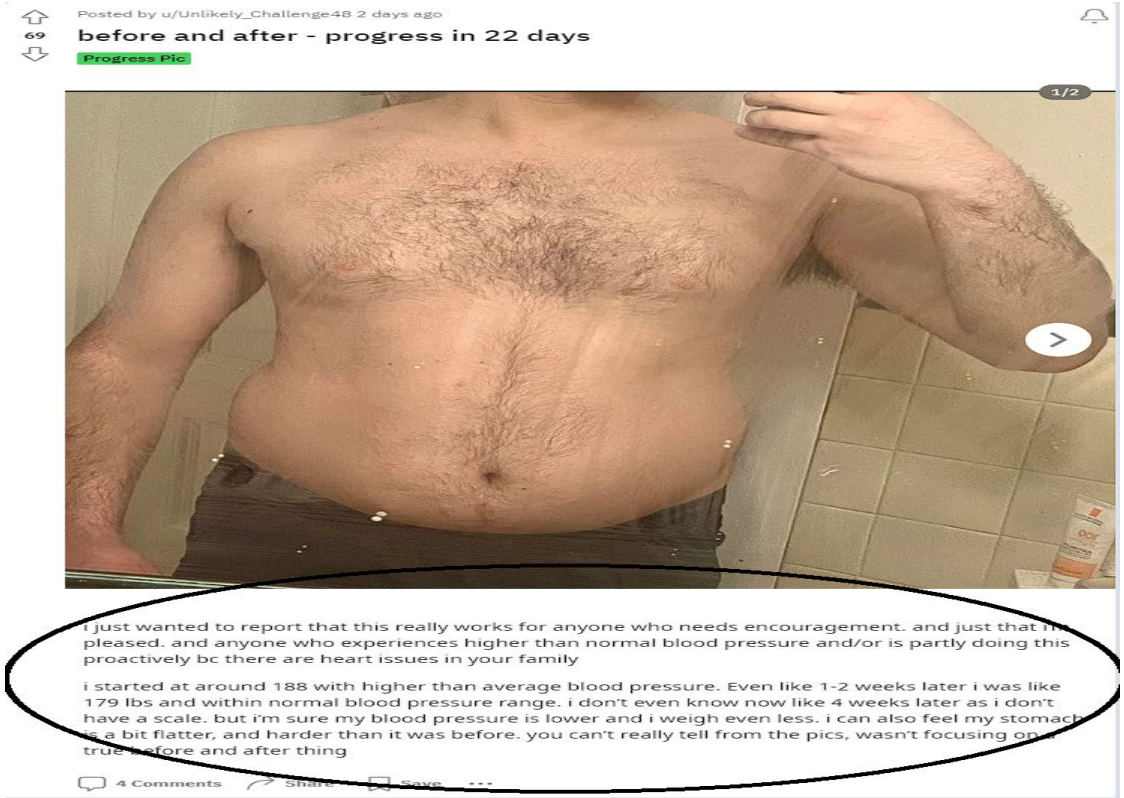


B

Important words

Show graph of top 10 important words

Reddit Habits: Text in Title Only



Future Work

- link to a private cloud when customer base increases
-
- Identify users from AN group with suicidal tendencies and hence preventing suicides
-
- Observe trends in these two subreddit groups
- Look into other online users platforms

Conclusion and Recommendations

Accuracy of model

- Using Multinomial NB TF-IDF, the accuracy is 0.985.

Key insights

- The keywords in reddits threads help us identify which group(IF or AN) the user belongs to
-

Business recommendations and applications:
- After being identified as IF or AN, the user is lead to suitable products and services.

Resources

Disordered Eating in Singapore

Increase in Diagnoses of Eating Disorders

Only 1.6 per cent of those screened positive for a eating disorder reported that they were receiving treatment

https://www.stra
rum-time-to-add
ating-disorders

https://www.straitstimes.com/singapore/new-programme-supports-parents-caring-for-children-with-eating-disorders-as-number-of-cases-rises

Problem Statement

Who We Are:

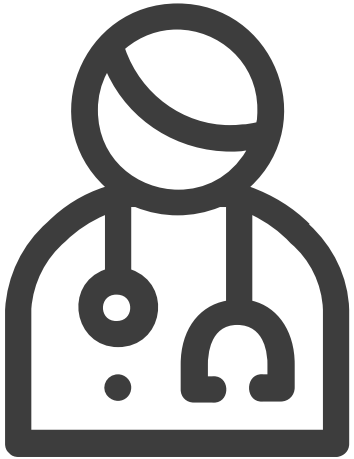
- 39 SIR Healthcare

Who You Are:

- In-House Health & Wellness Dept
- Business Partners

Goal:

- Successfully classify text-based posts into 1 of 2 classes
- Predictor, self-diagnostic for self and others, tool



Problem Statement

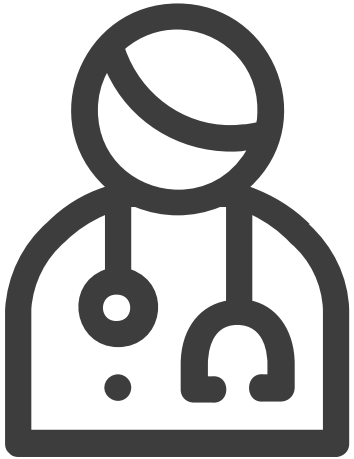
Who We Are:

- A healthcare group with an online presence
- We aim to identify the two group of users- Intermittent fasting and Anorexia Nervosa from inputs of these two subreddit groups

Target Customer: In-house services and our business partners

Goal:

- Successfully classify text-based posts into 1 of 2 classes
- Predictor, self-diagnostic for self and others, tool



https://www.8days.sg/entertainment/asian/mei-diet-dangerous-world-tour-769256

Intermittent Fasting by Celebrities

A-Mei, 50, Said To Have Lost 10kg For Her World Tour By Sticking To A Diet Doctors Are Telling People Not To Follow

Apparently her drastic meal plan can cause serious health problems in the long run.

Ilsa Chan
05 Apr 2023 at 16:30

A-Mei | diet

ow Terry Crews Uses Intermittent Fasting and Training to Sta

The jacked 'Brooklyn Nine-Nine' star has stuck to the same gym game plan for decades

BY EBENEZER SAMUEL, C.S.C.S. AND MELISSA MATTHEWS PUBLISHED: APR 15, 2019

SAVE ARTICLE

ILYA S. SAVENOK // GETTY IMAGES

There are plenty of ways to mix up your workout. Terry Crews isn't into any of them.

How do you build an action hero body like the one that Crews has sported in everything from Brooklyn Nine-Nine to The Expendables to Deadpool 2 and Arrested Development? By sticking to tried-and-true exercises for two

Trends

Don’t have to include all, just pick those you can find:

- Newspaper reports on the trends
- Statistics on IF/AN people - if these are becoming more popular these days?
- Any increasing trend in people practising IF becoming AN?

Contents

Background and Problem Statement

- similarities and differences
- health benefits of IF vs poor mental health

Background (1st Revision)

Definition used:

Intermittent Fasting (IF):

A popular health trend involving controlled periods of fasting and eating, often for weight management and wellness.

Anorexia Nervosa (AN):

A severe mental health disorder characterized by extreme dietary restriction, fear of gaining weight, and a distorted body image.

Statistics:

Research indicates that a portion of individuals who have experimented with Intermittent Fasting may develop Anorexia Nervosa. Exact numbers vary, but it's essential to highlight this potential risk. (INSERT NUMBERS HERE)

https://www.channelnewsasia.com/commentary/lose-weight-intermittent-fasting-time-health-3202436
https://www.channelnewsasia.com/cnainsider/does-intermittent-fasting-weight-loss-7-things-you-should-know-312701
https://cnalifestyle.channelnewsasia.com/wellness/orthorexia-nervosa-eating-disorder-obsession-healthy-food-364411
https://www.channelnewsasia.com/cna-insider/mental-health-what-schools-doing-help-students-2655911
https://www.straitstimes.com/life/eating-disorders-on-the-rise-amid-pandemic-related-stress

Background and Problem Statement

Set up the scenario - who are we, who you are

We are a group of healthcare specialist web designers and our audiences are representative for healthcare and mental wellbeing companies

Explain business need of project, project goals, potential impact

- IF: widespread popularity of intermittent fasting for health benefits
- AN: increase in mental health awareness in SG
- AN: high stress / prevalence of mental health conditions in SG
- distinguishing AN from IF for awareness and early intervention

https://www.channelnewsasia.com/commentary/lose-weight-intermittent-fasting-time-health-3202436
https://www.channelnewsasia.com/cnainsider/does-intermittent-fasting-weight-loss-7-things-you-should-know-312701
https://cnalifestyle.channelnewsasia.com/wellness/orthorexia-nervosa-eating-disorder-obsession-healthy-food-364411
https://www.channelnewsasia.com/cna-insider/mental-health-what-schools-doing-help-students-2655911
https://www.straitstimes.com/life/eating-disorders-on-the-rise-amid-pandemic-related-stress

Problem Statement (1st Revision)

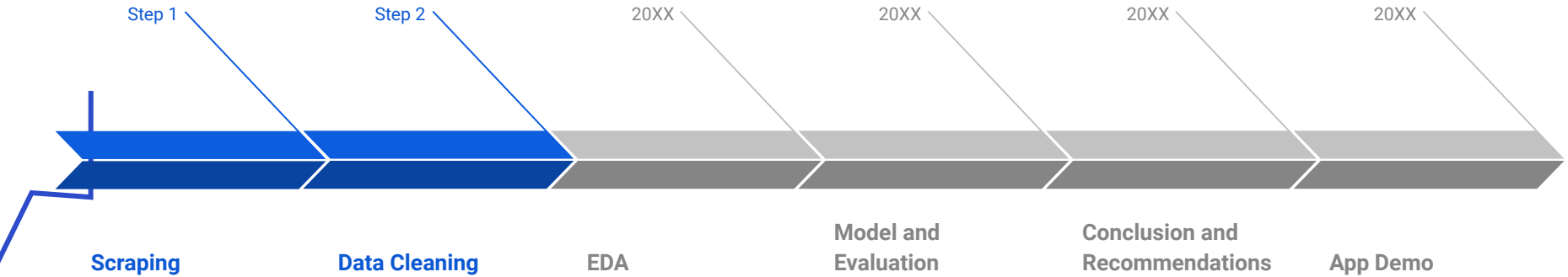
We are a healthcare specialist web designer group aiming to address the needs of healthcare and mental wellbeing companies. Our goal is to enhance their online presence and user experience.

Key Questions:

- How can we optimize web design for healthcare and mental wellbeing industries?
- What specific challenges do these companies face in their online presence?
- How can we effectively convey their services and mission through web design?

Workflow

Overview of data collection, cleaning, and analysis process



Scraping

Data extracted from reddit

Data Cleaning

Column selection

EDA

Lorem ipsum dolor sit amet, consectetur adipiscing.

Model and Evaluation

Lorem ipsum dolor sit amet, consectetur adipiscing.

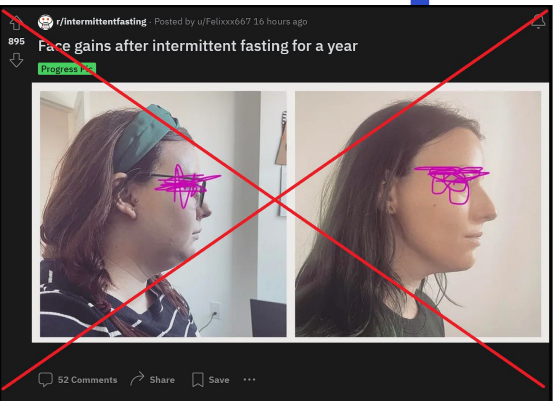
Conclusion and Recommendations

Lorem ipsum dolor sit amet, consectetur adipiscing.

App Demo

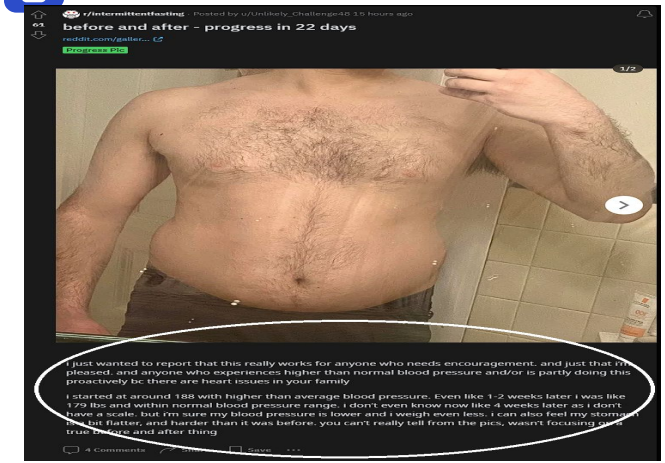
Lorem ipsum dolor sit amet, consectetur adipiscing.

Scraping

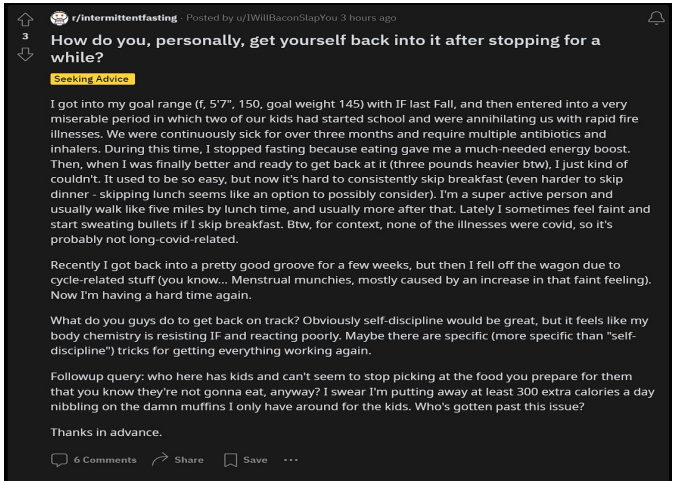


Worst of all: No text to extract at all. Only title.

Used praw in python to scrape reddit data



Seemingly ok: While there's an image, there's also text to extract.



Best one: This post has text for extraction..

Images taken from:

- https://www.reddit.com/r/intermittentfasting/comments/16stab9/before_and_after_progress_in_22_days/
- https://www.reddit.com/r/intermittentfasting/comments/16ssen3/face_gains_after_intermittent_fasting_for_a_year/
- https://www.reddit.com/r/intermittentfasting/comments/16tb590/how_do_you_personally_get_yourself_back_into_it/

Cleaning (Dataframe)

Cleaning of dataframe (columns: Title, Post Text, ID, Score, Total Comments, Post URL, Subreddit, Post Type, Time uploaded)

1. Removed Irrelevant Posts:
    a. Eliminated posts that were removed or deleted to ensure data accuracy.
2. Deduplicated Data:
    a. Removed duplicate posts to streamline the dataset.
3. Standardized Column Headers:
    a. Converted column headers to lowercase and used underscores (_) between words (lowercase, snake_case) for consistency and ease of reference.
4. Handled Missing Data:
    a. Addressed null values in the dataset to avoid data inconsistencies.
5. Merged 'Title' and 'Post' Columns:
    a. Combined information from the 'Title' and 'Post' columns into a new column named 'title_and_text' for improved data analysis and interpretation.

Cleaning (Text)

Cleaning of texts in 'title_&_post':

1. Remove Punctuation:
    a. Eliminate punctuation marks to simplify the text and aid in subsequent analysis.
2. Remove Stopwords:
    a. Exclude common stopwords (e.g., 'the,' 'and,' 'is') to extract meaningful keywords and phrases.
3. Lowercase:
    a. Convert all text to lowercase for consistency and easier comparisons.
4. Drop Hyperlinks:
    a. Remove hyperlinks from the text to ensure the analysis focuses on the actual content."

Trigrams

Intermittent Fasting	Anorexia
juic smoke etc	diagnos anorexia nervosa
sort new newer	tw tw tw
type fast water	bing restrict cycl
daili fast checkin	hate hate hate
fast hour etc	your sick enough
fast note far	photo ice cream
often throughout sort	mental health servic
fast water juic	low blood sugar
concern insight share	suffer anorexia nervosa
context fast etc	theyr sick enough

Models Deployed

Total: 9 models

Models	Vectorizer
Logistics Regression	CountVectorizer
Logistics Regression	N-gram (2,2)
Logistics Regression	TF-IDF
Multinomial Naive Bayes	CountVectorizer
Multinomial Naive Bayes	N-gram (2,2)
Multinomial Naive Bayes	TF-IDF
XGBoost	CountVectorizer
XGBoost	N-gram (2,2)
XGBoost	TF-IDF

Deployed: Multinomial CV

* All models have been optimised by tuning their hyperparameters before being run.

Model	Vectorizer	Accuracy (train)	Accuracy (test)	Precision (test)	Recall (test)	F1 (test)	ROC (test)	Mean (all scores)
Baseline: Logistic Regression	N-gram (2,2)	0.998	0.920	0.902	0.890	0.896	0.914	0.920
Multinomial	**CountVector**	**0.954**	**0.957**	**0.973**	**0.916**	**0.945**	**0.950**	**0.950**
Multinomial	N-gram (2,2)	0.954	0.957	0.973	0.916	0.944	0.950	0.949
Multinomial	TF-IDF	0.985	0.952	0.979	0.897	0.936	0.942	0.949
Logistic Regression	TF-IDF	0.992	0.947	0.972	0.890	0.929	0.937	0.945
XGBoost	N-gram (2,2)	0.996	0.940	0.958	0.884	0.919	0.930	0.938
XGBoost	CountVector	0.996	0.940	0.958	0.884	0.919	0.930	0.938
XGBoost	TF-IDF	0.996	0.935	0.951	0.877	0.913	0.924	0.933
Logistic Regression	N-gram (2,2)	0.999	0.920	0.902	0.890	0.896	0.914	0.920

Model Scores

* All models have been optimised by tuning their hyperparameters before being run.

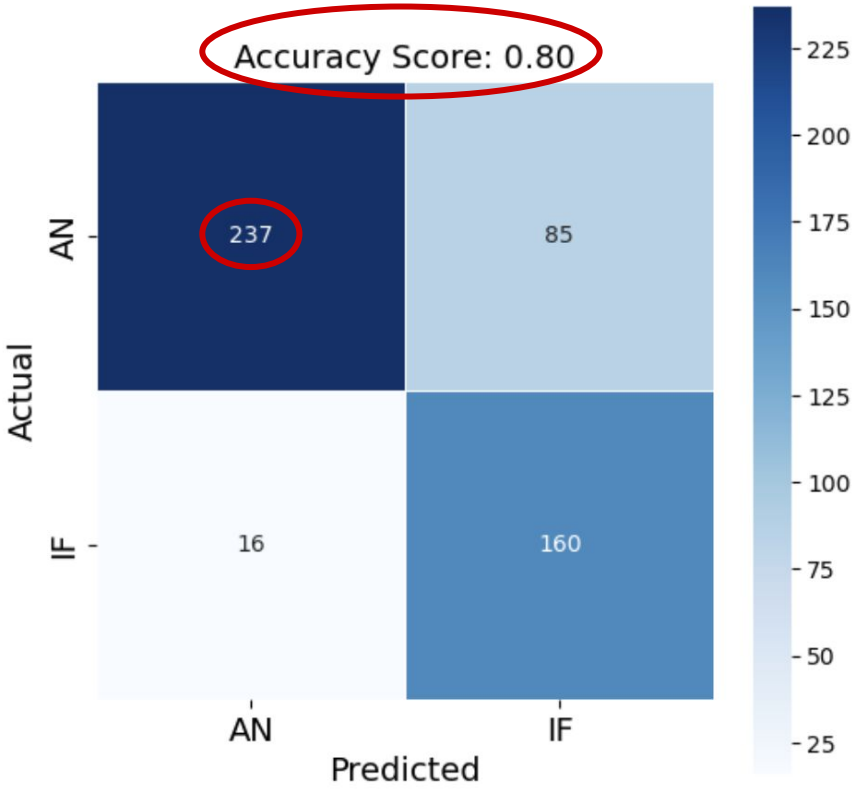
Model	Vectorizer	Accuracy (train)	Accuracy (test)	Precision (test)	Recall (test)	F1 (test)	ROC (test)	Mean (all scores)
Baseline: Logistic Regression	N-gram (2,2)	0.998	0.920	0.902	0.890	0.896	0.914	0.920
Multinomial	**CountVector**	**0.954**	**0.957**	**0.973**	**0.916**	**0.945**	**0.950**	**0.950**
Multinomial	N-gram (2,2)	0.954	0.957	0.973	0.916	0.944	0.950	0.949
Multinomial	TF-IDF	0.985	0.952	0.979	0.897	0.936	0.942	0.949
Logistic Regression	TF-IDF	0.992	0.947	0.972	0.890	0.929	0.937	0.945
XGBoost	N-gram (2,2)	0.996	0.940	0.958	0.884	0.919	0.930	0.938
XGBoost	CountVector	0.996	0.940	0.958	0.884	0.919	0.930	0.938
XGBoost	TF-IDF	0.996	0.935	0.951	0.877	0.913	0.924	0.933
Logistic Regression	N-gram (2,2)	0.999	0.920	0.902	0.890	0.896	0.914	0.920

Model Scores

* All models have been optimised by tuning their hyperparameters before being run.

Model	Vectorizer	Accuracy (train)	Accuracy (test)	Precision (test)	Recall (test)	F1 (test)	ROC (test)	Mean (all scores)
Baseline: Logistic Regression	N-gram (2,2)	0.998	0.920	0.902	0.890	0.896	0.914	0.920
Chosen: Multinomial	**CountVector**	**0.954**	**0.957**	**0.973**	**0.916**	**0.945**	**0.950**	**0.950**
Multinomial	N-gram (2,2)	0.954	0.957	0.973	0.916	0.944	0.950	0.949
Multinomial	TF-IDF	0.985	0.952	0.979	0.897	0.936	0.942	0.949
Logistic Regression	TF-IDF	0.992	0.947	0.972	0.890	0.929	0.937	0.945
XGBoost	N-gram (2,2)	0.996	0.940	0.958	0.884	0.919	0.930	0.938
XGBoost	CountVector	0.996	0.940	0.958	0.884	0.919	0.930	0.938
XGBoost	TF-IDF	0.996	0.935	0.951	0.877	0.913	0.924	0.933
Logistic Regression	N-gram (2,2)	0.999	0.920	0.902	0.890	0.896	0.914	0.920

Logistic Regression with Ngram (2,2)



Multinomial NB with TF-IDF