FF

FINESSE FIT SOLUTIONS

NLP BINARY TEXT CLASSIFICATION

GYM AND YOGA

FINESSE FIT SOLUTIONS

HEALTH & FITNESS INDUSTRY SINGAPORE

" The Singaporean gyms, health and fitness clubs market is expected to **generate total revenues of $0.2bn in 2021, representing a compound annual growth rate (CAGR) of 3.6% between 2017 and 2021.**
The lockdown and social-distancing measures imposed to contain the ongoing pandemic have prevented gyms and health and fitness club operators from generating revenues from new and existing users.
The shift of the millennial generation to healthier lifestyles, which has brought the importance of physical exercise to the fore, has been the main driver of this market’s expansion over the last decade." - MarketResearch.com

“As a new brand, it meant it was **harder to attract customers and staff**. This, coupled with the shortage of manpower, (meant) we **could not sustain** the business and we made the difficult decision to **wind down.**” - Superfly Studios

PROBLEM STATEMENT

- With the loosening of Covid restrictions, how do Gyms and Yoga Studios plan to attract old & new customers?

- Traditional vs Digital marketing?

  Traditional-
  - **Print Forms**
  - **TV/Radio**

  Digital-
  - **Email**
  - **Social Media**
  - **SEO**
  - **Influencer**

- Lead Time for the public to receive the message?

logo
GET IN
SHAPE
& BE HEALTHY
25% OFF
OUR SERVICES
› PERSONAL TRAINING
› FITNESS PROGRAMS
› DIET GUIDELINES
› YOGA
› AND MORE...
CALL US
1.800.123.45678
7123 Street Name, Landmark Tower, Country Name - 412454
contact@website.com
www.website.com
#fitnessclub
@fitnessclub

Lorem ipsum dolor sit amet, consectetur adipiscing elit. In nulla ligula, **maximus vitae sodales finibus,** pretium et massa.

Fusce sed magna viverra tortor feugiat dictum. Suspendisse Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. In auctor, orci a lacinia imperdiet,

OVERVIEW

- Background Analysis
  - Problems faced by Gym and Yoga Studios.
- Objective
  - Program that can effectively distinguish social media posts based on their topics.
- Process Workflow
  - Steps in creating the program.
- Conclusions/Recommendations

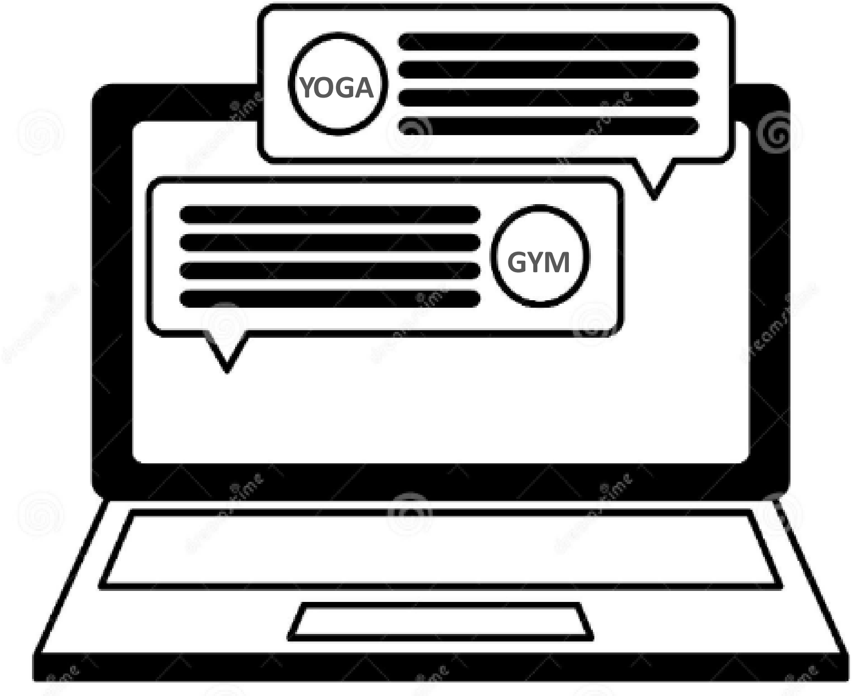
BACKGROUND

- The pandemic and rising inflation costs pummeled the Gym and Yoga industry :
  - Decreased revenue.
  - Decreased membership.
  - Increased competition fror

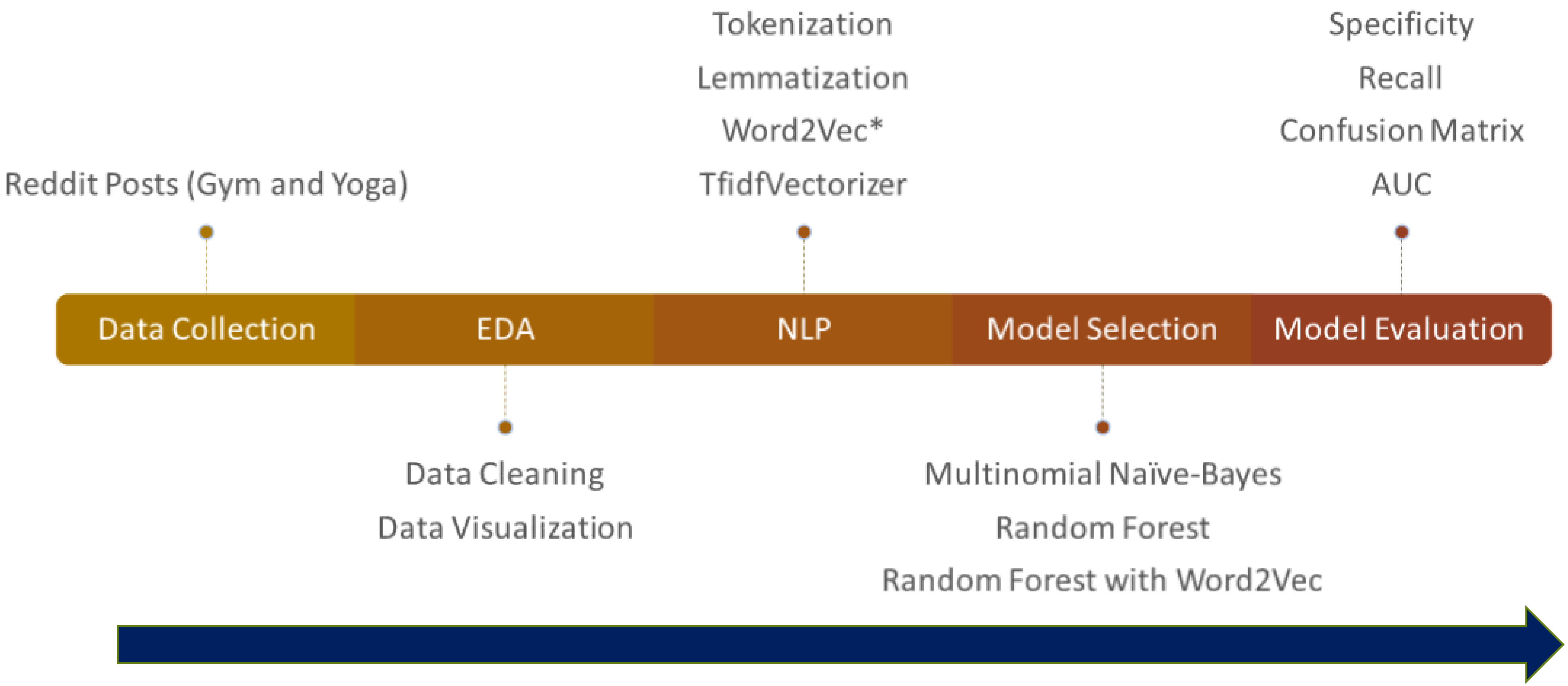
- Loosening of Covid restri
  - People can go Gym and/or
- Fast, efficient and low co crowds back.
  - Reaches the target audience instantly.
  - Can be customized for different niche groups.
  - Most importantly....

OBJECTIVE

- Objective of this project:
  - To get Gym and Yoga related posts from Reddit.
  - NLP to process the texts.
  - Build a classification model that can accurately distinguish between posts related to Yoga and Gym from Reddit.
  - Send targeted marketing materials to appropriate users.



PROCESS WORKFLOW



Data Collection

Reddit Posts (Gym and Yoga)

EDA

Data Cleaning

Data Visualization

NLP

Tokenization

Lemmatization

Word2Vec*

TfidfVectorizer

Model Selection

Multinomial Naïve-Bayes

Random Forest

Random Forest with Word2Vec

Model Evaluation

Specificity

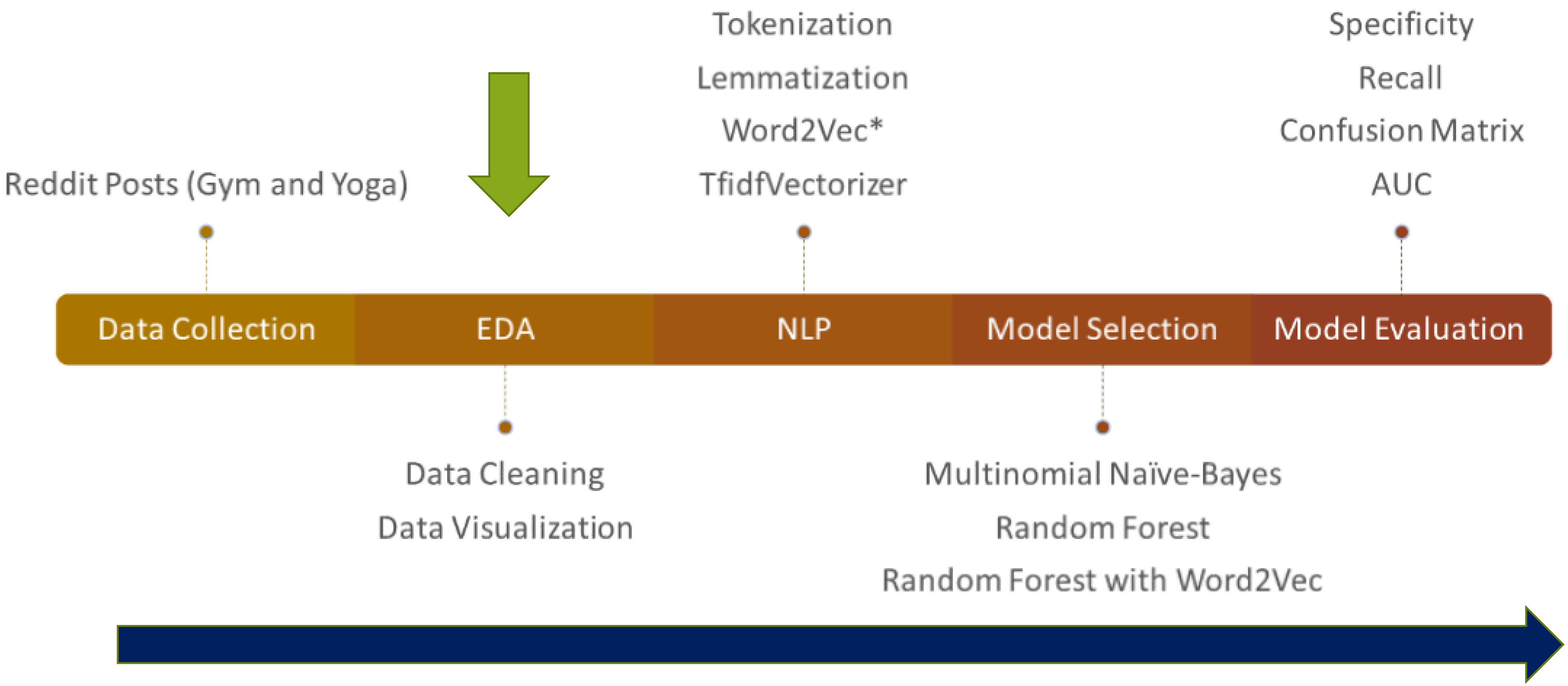
Recall

Confusion Matrix

AUC

FINESSE FIT SOLUTIONS

PROCESS WORKFLOW



Data Collection
Reddit Posts (Gym and Yoga)

EDA
Data Cleaning
Data Visualization

NLP
Tokenization
Lemmatization
Word2Vec*
TfidfVectorizer

Model Selection
Multinomial Naïve-Bayes
Random Forest
Random Forest with Word2Vec

Model Evaluation
Specificity
Recall
Confusion Matrix
AUC

EDA

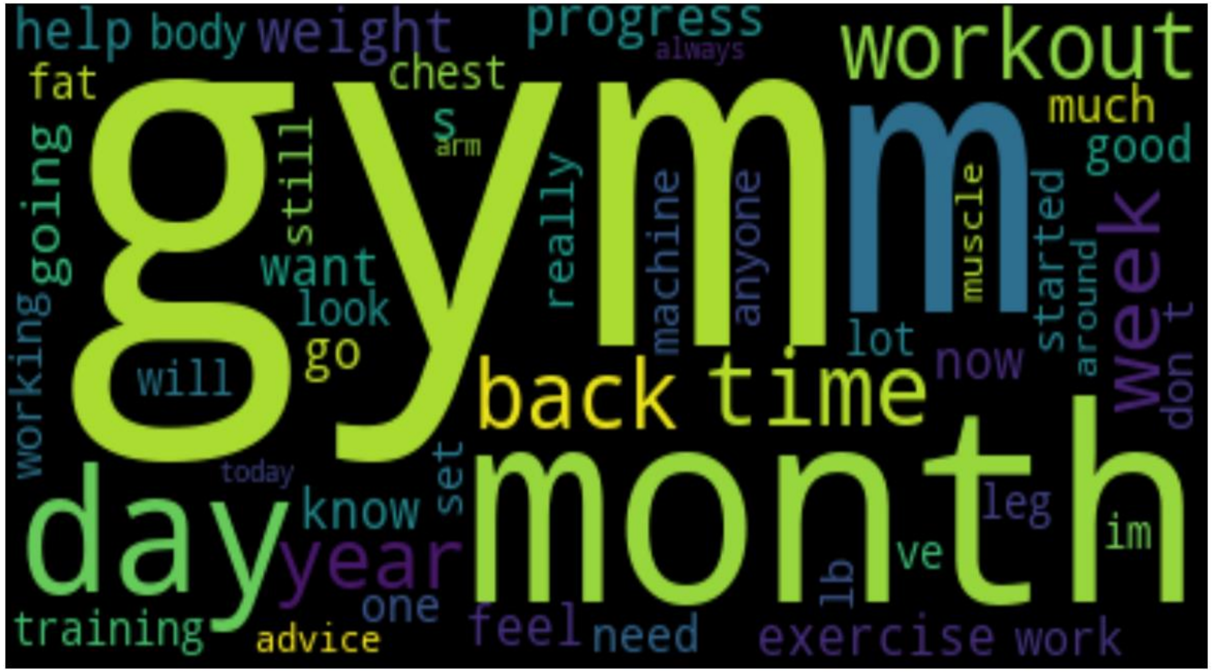
- Data Cleaning
  - Remove null posts.
  - Deduplicate.

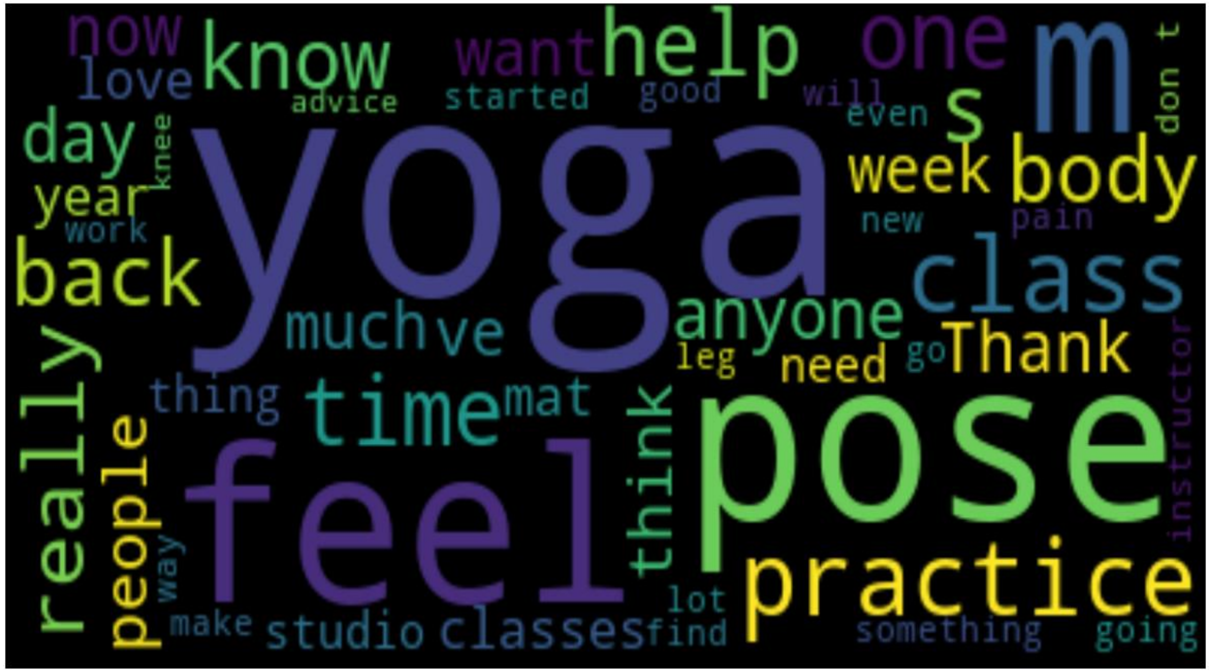
DEDUPLICATION

MISSING POSTS

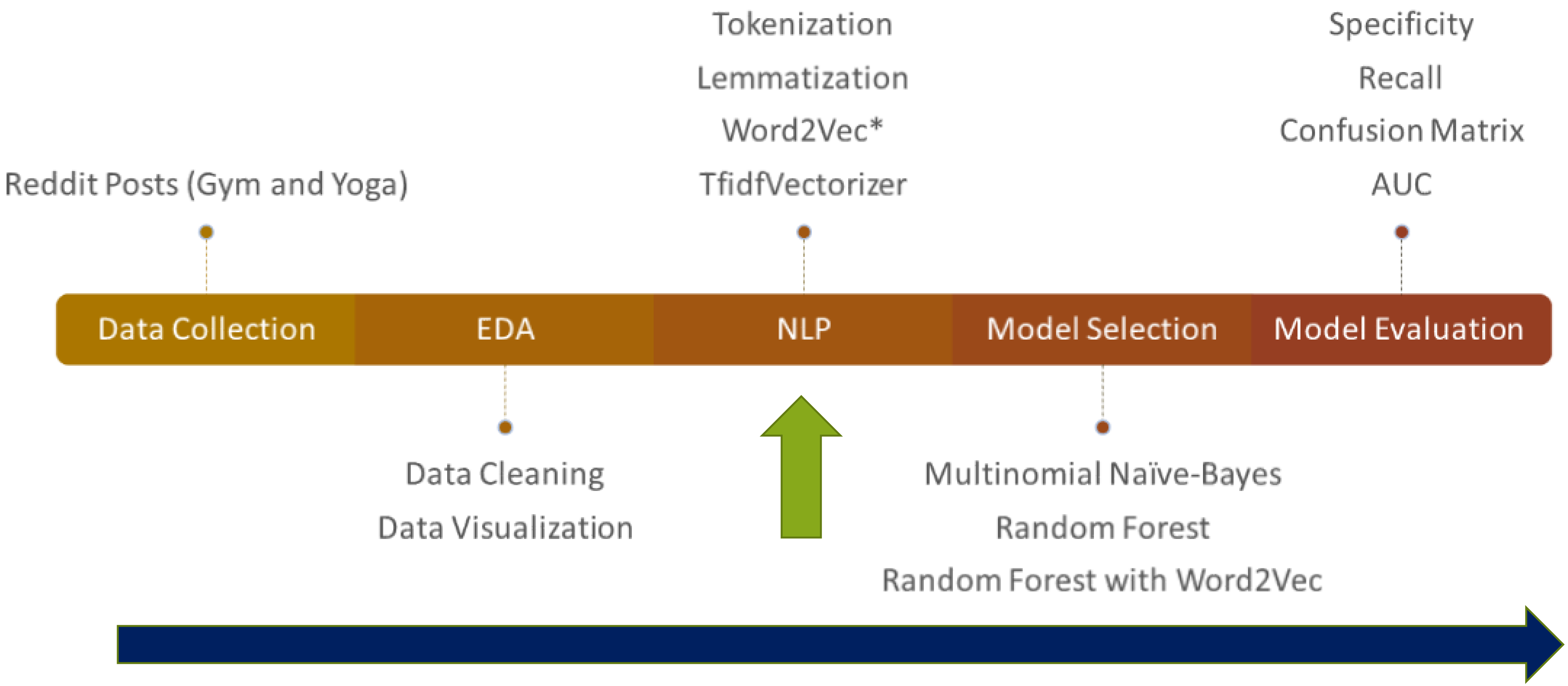
EDA

- Data Visualization (Word Cloud)





PROCESS WORKFLOW



Data Collection: Reddit Posts (Gym and Yoga)

EDA: Data Cleaning, Data Visualization

NLP: Tokenization, Lemmatization, Word2Vec*, TfidfVectorizer

Model Selection: Multinomial Naïve-Bayes, Random Forest, Random Forest with Word2Vec

Model Evaluation: Specificity, Recall, Confusion Matrix, AUC

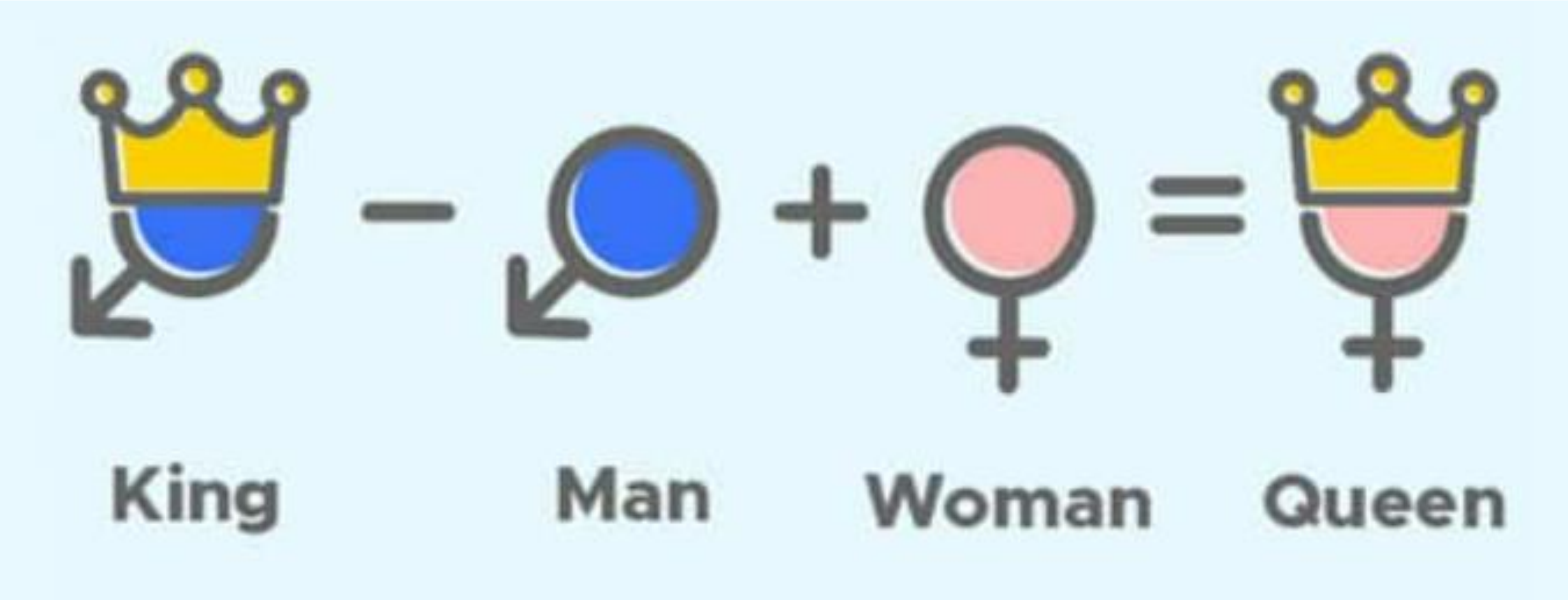
NLP

M gets a

tor

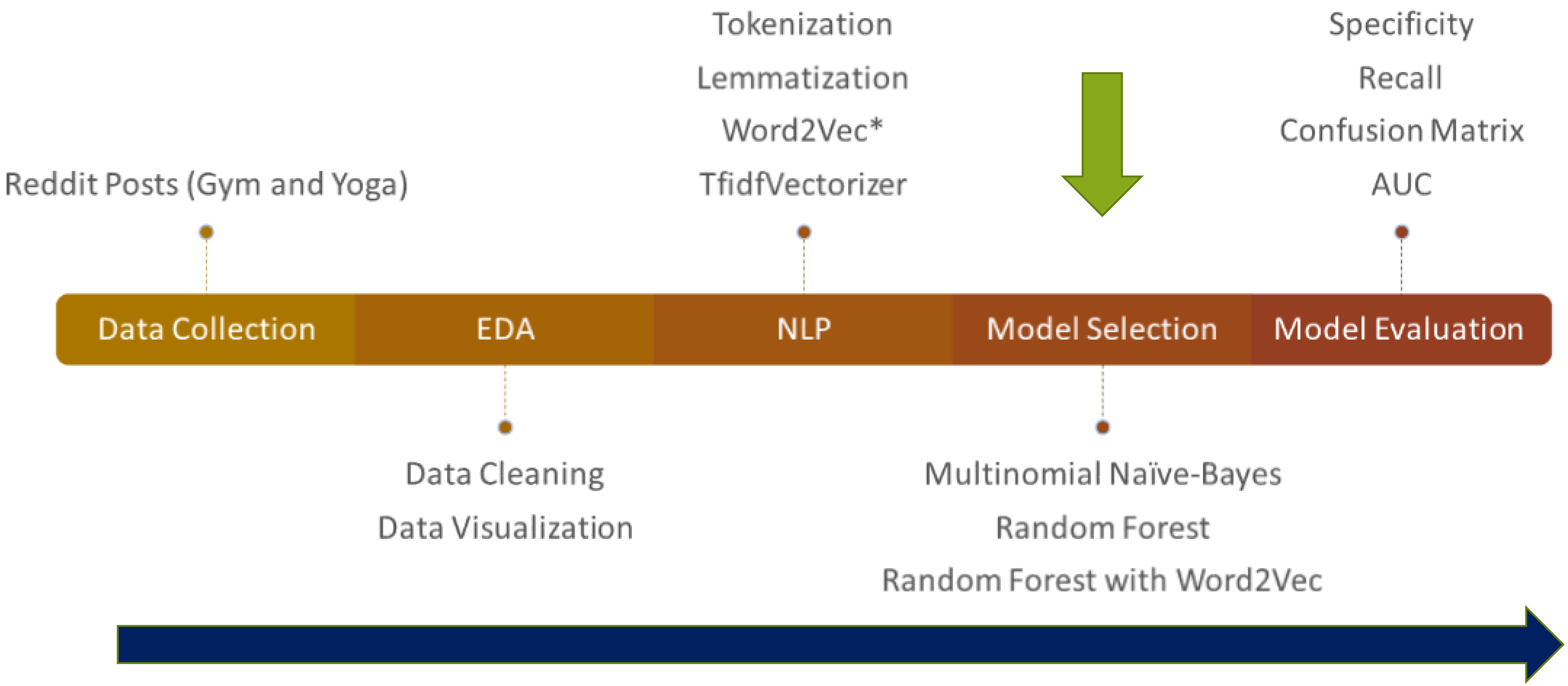
p rds...

ctors



- Word2Vec
  - Analyze the words in each post.
  - Words represented in high-dimensional vector space.
  - Vectors are used to identify terms that are similar based on words used.

PROCESS WORKFLOW



Reddit Posts (Gym and Yoga)

Data Collection

EDA

Data Cleaning

Data Visualization

NLP

Tokenization

Lemmatization

Word2Vec*

TfidfVectorizer

Model Selection

Multinomial Naïve-Bayes

Random Forest

Random Forest with Word2Vec

Model Evaluation

Specificity

Recall

Confusion Matrix

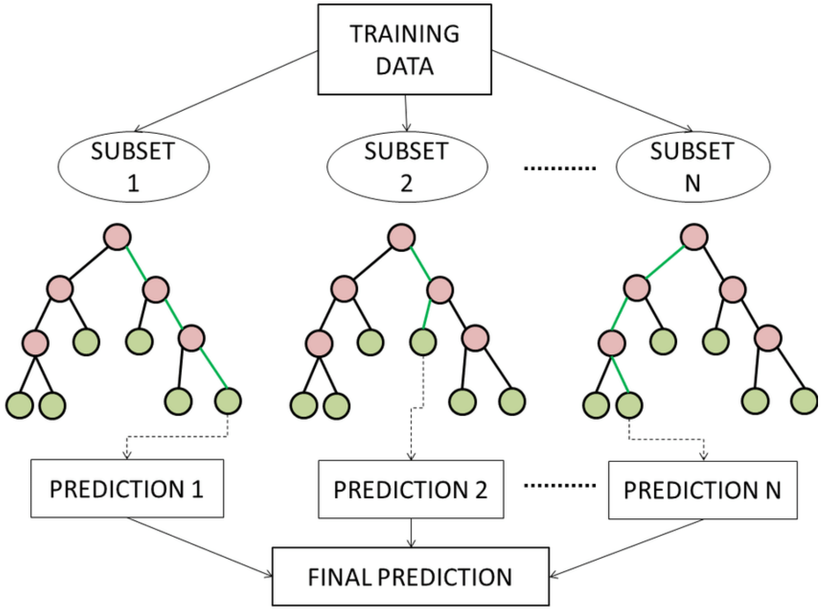
AUC

MODEL SELECTION

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

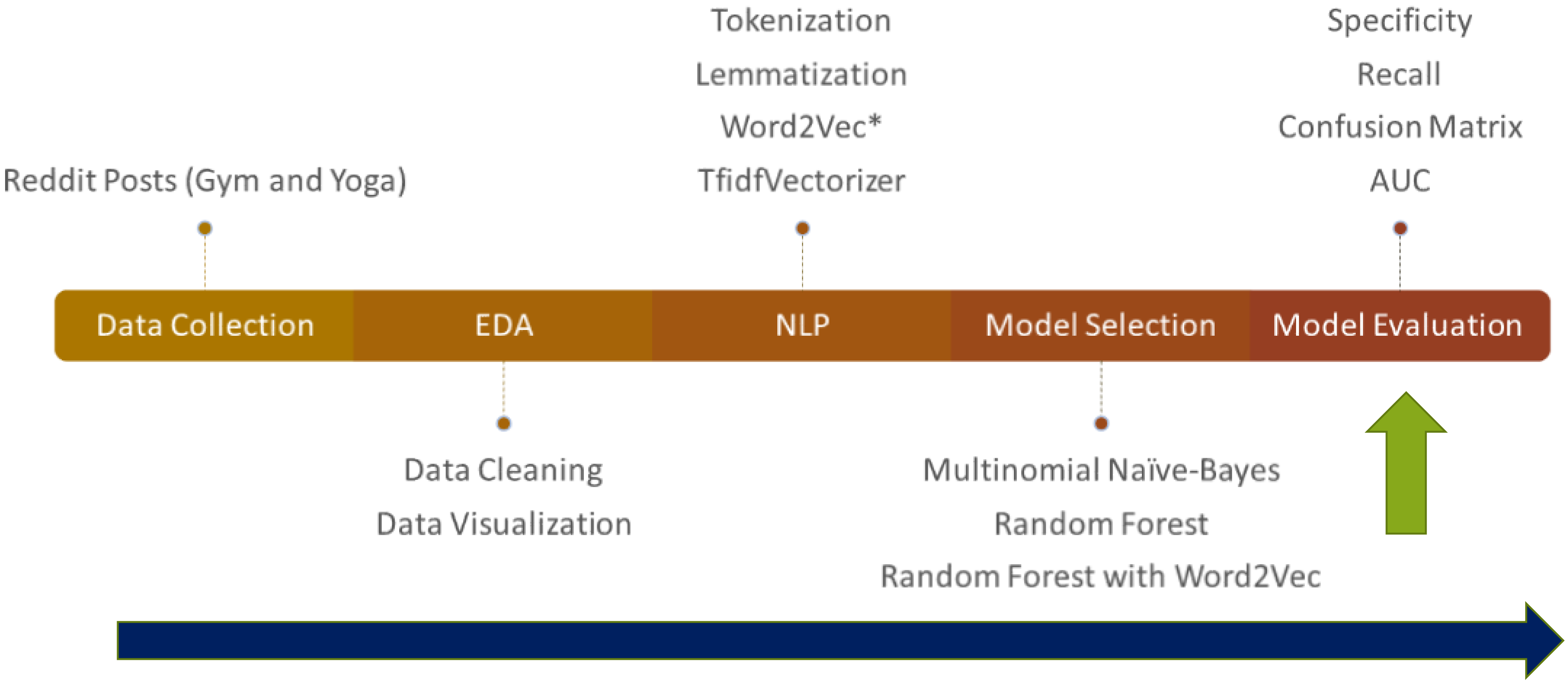
- Naïve-Bayes
  - Probability of a post belong to a category.
  - "Naïve" assumptions.

- Random Forest
  - Multiple Decision Trees.
  - Random select subsets of data and features.
  - Reduce variance of model.
  - Combine each subset prediction to final prediction.



FINESSE FIT SOLUTIONS

PROCESS WORKFLOW



Reddit Posts (Gym and Yoga)

Tokenization
Lemmatization
Word2Vec*
TfidfVectorizer

Specificity
Recall
Confusion Matrix
AUC

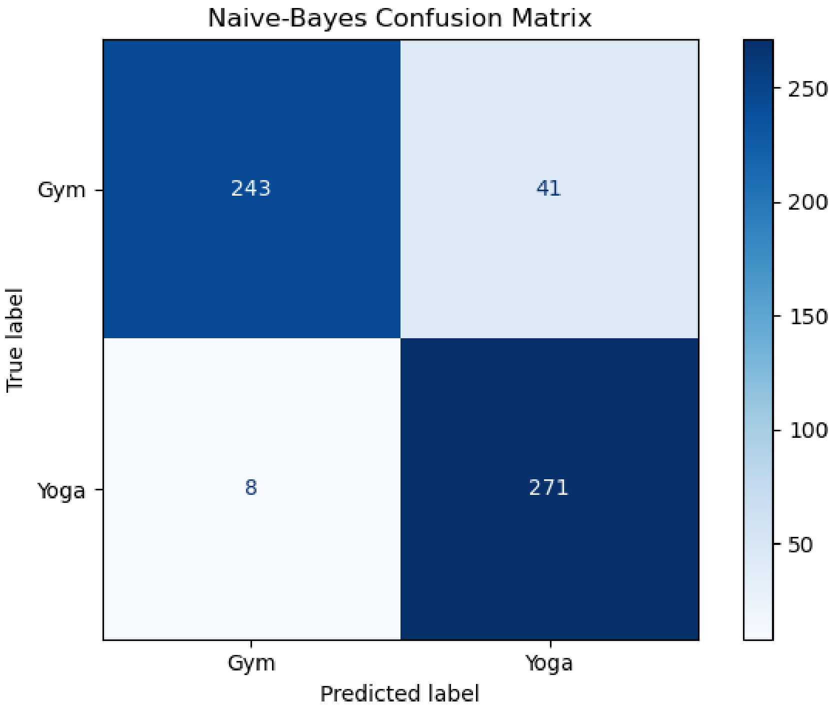
Data Collection | EDA | NLP | Model Selection | Model Evaluation

Data Cleaning
Data Visualization

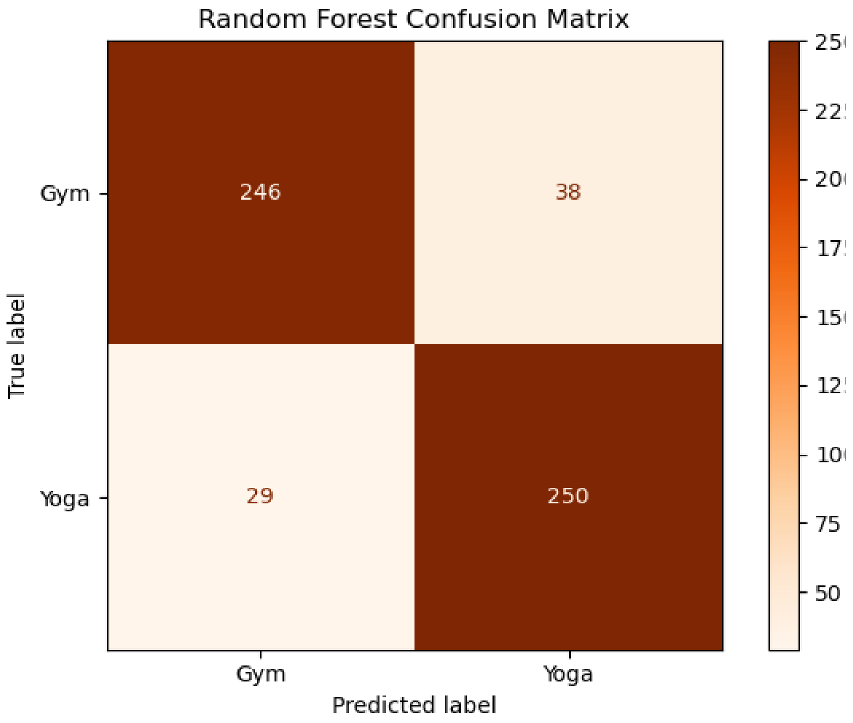
Multinomial Naïve-Bayes
Random Forest
Random Forest with Word2Vec

MODEL EVALUATION

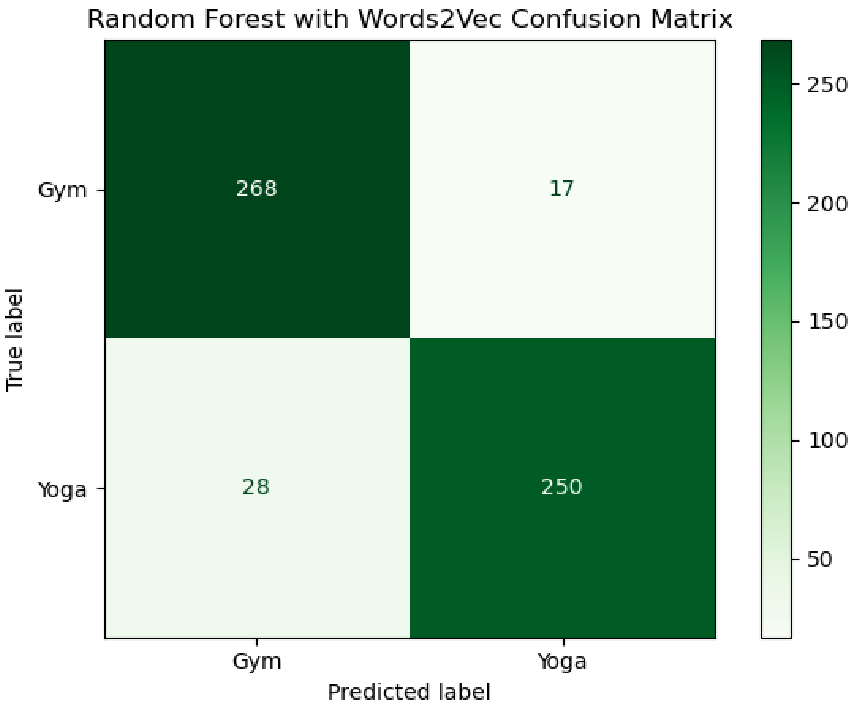
- Confusion Matrix Plots



FP=41 TP=271
FN=8 TN=243



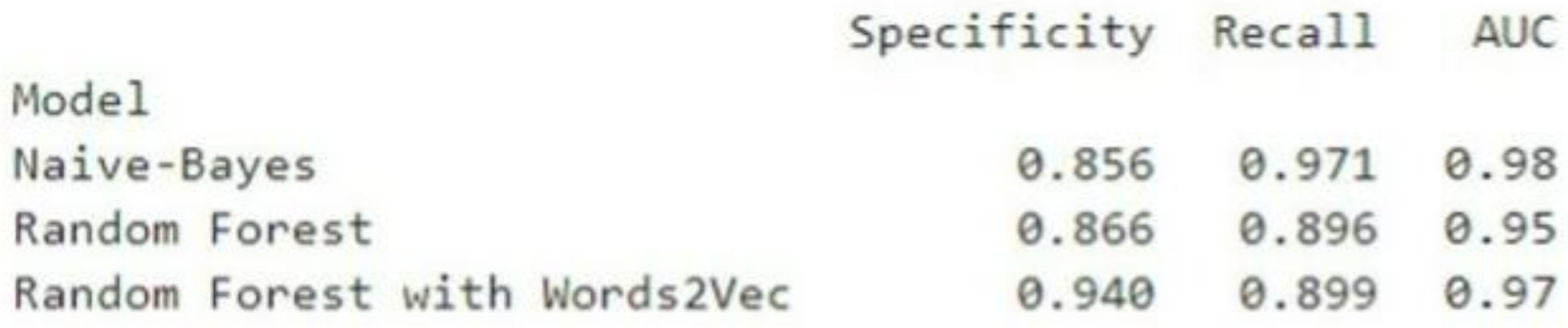
FP=38 TP=250
FN=29 TN=246



FP=17 TP=250
FN=28 TN=268

MODEL EVALUATION

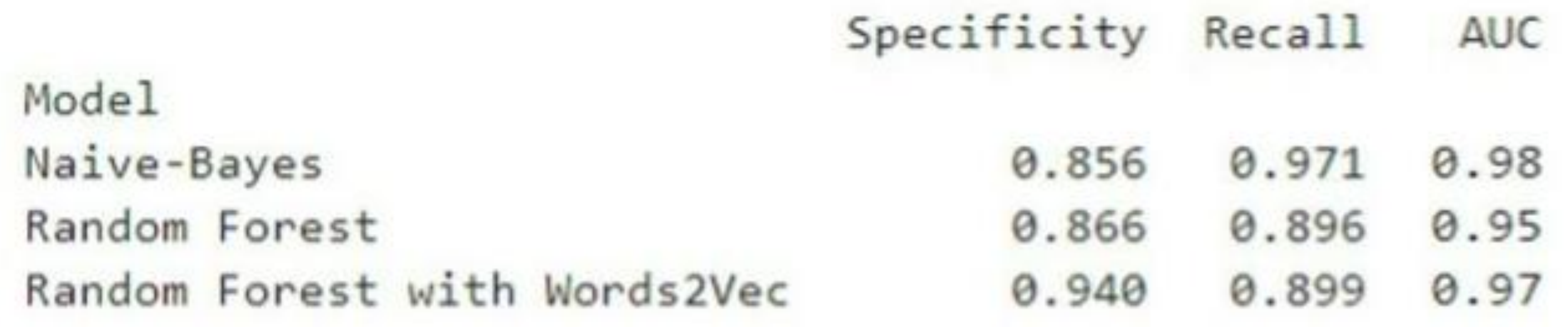
- Classification Model Scores
  - Recall: proportion of Yoga posts correctly identified
  - Specificity: proportion of Gym posts correctly identified.
  - AUC: measure how well the model distinguish between different classes of data.



```
                              Specificity  Recall   AUC
Model
Naive-Bayes                         0.856   0.971  0.98
Random Forest                       0.866   0.896  0.95
Random Forest with Words2Vec        0.940   0.899  0.97
```

CONCLUSIONS

- Random Forest with Word2Vec embedding
  - It has the highest specificity, and a very good AUC.
  - It's recall is in between Naïve-Bayes and Random Forest.
  - This model is the best performing model among the three models evaluated.
  - Combination of Word2Vec ability to seek **semantic relationships between words** and Random Forest ability to **reduce overfitting**.
  - Can effectively distinguish between different posts.



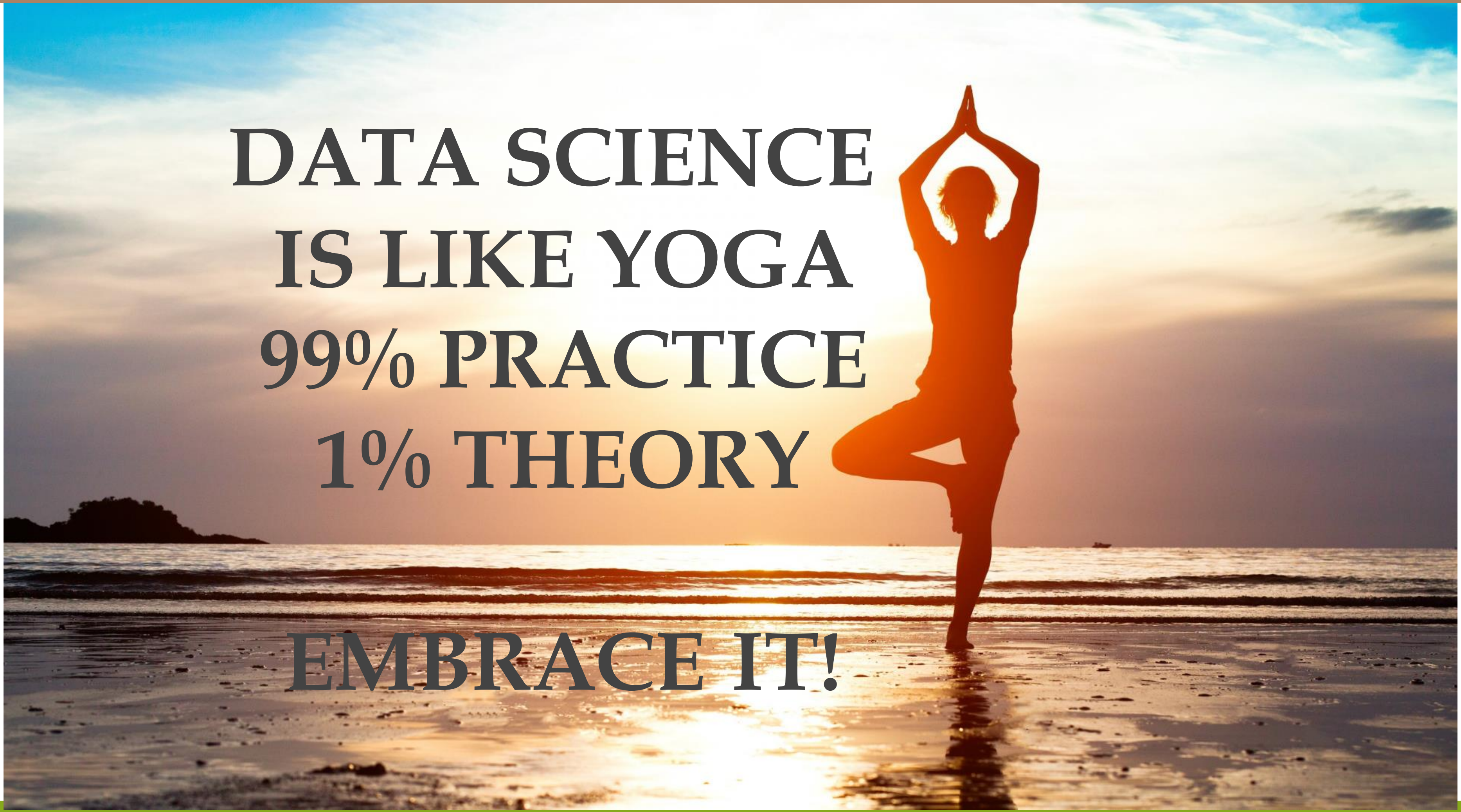
```
                              Specificity  Recall   AUC
Model
Naive-Bayes                         0.856   0.971  0.98
Random Forest                       0.866   0.896  0.95
Random Forest with Words2Vec        0.940   0.899  0.97
```

FINESSE FIT SOLUTIONS

RECOMMENDATIONS

- Can be used on other social media platforms (Twitter, Instagram).
- Insights into habits & interests:
  - ➢Yoga posts:
    - ○New & Different yoga techniques.
    - ○Yoga products such as mats or straps.
  - ➢Gym posts:
    - ○Favorite Protein bars/Shakes.
    - ○Gym accessories such as wrist straps or weightlifting belt.

Q&A



DATA SCIENCE
IS LIKE YOGA
99% PRACTICE
1% THEORY

EMBRACE IT!