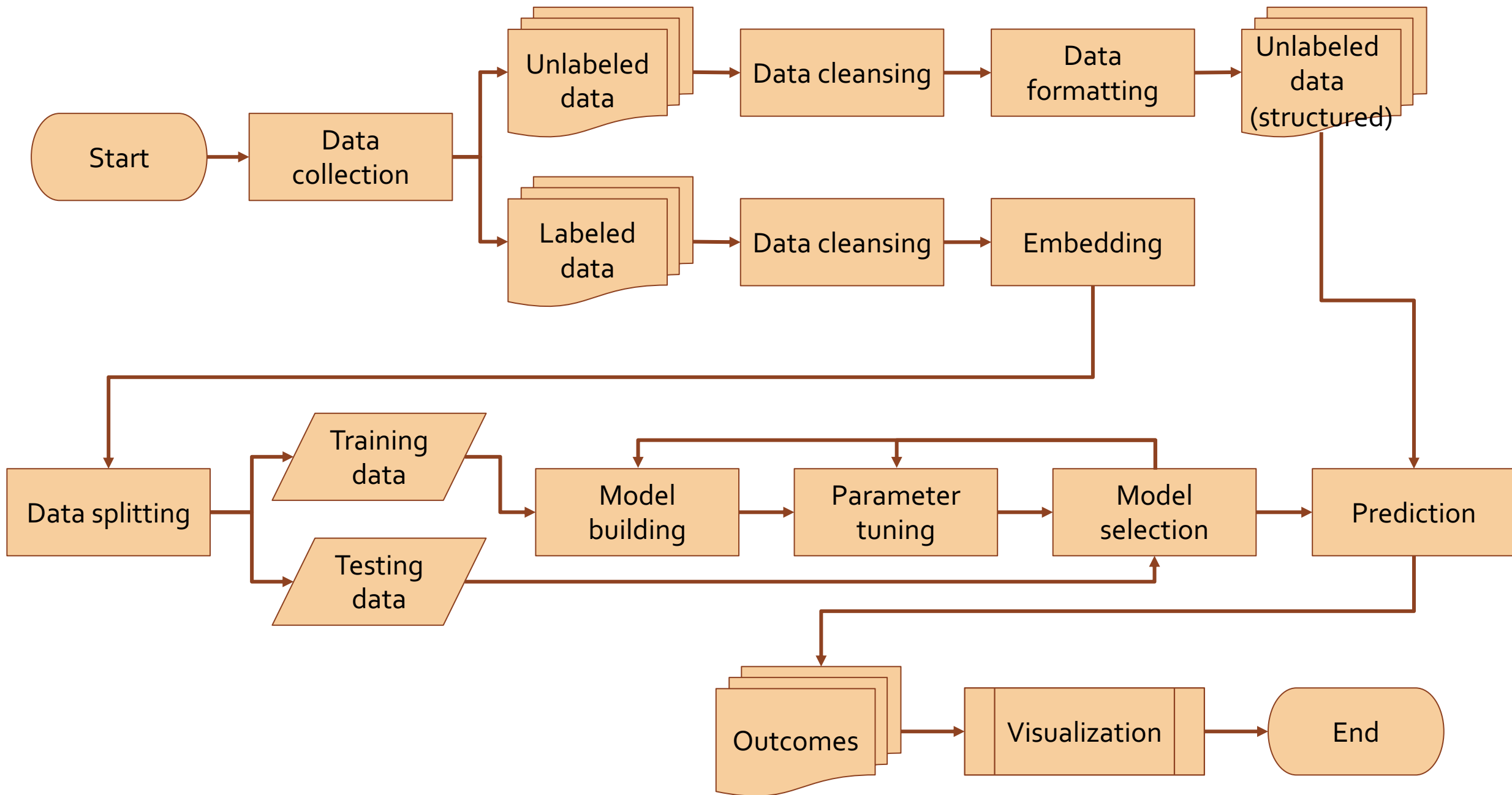


Big-Five Personality Prediction of Movie Characters -- Experiment Process

Yung-Chun Chen, Sep. 2023





Data

Labeled Data

- **essays.csv** (Pennebaker & King 1999)
 - 2400 stream-of-consciousness texts labelled with personality
- **mypersonality_final.csv** (Celli, Pianesi, Stillwell & Kosinski, 2013)
 - 10000 Facebook status updates of 250 users labelled with personality (with scores)
- Source: [here](#)
- **Questions: Which one performs better? Which one suits my unlabeled data?**

Unlabeled Data

- [Internet Movie Database](#)
- Web crawler from [GitHub](#)
 - Scraping, parsing, formatting
 - @misc{Saha_Movie_Script_Database_2021,
author = {Saha, Aveek},
month = {7},
title = {{Movie Script Database}},
url = {https://github.com/Aveek-Saha/Movie-Script-Database},
year = {2021} }

YOUNG ANNA
Elsa. Psst. Elsa! Psst.

Elsa doesn't stir. Anna sits on Elsa and bounces.

YOUNG ANNA (CONT'D)
Wake up. Wake up. Wake up.

YOUNG ELSA
(grumbling)
Anna, go back to sleep.

Anna rolls onto her back and spreads all her weight on Elsa.

3

FROZEN - J. Lee

YOUNG ANNA
(drama queen-ish)
I just can't. The sky's awake, so
I'm awake, so we have to play.

YOUNG ELSA
...Go play by yourself.

Elsa shoves Anna off the bed.

Anna lands butt to floor, sighs, defeated. But then she gets
an idea. She hops back on the bed and lifts one of Elsa's
eyelids.

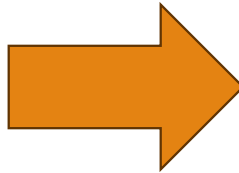
YOUNG ANNA
(mischievously)
Do you want to build a snowman?

Elsa's eyes both pop open. She smiles.

INT. CASTLE STAIRCASE -- NIGHT

Anna, now wearing snow boots, pulls Elsa by the hand.

YOUNG ANNA
Come on, come on, come on, come on.



YOUNG ANNA=>Elsa. Psst. Elsa! Psst.
YOUNG ANNA=>Wake up. Wake up. Wake up.
YOUNG ELSA=>Anna, go back to sleep.
YOUNG ANNA=>I just can't. The sky's awake, so I'm awake, so we have to play.
YOUNG ELSA=>...Go play by yourself.
YOUNG ANNA=>Do you want to build a snowman?
YOUNG ANNA=>Come on, come on, come on, come on.
YOUNG ANNA=>Do the magic! Do the magic!
YOUNG ANNA=>This is amazing!
YOUNG ELSA=>Watch this!
YOUNG ELSA=>Hi, I'm Olaf and I like warm hugs.
YOUNG ANNA=>I love you, Olaf.
YOUNG ANNA=>Catch me!
YOUNG ELSA=>Gotcha!
YOUNG ANNA=>Again! Again!
YOUNG ELSA=>Slow down!
YOUNG ELSA=>ANNA!
YOUNG ELSA=>MAMA! PAPA!
KING=>Elsa, what have you done? This is getting out of hand!
QUEEN=>Anna!
ELSA=>It was an accident. I'm sorry, Anna.
QUEEN=>She's ice cold.
KING=>...I know where we have to go.
KRISTOFF=>Ice?
YOUNG KRISTOFF=>Faster, Sven!
KING=>Please, help. My daughter!
YOUNG KRISTOFF=>Trolls...?



Data Cleansing

Text

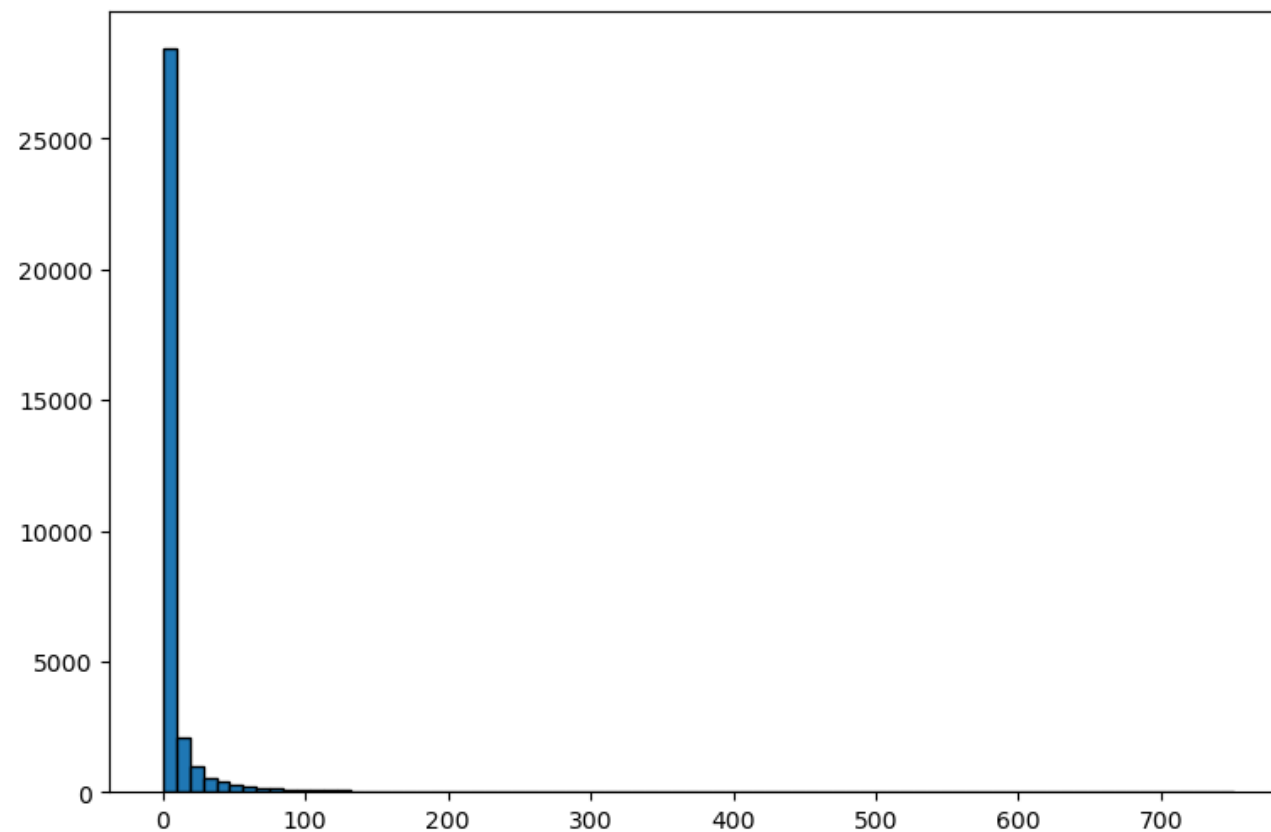
Training & predicting data

- Removed punctuations, numbers, non-sense words, stop words, and made them lower

Predicting data

- Found main characters: characters whose number of lines exceeds the **95th percentile**.
- Applied **text-embedding-ada-002** transforming model from OpenAI
- Dropped files < 5mb

Distribution of lines



	line_num
count	34031.000000
mean	10.087068
std	33.262733
min	0.000000
25%	1.000000
50%	1.000000
75%	5.000000
80%	7.000000
90%	19.000000
95%	43.000000
max	751.000000



Model Selection

Model Comparison

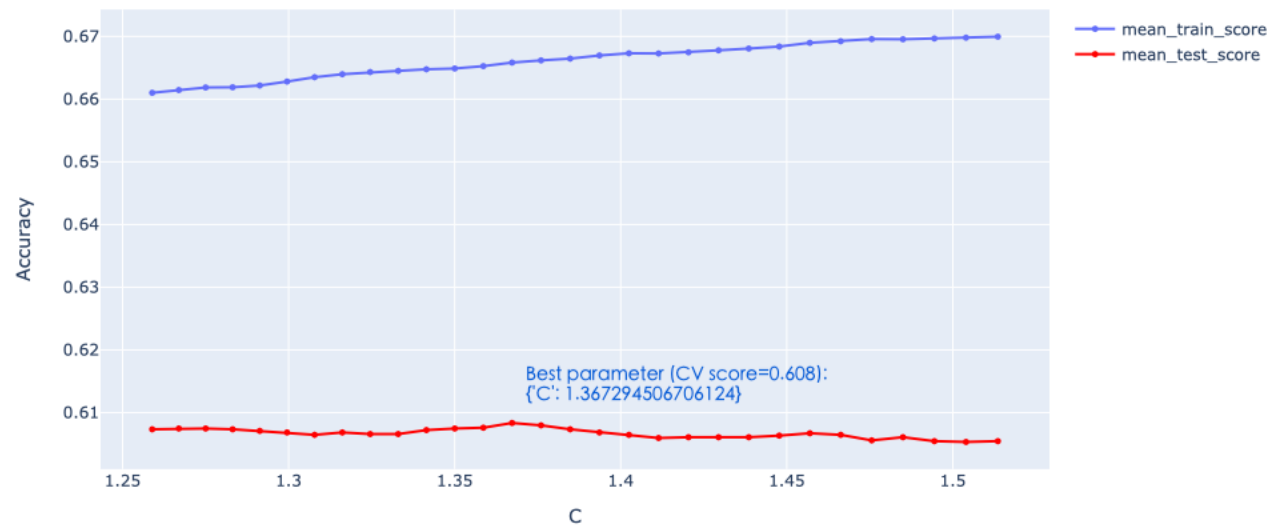
1. Classification vs. regression
2. Random forest vs. SVM vs. Adaboost vs. MLPClassifier accuracy & overfitting
3. Different training datasets
4. Different methods of data cleansing

remove punctuations	remove stop words	rank
T	F	2
T	T	1
F	T	4
F	F	3

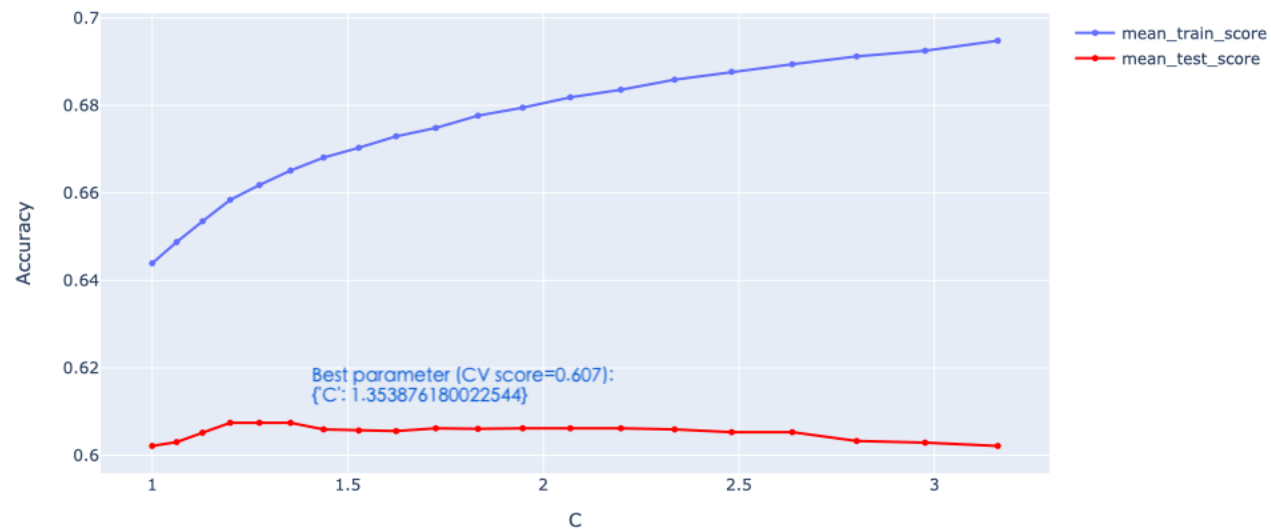
5. SVM parameters:
linear vs. rbf vs. poly
scale vs. auto
6. Gridsearch: find the best C

Different training datasets

Essays: linear - scale

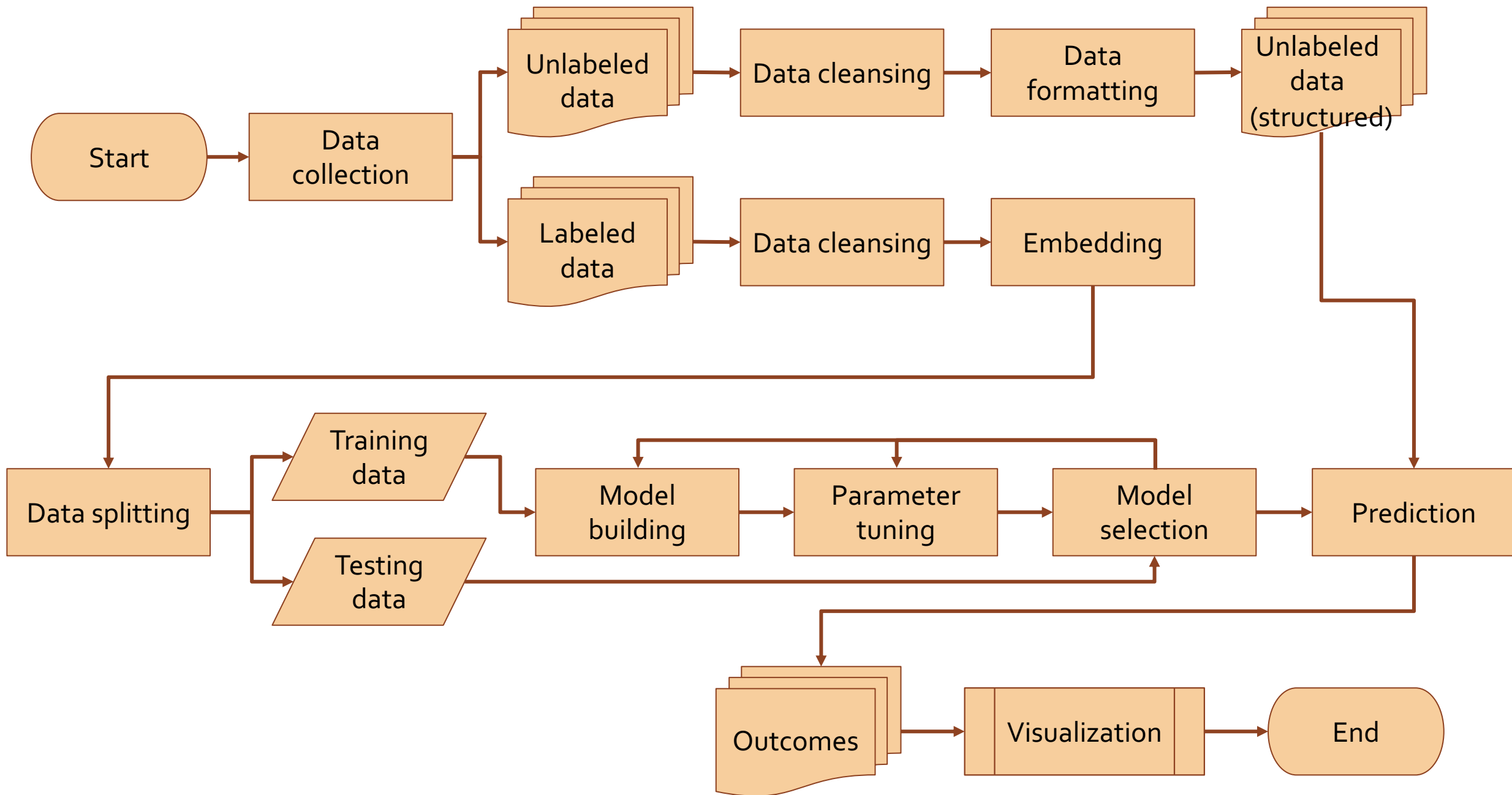


Mypersonality: linear - scale





The process is NOT linear





Further Improvement

Further Improvement

1. Data cleansing and structuring
2. Apply sentiment analysis for more features
3. Will the length of text affect?
4. Different models