NLP Framework

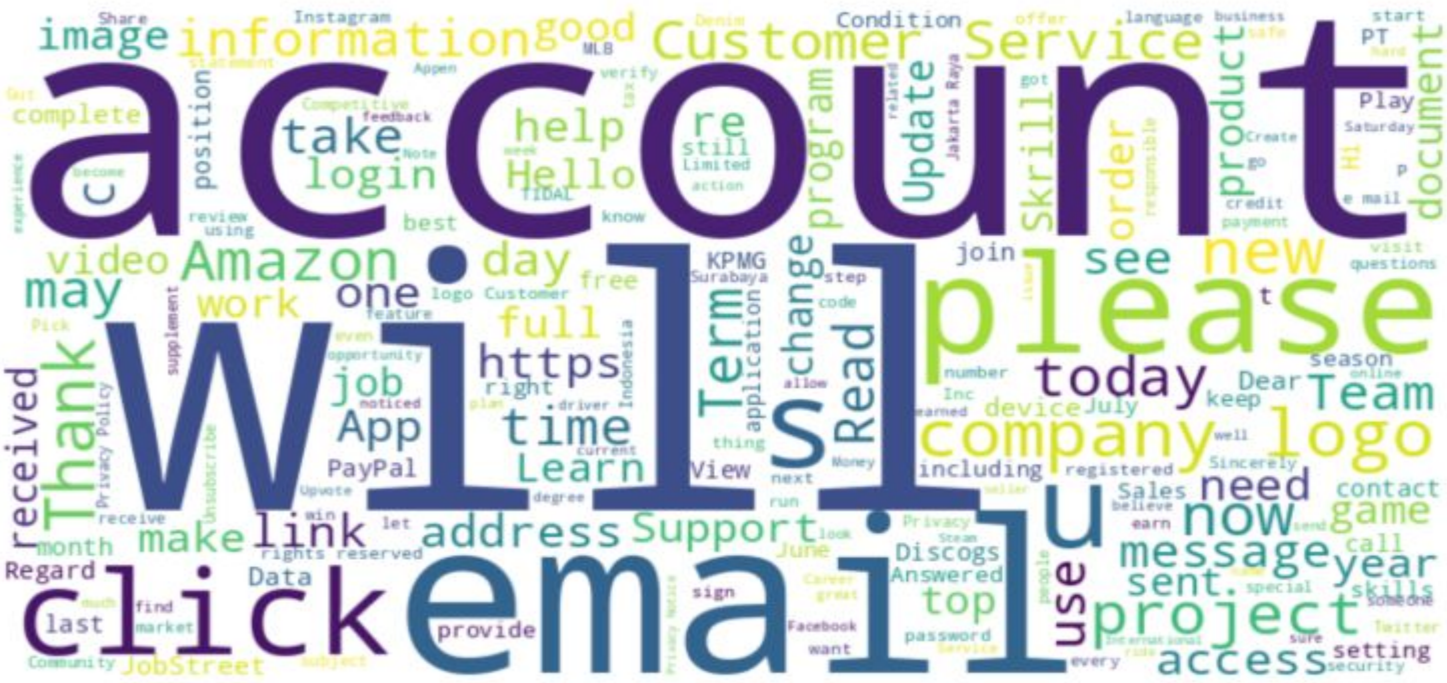
(Natural Language Processing)

Developer:

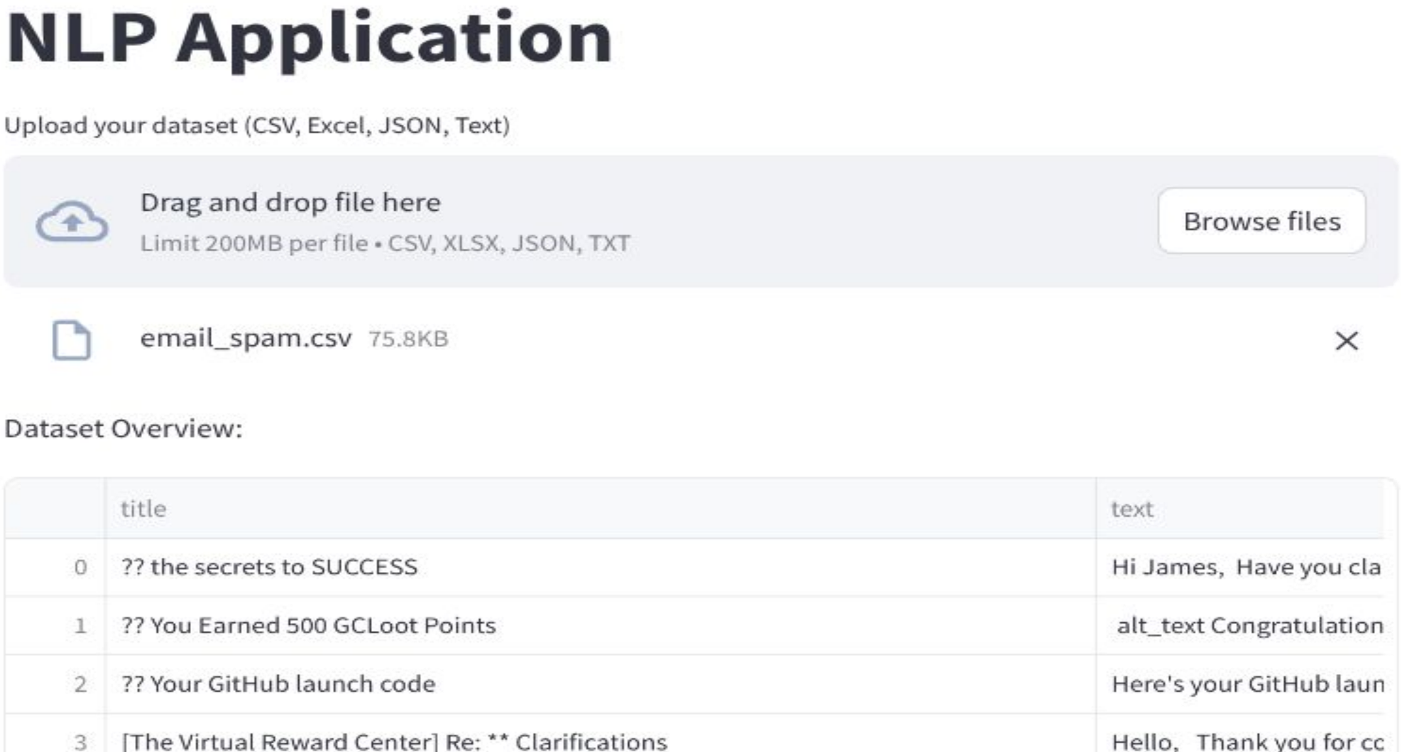
TC Antony - Data Scientist

Step 1: Word Cloud (for overall understanding about data)

Word Cloud



NLP Framework



NLP Application

Upload your dataset (CSV, Excel, JSON, Text)

Drag and drop file here

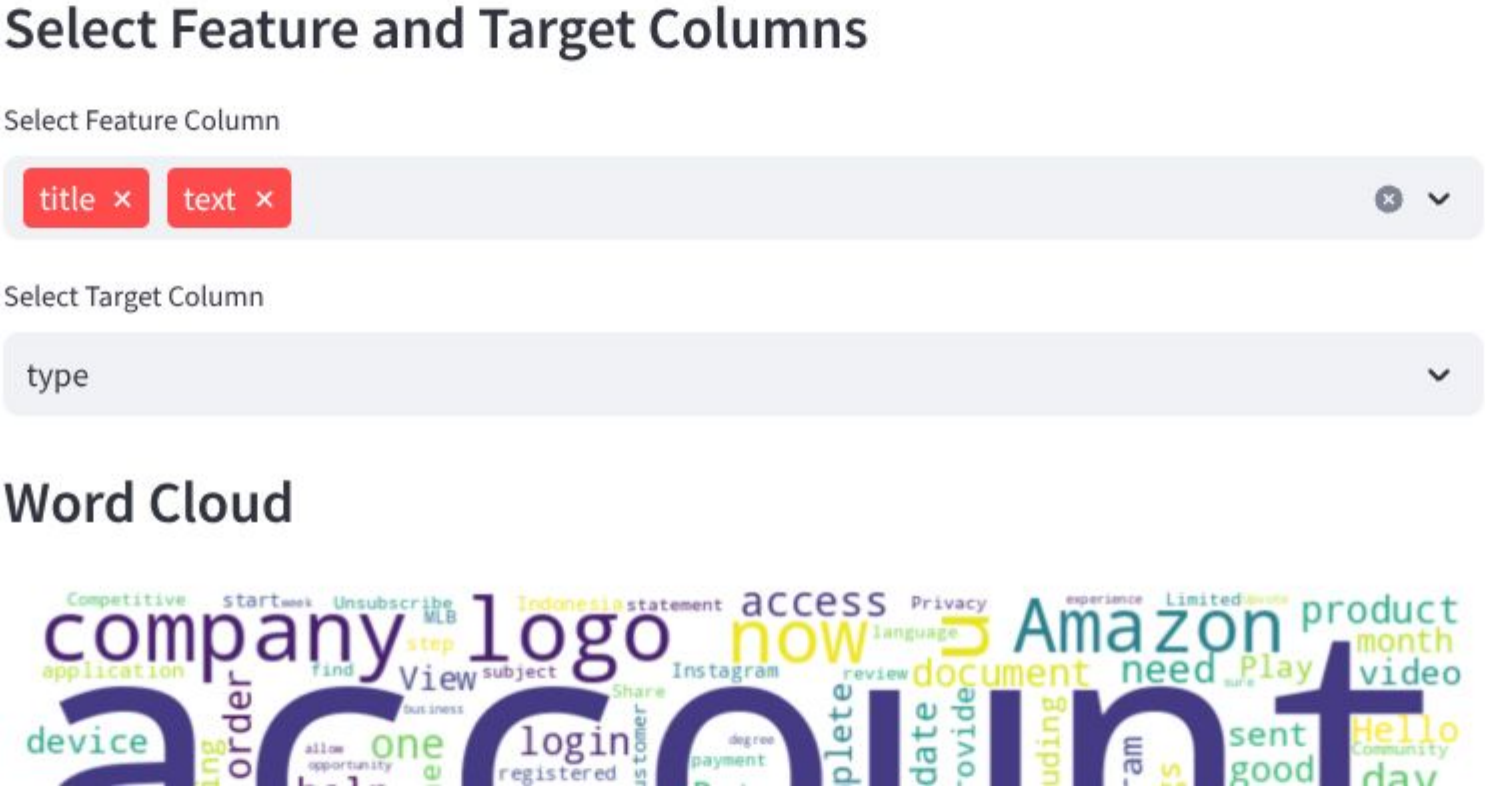
Limit 200MB per file • CSV, XLSX, JSON, TXT

email_spam.csv 75.8KB

Dataset Overview:

	title	text
0	?? the secrets to SUCCESS	Hi James, Have you cla
1	?? You Earned 500 GCLoot Points	alt_text Congratulation
2	?? Your GitHub launch code	Here's your GitHub laun
3	[The Virtual Reward Center] Re: ** Clarifications	Hello, Thank you for cc

Feature and Target Column Selection



Select Feature and Target Columns

Select Feature Column

title × text ×

Select Target Column

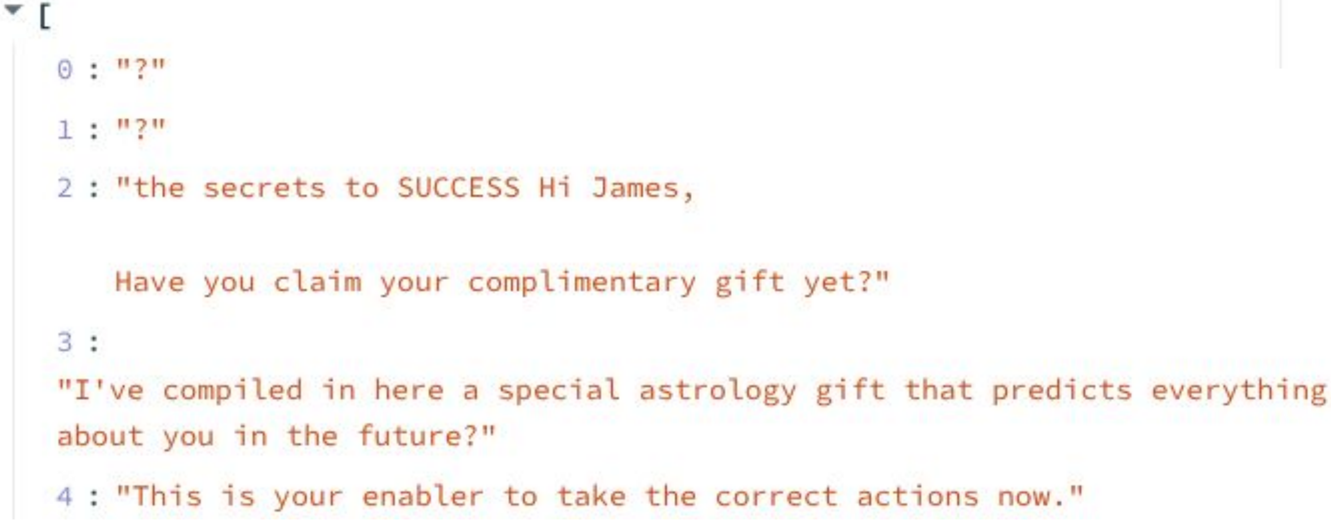
type

Word Cloud

Step 2: Preprocessing

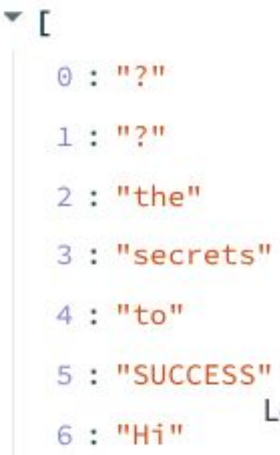
1) Tokenization

- Sentence Tokenization (Splitting text into sentences)
- Word Tokenization (Splitting sentences into words)
- Letter Tokenization (Splitting words into characters)



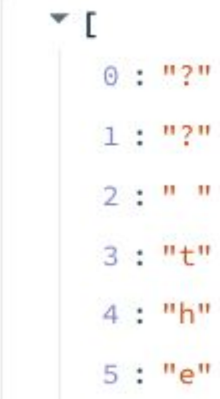
```
[
  0 : "?"
  1 : "?"
  2 : "the secrets to SUCCESS Hi James,

      Have you claim your complimentary gift yet?"
  3 :
  "I've compiled in here a special astrology gift that predicts everything
  about you in the future?"
  4 : "This is your enabler to take the correct actions now."
```

Word Tokenization:



```
[
  0 : "?"
  1 : "?"
  2 : "the"
  3 : "secrets"
  4 : "to"
  5 : "SUCCESS"
  6 : "Hi"
```

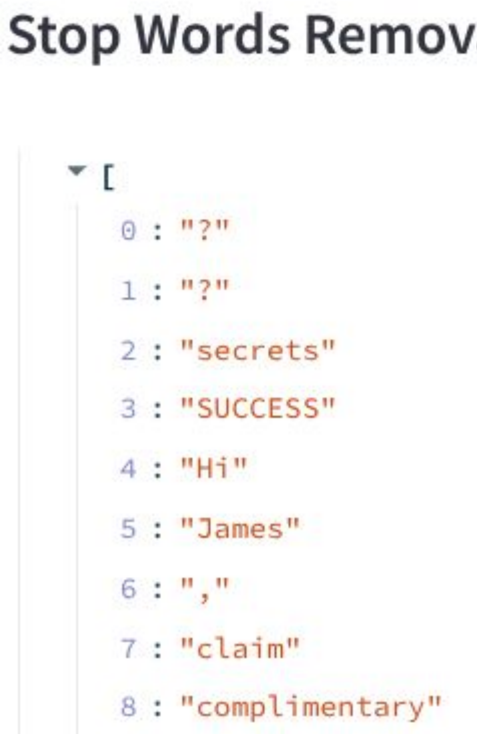
Letter Tokenization:



```
[
  0 : "?"
  1 : "?"
  2 : " "
  3 : "t"
  4 : "h"
  5 : "e"
```

2) Stop Words Removal

Removing common words (e.g., "the", "is", "and") to improve model efficiency



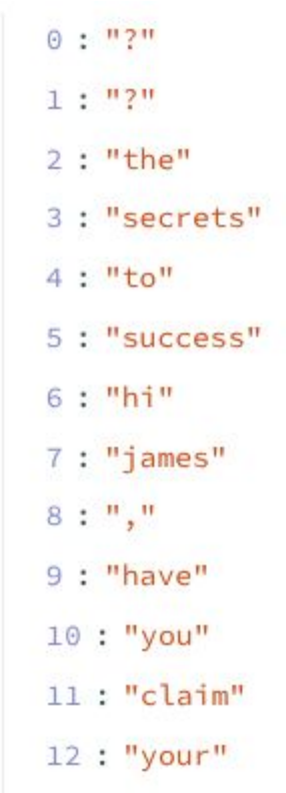
Stop Words Removal

```
▾ [
   0 : "?"
   1 : "?"
   2 : "secrets"
   3 : "SUCCESS"
   4 : "Hi"
   5 : "James"
   6 : ","
   7 : "claim"
   8 : "complimentary"
```

Compare this to previous slide the is like words are removed

3) Lowercasing

Converting text to lowercase to maintain consistency.

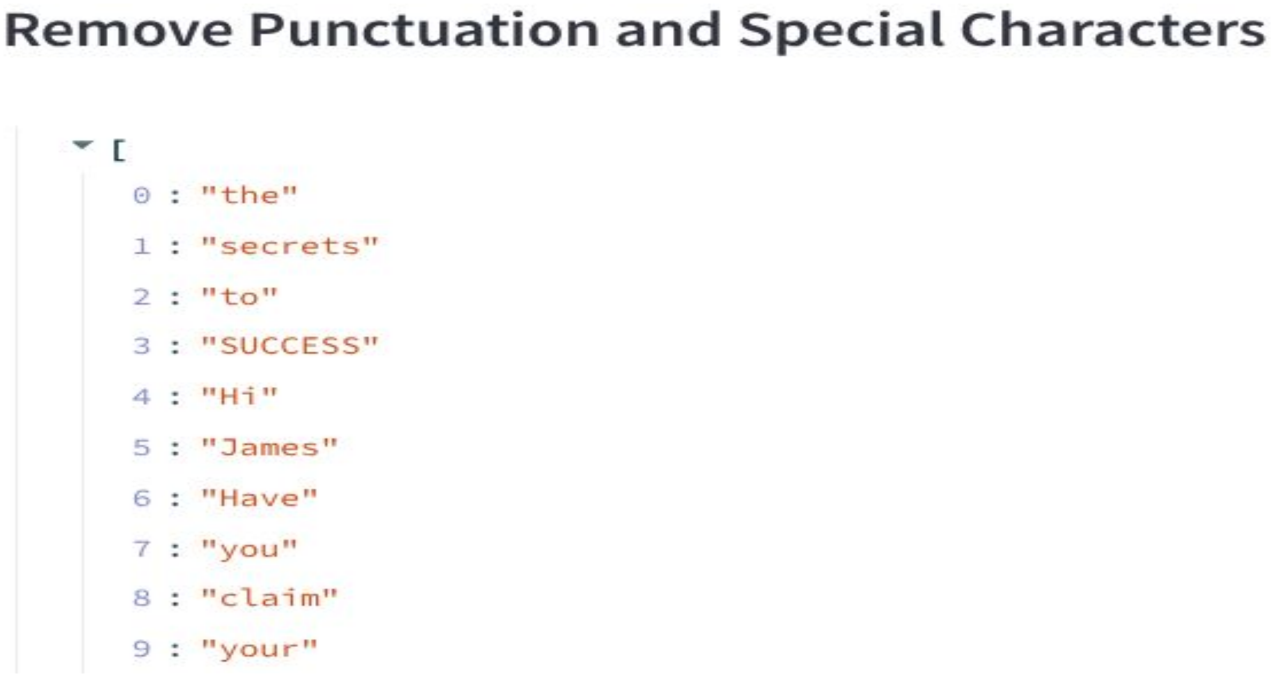


```
0 : "?"
1 : "?"
2 : "the"
3 : "secrets"
4 : "to"
5 : "success"
6 : "hi"
7 : "james"
8 : ","
9 : "have"
10 : "you"
11 : "claim"
12 : "your"
```

Compare this to previous slide - success, hi, james these words are convert uppercase to lowercase

4) Remove Punctuation

Removing special characters (e.g., ".", "!", "?") to simplify processing.



Remove Punctuation and Special Characters

```
▾ [
    0 : "the"
    1 : "secrets"
    2 : "to"
    3 : "SUCCESS"
    4 : "Hi"
    5 : "James"
    6 : "Have"
    7 : "you"
    8 : "claim"
    9 : "your"
```

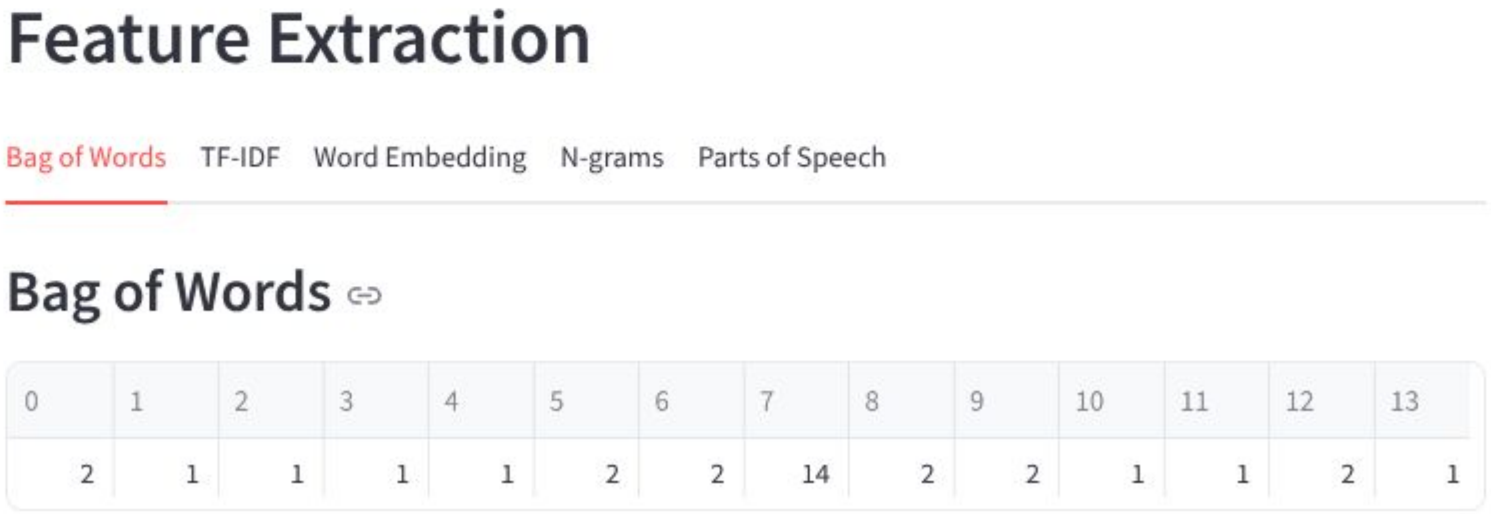
5) Lemmatization

Converting words to their base forms (e.g., "running" → "run").

Step 3: Feature Extraction

Bag of Words (BoW)

- Converts text into a frequency-based numerical representation.



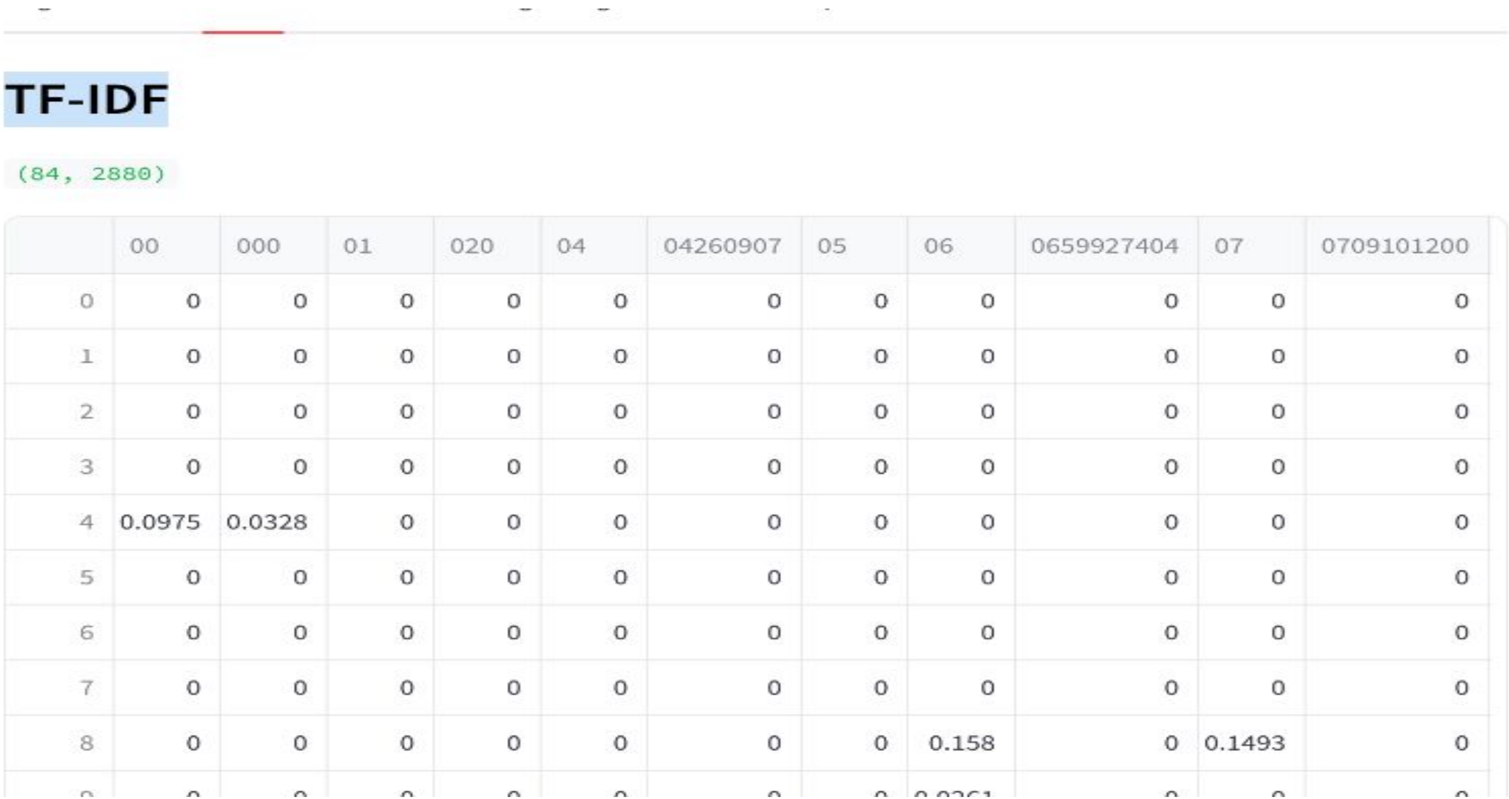
Feature Extraction

Bag of Words

0	1	2	3	4	5	6	7	8	9	10	11	12	13
2	1	1	1	1	2	2	14	2	2	1	1	2	1

2) TF-IDF

TF-IDF (Term Frequency - Inverse Document Frequency)



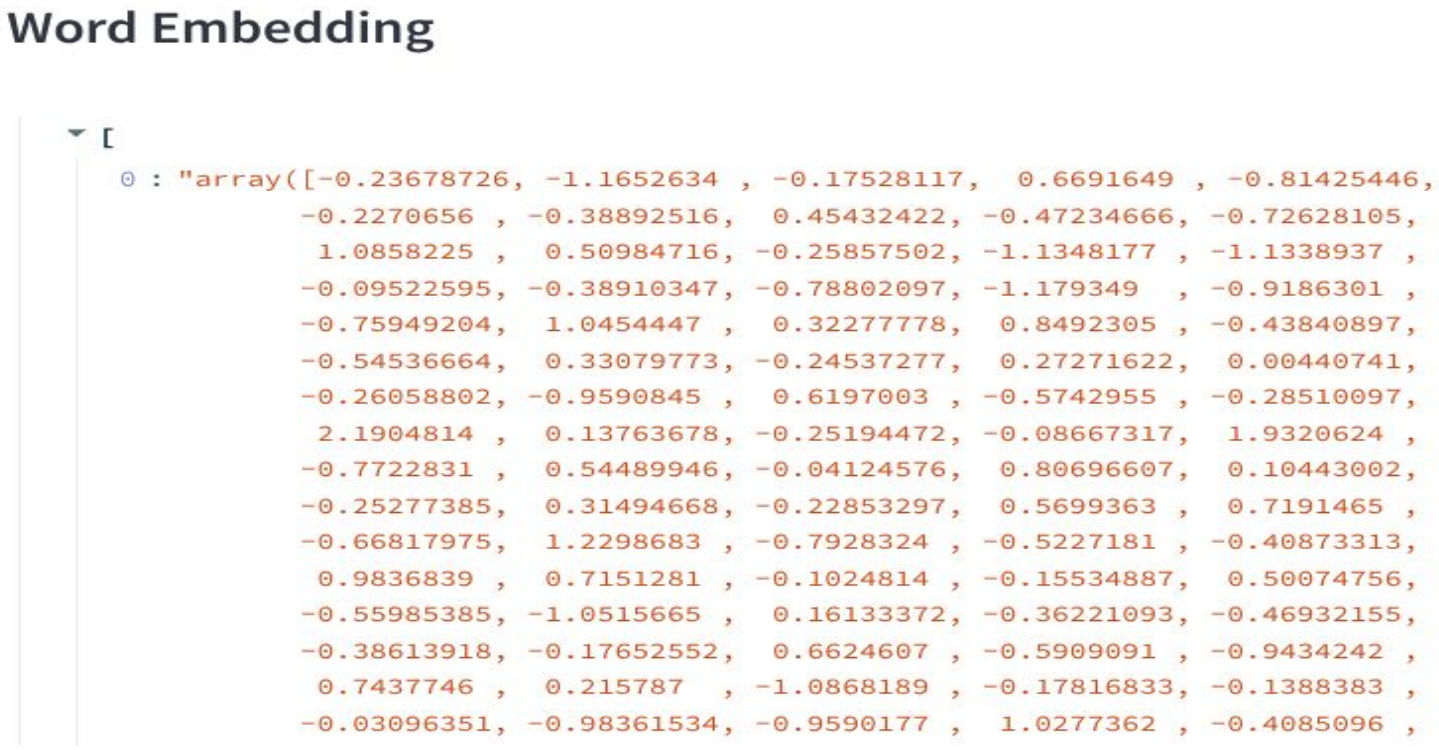
TF-IDF

```
(84, 2880)
```

	00	000	01	020	04	04260907	05	06	0659927404	07	0709101200
0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0
4	0.0975	0.0328	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0.158	0	0.1493	0
9	0	0	0	0	0	0	0	0.0261	0	0	0

3) Word Embedding (Using spaCy)

- Converts words into dense vector representations for capturing semantic meaning.



Word Embedding

```
▾ [
  0 : "array([-0.23678726, -1.1652634 , -0.17528117,  0.6691649 , -0.81425446,
             -0.2270656 , -0.38892516,  0.45432422, -0.47234666, -0.72628105,
              1.0858225 ,  0.50984716, -0.25857502, -1.1348177 , -1.1338937 ,
             -0.09522595, -0.38910347, -0.78802097, -1.179349  , -0.9186301 ,
             -0.75949204,  1.0454447 ,  0.32277778,  0.8492305 , -0.43840897,
             -0.54536664,  0.33079773, -0.24537277,  0.27271622,  0.00440741,
             -0.26058802, -0.9590845 ,  0.6197003 , -0.5742955 , -0.28510097,
              2.1904814 ,  0.13763678, -0.25194472, -0.08667317,  1.9320624 ,
             -0.7722831 ,  0.54489946, -0.04124576,  0.80696607,  0.10443002,
             -0.25277385,  0.31494668, -0.22853297,  0.5699363 ,  0.7191465 ,
             -0.66817975,  1.2298683 , -0.7928324 , -0.5227181 , -0.40873313,
              0.9836839 ,  0.7151281 , -0.1024814 , -0.15534887,  0.50074756,
             -0.55985385, -1.0515665 ,  0.16133372, -0.36221093, -0.46932155,
             -0.38613918, -0.17652552,  0.6624607 , -0.5909091 , -0.9434242 ,
              0.7437746 ,  0.215787  , -1.0868189 , -0.17816833, -0.1388383 ,
             -0.03096351, -0.98361534, -0.9590177 ,  1.0277362 , -0.4085096 ,
```

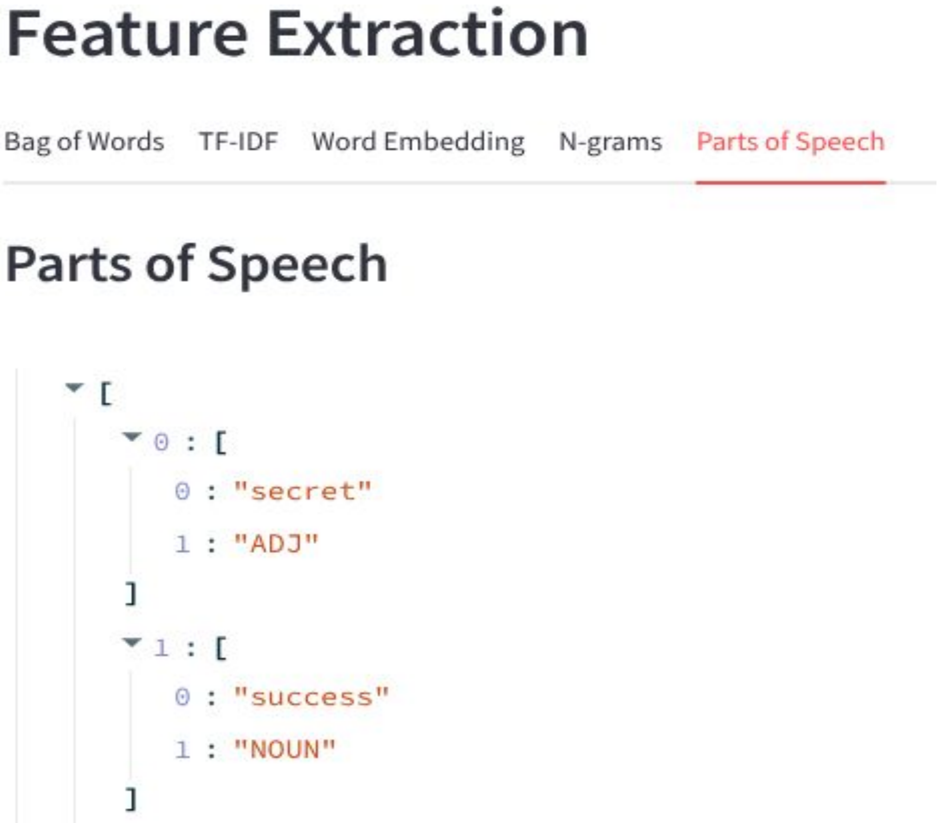
4) **N-grams**

N-grams

- Generates:
  - **Unigrams** (single words)
  - **Bigrams** (two-word combinations)
  - **Trigrams** (three-word combinations)

5) POS (Part of Speech)

Identifies the grammatical role of words (e.g., noun, verb, adjective).



Feature Extraction

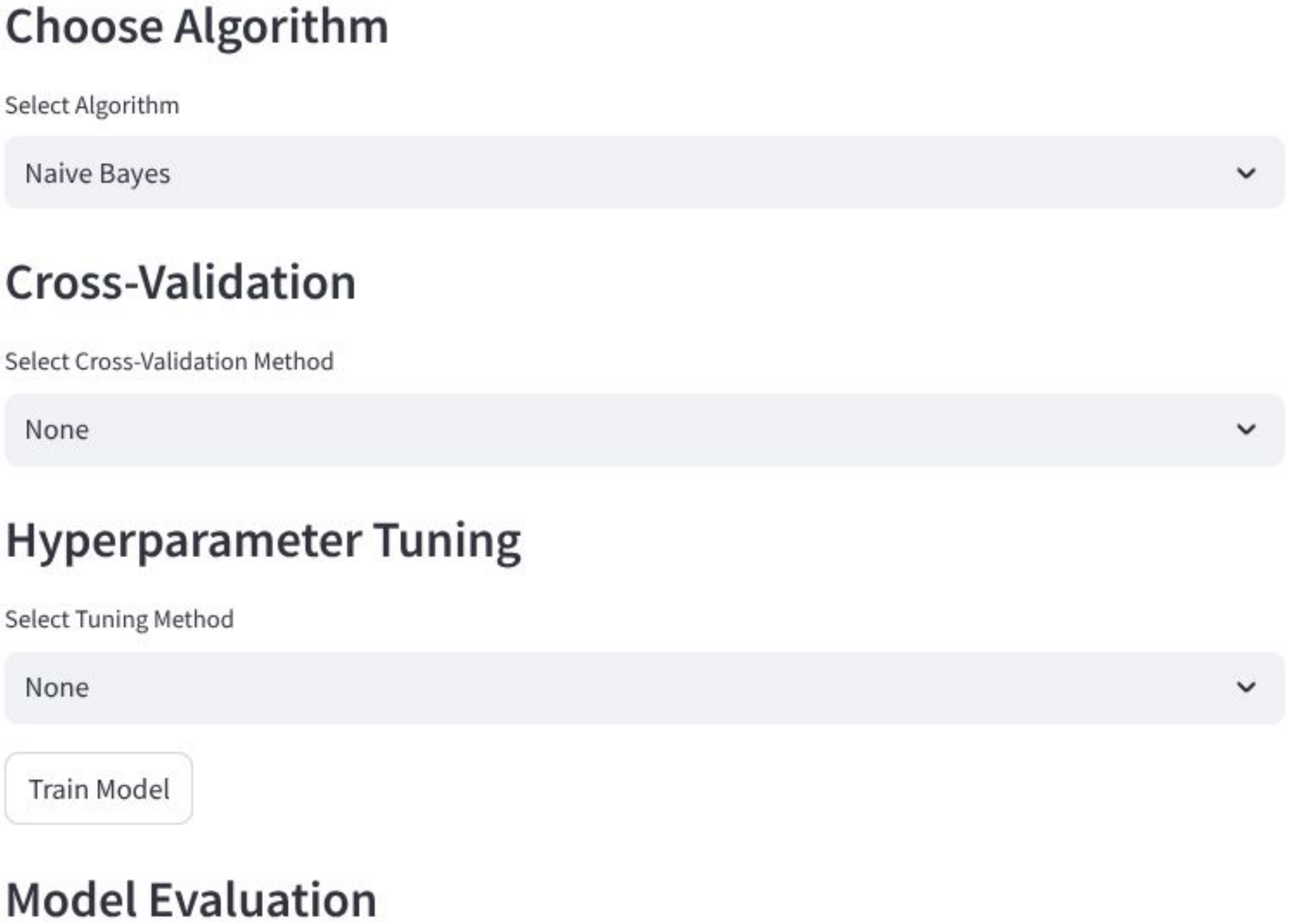
Parts of Speech

```
▼ [
   ▼ 0 : [
        0 : "secret"
        1 : "ADJ"
     ]
   ▼ 1 : [
        0 : "success"
        1 : "NOUN"
     ]
```

Step 4: Model Training

Machine Learning Algorithms Used

- **Naive Bayes (MultinomialNB)**
  - Best suited for text classification tasks.
- **Support Vector Machine (SVM)**
  - Effective in high-dimensional spaces.
- **Decision Tree Classifier**
  - Splits data into a hierarchical structure.
- **Random Forest Classifier**
  - An ensemble method that improves accuracy.
- **Gradient Boosting Classifier**
  - Boosts weak learners iteratively for better performance.



Choose Algorithm

Select Algorithm

Naive Bayes

Cross-Validation

Select Cross-Validation Method

None

Hyperparameter Tuning

Select Tuning Method

None

Model Evaluation

4. Cross-Validation Techniques

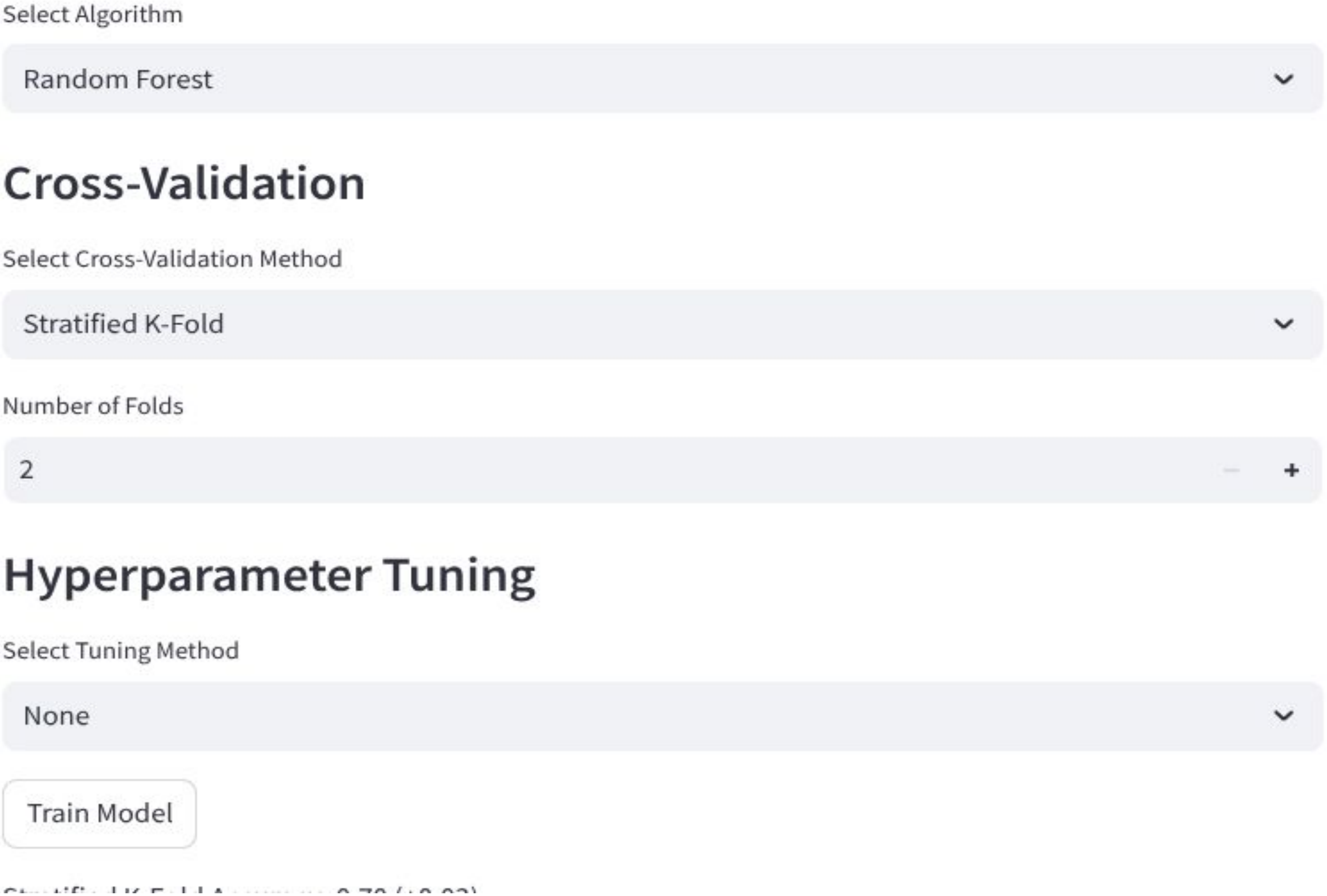
Cross-validation is used to evaluate model performance.

- **K-Fold Cross-Validation**
  - Splits data into ‘K’ parts and trains the model multiple times.
- **Stratified K-Fold**
  - Ensures each fold has a balanced class distribution.
- **Hold-Out Method**
  - Splits data into training and testing sets (e.g., 80% training, 20% testing).

5. Hyperparameter Tuning

Optimizes model performance by selecting the best parameters.

- **Grid Search**
  - Exhaustively searches over all parameter combinations.
- **Random Search**
  - Randomly selects parameter combinations for faster tuning.



Select Algorithm

Random Forest

Cross-Validation

Select Cross-Validation Method

Stratified K-Fold

Number of Folds

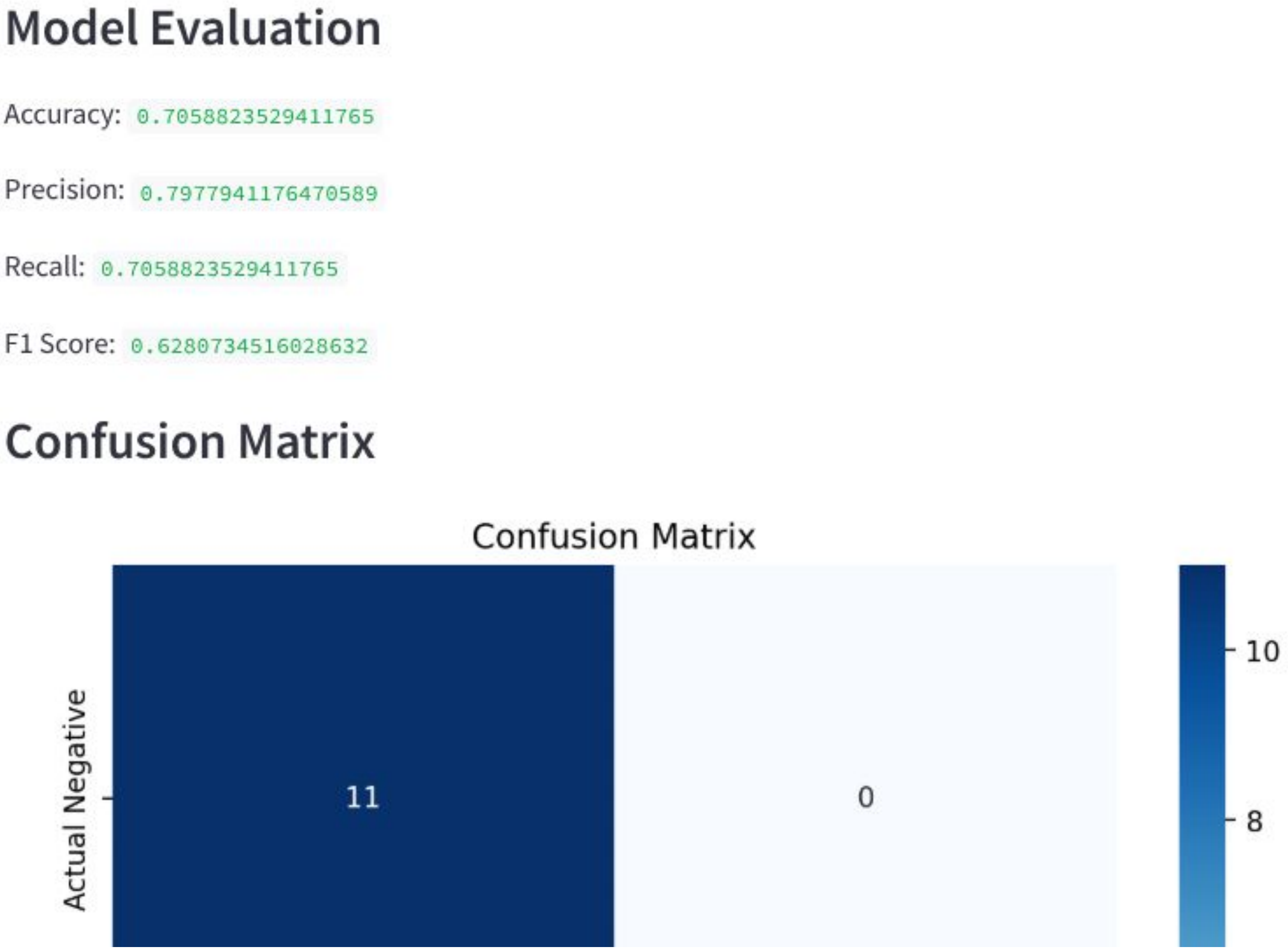
2

Hyperparameter Tuning

Select Tuning Method

None

Stratified K-Fold Accuracy: 0.79 (±0.03)



Model Evaluation

Accuracy: `0.7058823529411765`

Precision: `0.7977941176470589`

Recall: `0.7058823529411765`

F1 Score: `0.6280734516028632`

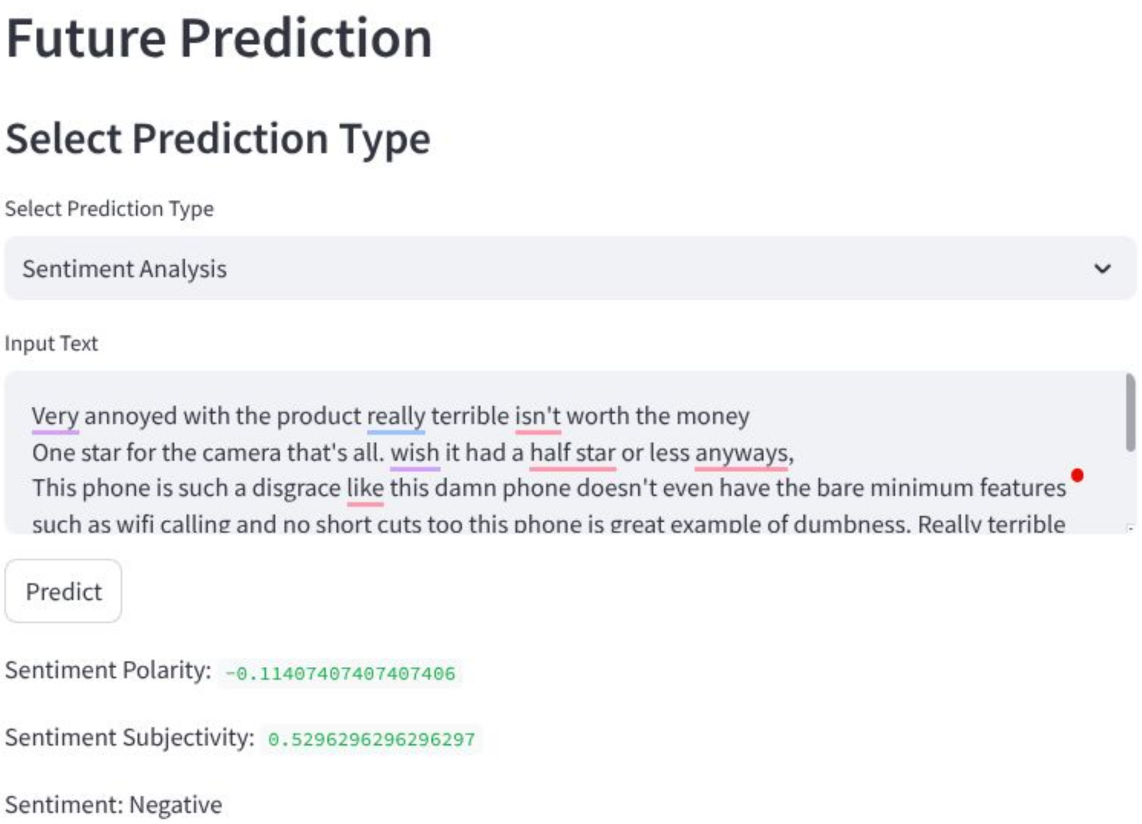
Confusion Matrix

Confusion Matrix Explanation

- **True Positive (TP)**: The model correctly predicted the positive class.
- **False Positive (FP)**: The model incorrectly predicted the positive class (Type I Error).
- **True Negative (TN)**: The model correctly predicted the negative class.
- **False Negative (FN)**: The model incorrectly predicted the negative class (Type II Error).

- **True Positives (TP)**: 1
- **False Positives (FP)**: 0
- **True Negatives (TN)**: 11
- **False Negatives (FN)**: 5

Sentiment Analysis



Future Prediction

Select Prediction Type

Select Prediction Type

Sentiment Analysis

Input Text

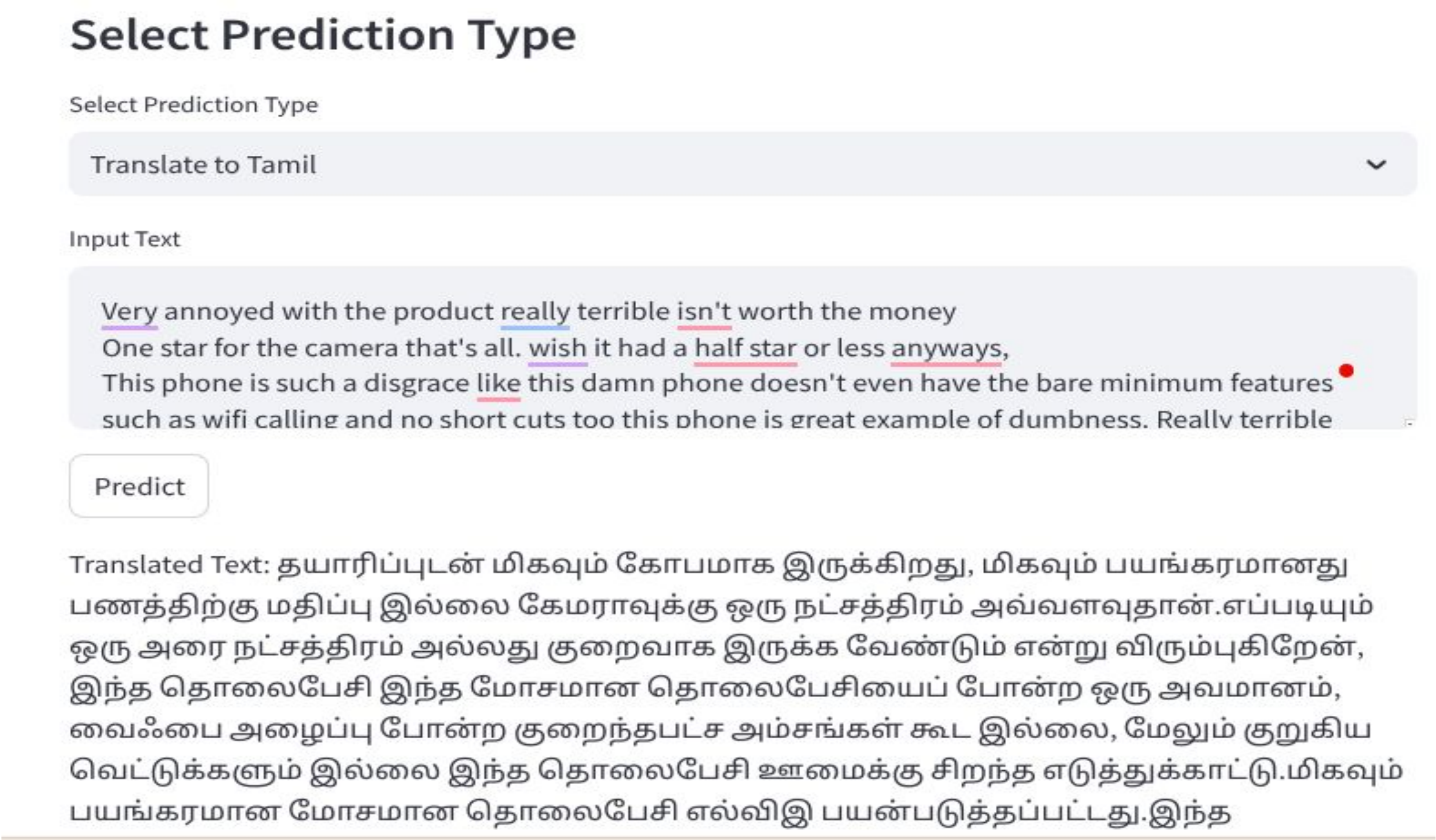
Very annoyed with the product really terrible isn't worth the money
One star for the camera that's all. wish it had a half star or less anyways,
This phone is such a disgrace like this damn phone doesn't even have the bare minimum features
such as wifi calling and no short cuts too this phone is great example of dumbness. Really terrible

Sentiment Polarity: `-0.11407407407407406`

Sentiment Subjectivity: `0.5296296296296297`

Sentiment: Negative

Translator



Select Prediction Type

Select Prediction Type

Translate to Tamil

Input Text

Very annoyed with the product really terrible isn't worth the money
One star for the camera that's all. wish it had a half star or less anyways,
This phone is such a disgrace like this damn phone doesn't even have the bare minimum features
such as wifi calling and no short cuts too this phone is great example of dumbness. Really terrible

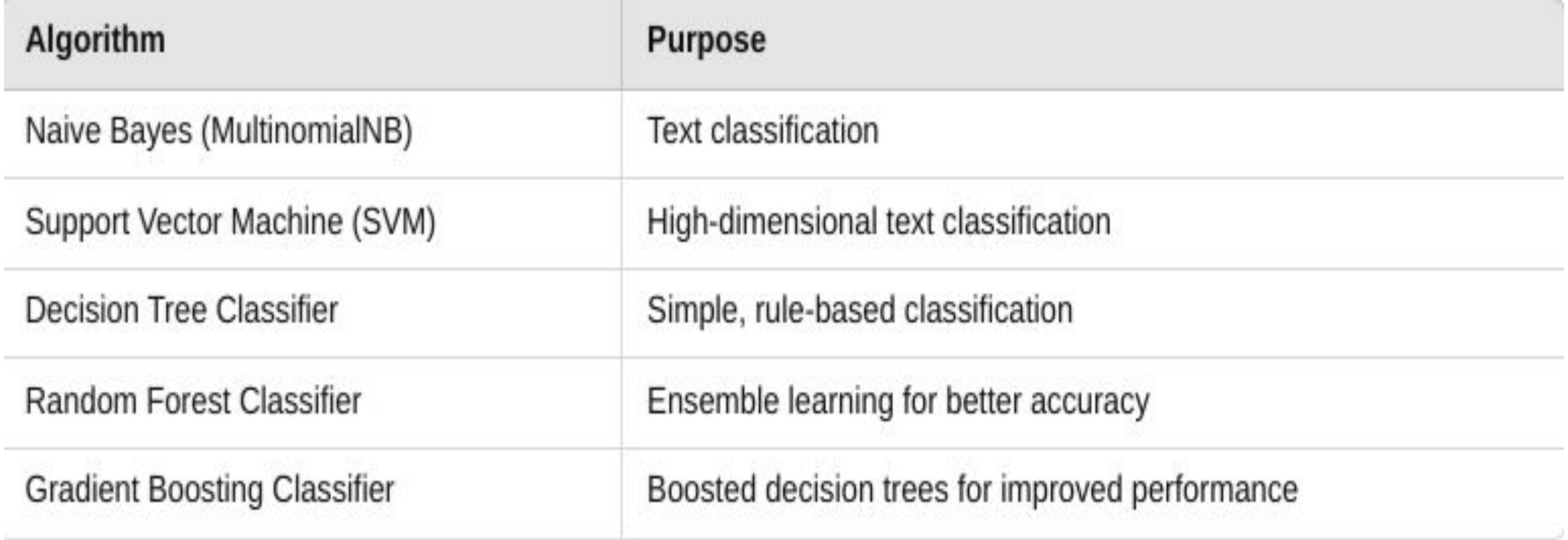
Translated Text: தயாரிப்புடன் மிகவும் கோபமாக இருக்கிறது, மிகவும் பயங்கரமானது பணத்திற்கு மதிப்பு இல்லை கேமராவுக்கு ஒரு நட்சத்திரம் அவ்வளவுதான்.எப்படியும் ஒரு அரை நட்சத்திரம் அல்லது குறைவாக இருக்க வேண்டும் என்று விரும்புகிறேன், இந்த தொலைபேசி இந்த மோசமான தொலைபேசியைப் போன்ற ஒரு அவமானம், வைஃபை அழைப்பு போன்ற குறைந்தபட்ச அம்சங்கள் கூட இல்லை, மேலும் குறுகிய வெட்டுக்களும் இல்லை இந்த தொலைபேசி ஊமைக்கு சிறந்த எடுத்துக்காட்டு.மிகவும் பயங்கரமான மோசமான தொலைபேசி எல்விஇ பயன்படுத்தப்பட்டது.இந்த

6. Future Prediction

The trained model makes predictions based on user input.

Prediction Options

- **Spam/Ham Detection** (Classifies text as spam or not)
- **Sentiment Analysis** (Determines positive, negative, or neutral sentiment)
- **Topic Prediction** (Identifies the main topic of the text)
- **Document Classification** (Classifies input text into predefined categories)
- **Language Detection** (Detects the language of the text)
- **Text Translation** (Translates text to Tamil using Google Translate)



Algorithm	Purpose
Naive Bayes (MultinomialNB)	Text classification
Support Vector Machine (SVM)	High-dimensional text classification
Decision Tree Classifier	Simple, rule-based classification
Random Forest Classifier	Ensemble learning for better accuracy
Gradient Boosting Classifier	Boosted decision trees for improved performance