Problem statement

↓

Problems ← Data → Data formats

- json
- images
- .txt
- .csv

↓

Initial preprocessing

- Language detection
- Language translation

↓

EDA

(A) Ngrams

- Unigram
- Bigram
- Trigram

→ Domain Specific stopwords

→ Keyphrases - Root cause analysis

(B) Word cloud

(C) Keyphrase extraction

- RAKE
- YAKE

↓

Preprocessing

→ (1) Remove spaces,
null, blanklines

- ① Remove spaces, newlines, blanklines
- ② Contraction mapping
- ③ Handling accented characters
- ④ Cleaning
  - ⓐ Tokenization
    - Sentence
    - word
    - Whitespace
    - Regex tokenizer
  - ⓑ Remove Punctuation
  - ⓒ Remove stopwords
    - language specific
    - domain specific
  - ⓓ len(word) > 2
  - ⓔ Normalization
- ⑤ Autocorrection
  - autocorrect
  - text blob
  - Symspell
- ⑥ Stemmer

Lemmatizer

↳ (6) Stemmer

Lemmatizer

Porter stemmer

Snowball stemmer

↓

Clean data

Target yes ↙ ↘ No Target

evaluation:

- Silhouette score
- Silhouette visualizer
- Dunn index
- Kappa index

(1) Clustering

(a) Kmeans - Count Vectorizer

- TFIDF
- Word 2 Vec

(b) Hierarchical clustering

(c) DB scan

(d) SOM → Self Organising Maps

(2) Supportive column

- Star / Rating column

↓

Training data

↓

Modelling ⟶ (1) Logistic Regression

↓

Modelling → (1) Logistic Regression

(2) SVM

(3) Decision tree classifier

(4) Adaboost

(5) Naive Bayes

(6) RNN, LSTM

↓

evaluation

↓

Create pickle file

↓

Deployment

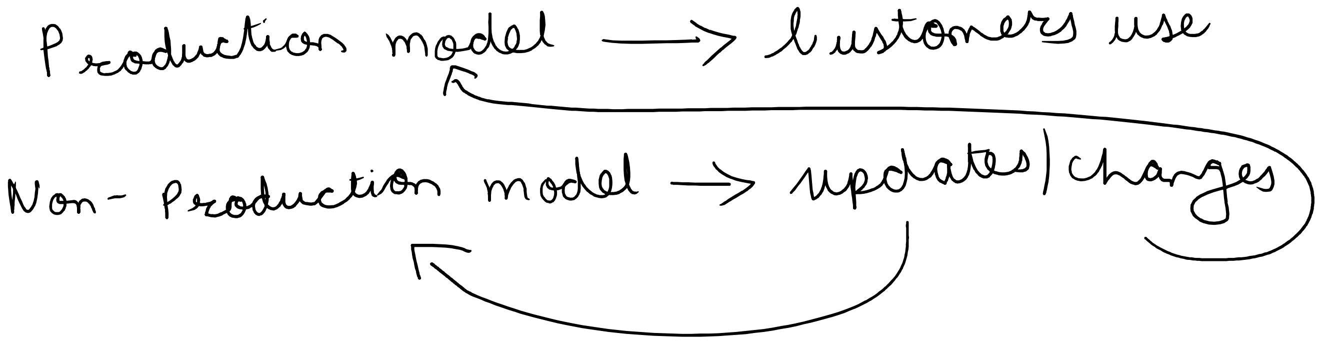
↓

CICD

Model Build → Deployment

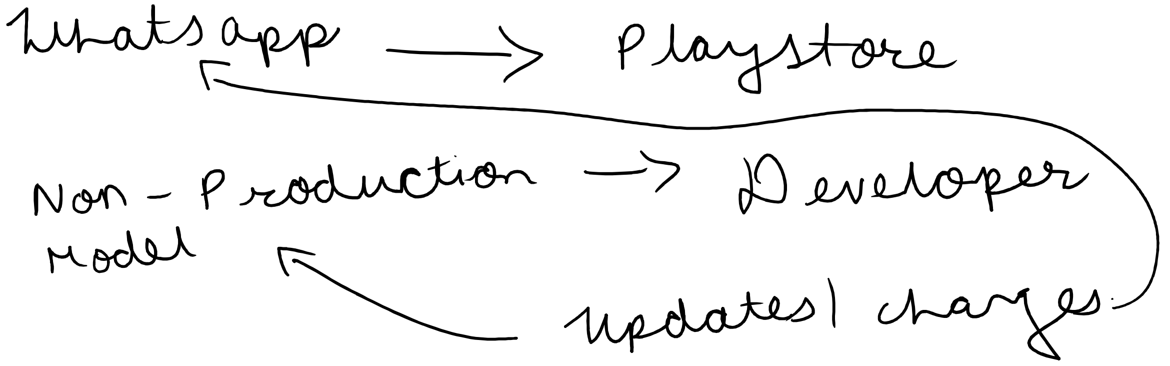
3 - 6 Months

CICD → Continuous integration,
Continuous deployment



Production model → Customers use

Non-Production model → updates/changes



Whatsapp → Playstore

Non-Production model → Developer

updates/changes