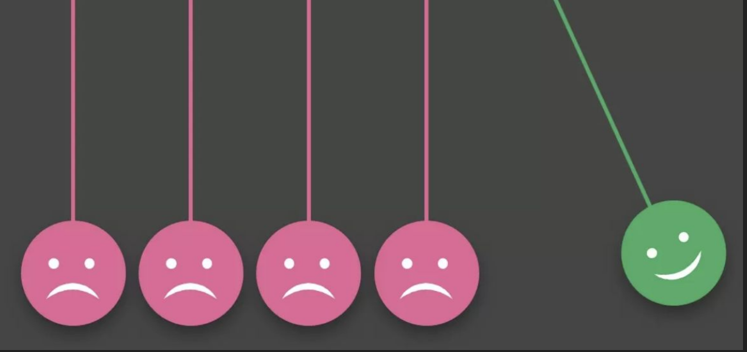


Sentiment Analysis

Social Media Based

Paola Garay

Business case

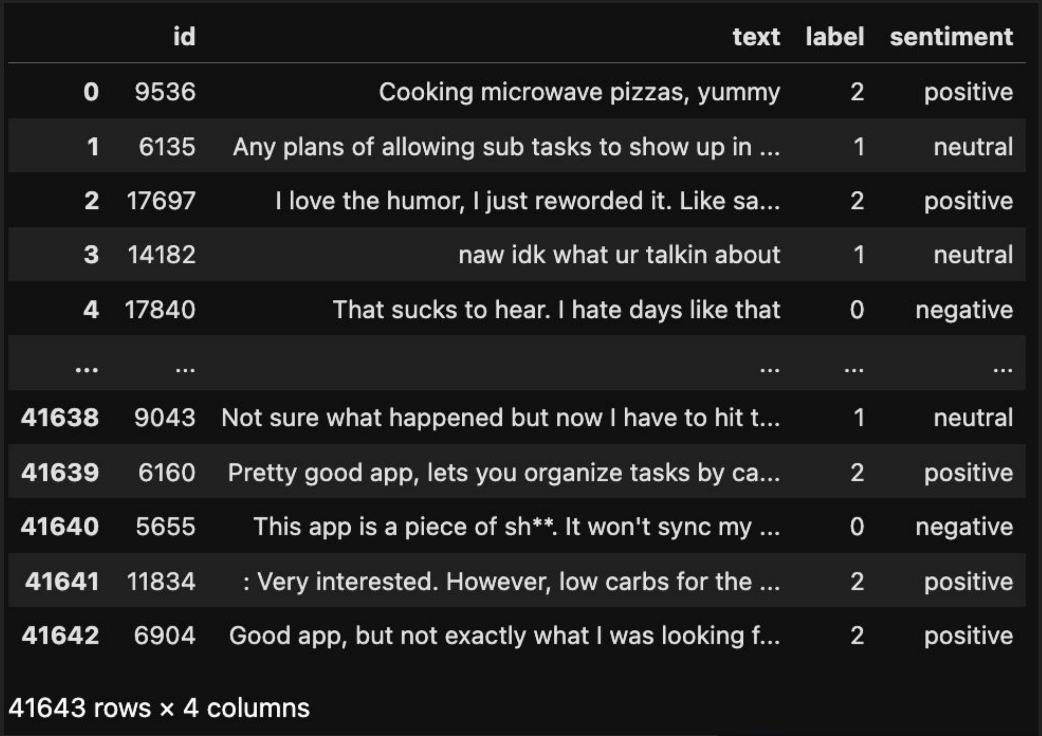
Offer companies insights from **social media comments:**

- **Campaign Sentiment Evaluation**: Analyze user feedback on campaigns across social media.
- **Sentiment Tracking**: Monitor shifts in customer sentiment over time.
- **Competitor Benchmarking**: Compare brand perception with competitors' social media sentiment.

Visualization

Demo: http://localhost:8501/

Data collection



	id	text	label	sentiment
0	9536	Cooking microwave pizzas, yummy	2	positive
1	6135	Any plans of allowing sub tasks to show up in ...	1	neutral
2	17697	I love the humor, I just reworded it. Like sa...	2	positive
3	14182	naw idk what ur talkin about	1	neutral
4	17840	That sucks to hear. I hate days like that	0	negative
...	...	...	...	...
41638	9043	Not sure what happened but now I have to hit t...	1	neutral
41639	6160	Pretty good app, lets you organize tasks by ca...	2	positive
41640	5655	This app is a piece of sh**. It won't sync my ...	0	negative
41641	11834	: Very interested. However, low carbs for the ...	2	positive
41642	6904	Good app, but not exactly what I was looking f...	2	positive

41643 rows × 4 columns

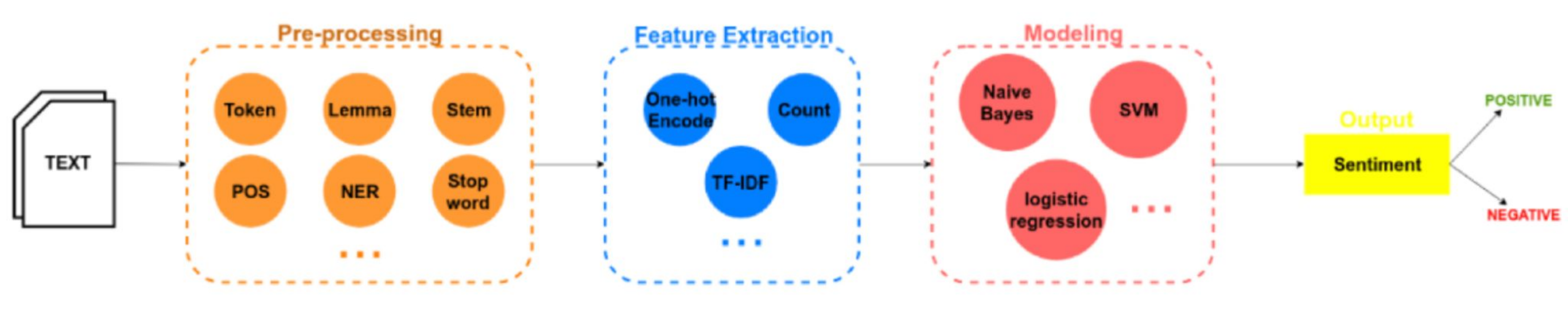
🤗 **Hugging Face**

Positive — Label: **2**

Neutral — Label: **1**

Negative — Label: **0**

Summary


TEXT
Pre-processing
Token
Lemma
Stem
POS
NER
Stop word
...
Feature Extraction
One-hot Encode
Count
TF-IDF
...
Modeling
Naive Bayes
SVM
logistic regression
...
Output
Sentiment
POSITIVE
NEGATIVE

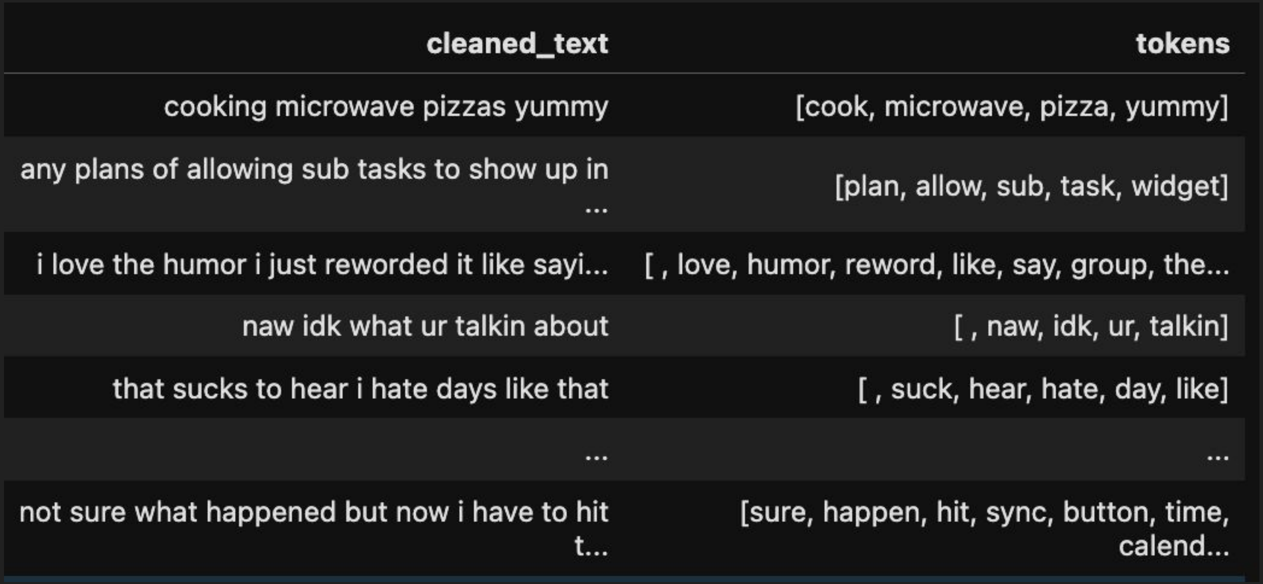
Pre- Preprocessing

- Removing URLs, mentions, hashtags, punctuation in '**text**' column, and convert it to lowercase.

- Remove **stop words**

- **Tokenization**
- **Lemmatization**

spaCy



	cleaned_text	tokens
	cooking microwave pizzas yummy	[cook, microwave, pizza, yummy]
	any plans of allowing sub tasks to show up in ...	[plan, allow, sub, task, widget]
	i love the humor i just reworded it like sayi...	[, love, humor, reword, like, say, group, the...
	naw idk what ur talkin about	[, naw, idk, ur, talkin]
	that sucks to hear i hate days like that	[, suck, hear, hate, day, like]
	...	...
	not sure what happened but now i have to hit t...	[sure, happen, hit, sync, button, time, calend...

Data Cleaning and Preprocessing

advanced Natural Language Processing (NLP) library designed for large-scale text processing.

Key Features of spaCy:

1. **Tokenization**: Breaks down text into words, punctuation, or symbols (called tokens).
2. **Part-of-Speech Tagging**: Assigns tags like noun, verb, adjective to each word.
3. **Named Entity Recognition (NER)**: Identifies entities such as names of people, organizations, or locations.
4. **Dependency Parsing**: Shows how words in a sentence are related.
5. **Pre-trained Models**: spaCy provides pre-built models for multiple languages to perform various NLP tasks.
6. **Lemmatization**: Reduces words to their base or root form.

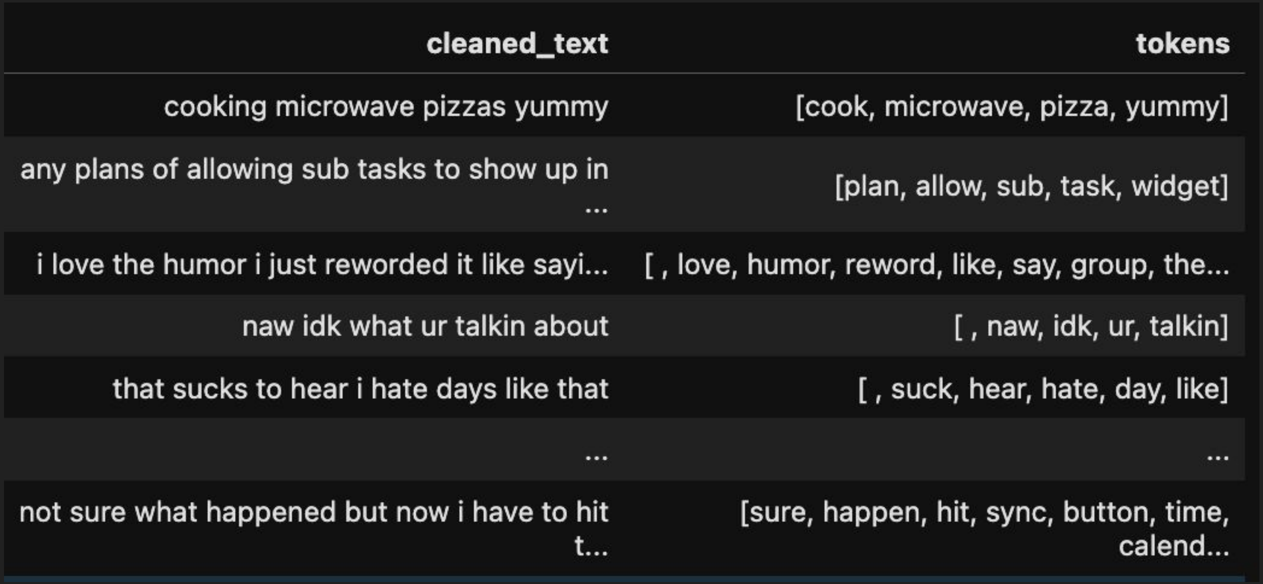
Pre- Preprocessing

- Removing URLs, mentions, hashtags, punctuation in '**text**' column, and convert it to lowercase.

- Remove **stop words**

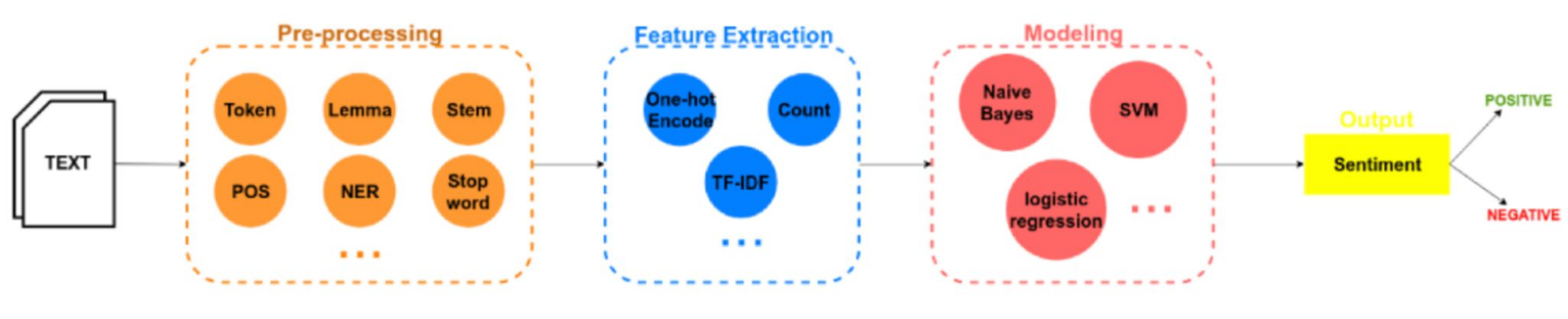
- **Tokenization**
- **Lemmatization**

spaCy



cleaned_text	tokens
cooking microwave pizzas yummy	[cook, microwave, pizza, yummy]
any plans of allowing sub tasks to show up in ...	[plan, allow, sub, task, widget]
i love the humor i just reworded it like sayi...	[, love, humor, reword, like, say, group, the...
naw idk what ur talkin about	[, naw, idk, ur, talkin]
that sucks to hear i hate days like that	[, suck, hear, hate, day, like]
...	...
not sure what happened but now i have to hit t...	[sure, happen, hit, sync, button, time, calend...

Summary

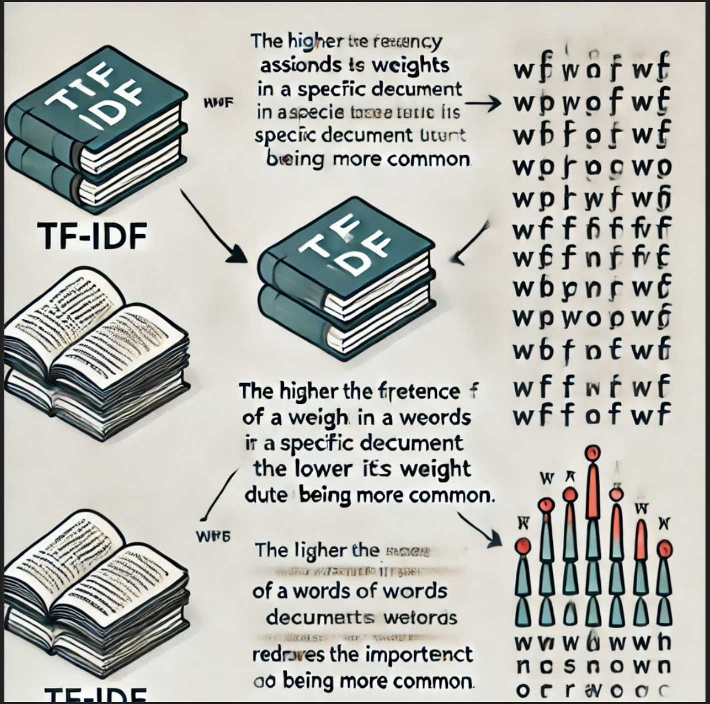

TEXT
Pre-processing
Token
Lemma
Stem
POS
NER
Stop word
...
Feature Extraction
One-hot Encode
Count
TF-IDF
...
Modeling
Naive Bayes
SVM
logistic regression
...
Output
Sentiment
POSITIVE
NEGATIVE

Fitting the **TF-IDF vectorizer**

(Term Frequency-Inverse Document Frequency)

- **Learning the Vocabulary:** The vectorizer scans through your **training data** and learns all the unique words (tokens) that appear.

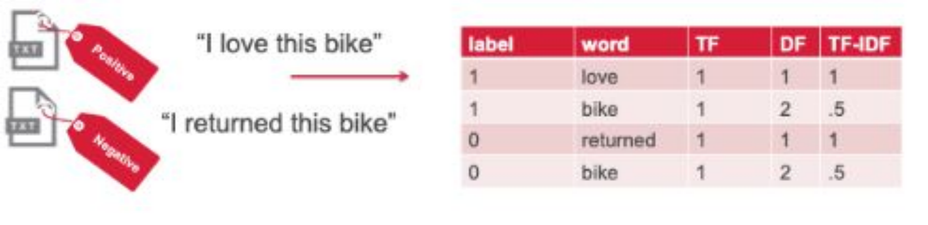
- **Assigning Weights (Importance)** to Words: It computes how frequent a word is in each document (term frequency, TF) and how rare or common that word is across the entire dataset (inverse document frequency, IDF).


TF-IDF
The higher tre frequency assionds its weights in a specfic decument in a specie tooea tatic its specfic document utunt being more common
The higher the fretence of a weigh in a weords ir a specfic decument the lower its weight dute being more common.
The ligher the of a words of words decuments wetords redroves the importenct ao being more common.

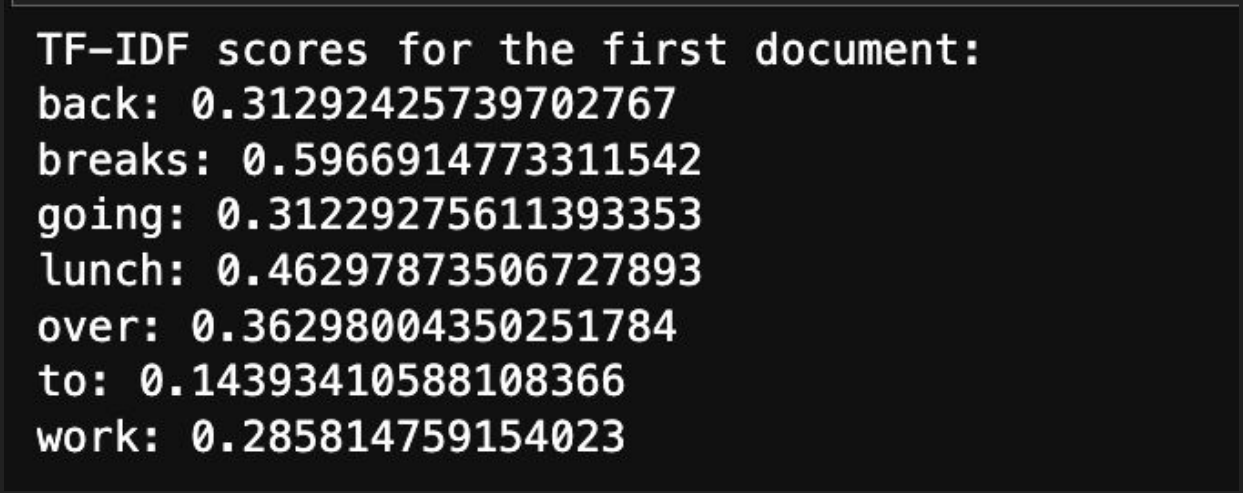
Vectorize the Text Using **TF-IDF (feature extraction)**

Use **TF-IDF** to convert the text into a numerical format:

- Transform the training and testing data

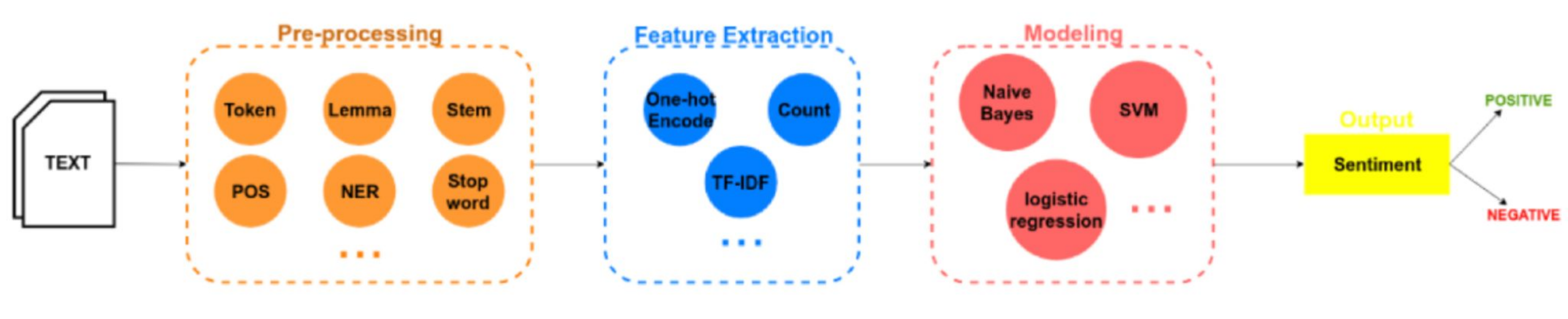


label	word	TF	DF	TF-IDF
1	love	1	1	1
1	bike	1	2	.5
0	returned	1	1	1
0	bike	1	2	.5



```
TF–IDF scores for the first document:
back: 0.31292425739702767
breaks: 0.596691477331542
going: 0.31229275611393353
lunch: 0.46297873506727893
over: 0.36298004350251784
to: 0.14393410588108366
work: 0.285814759154023
```

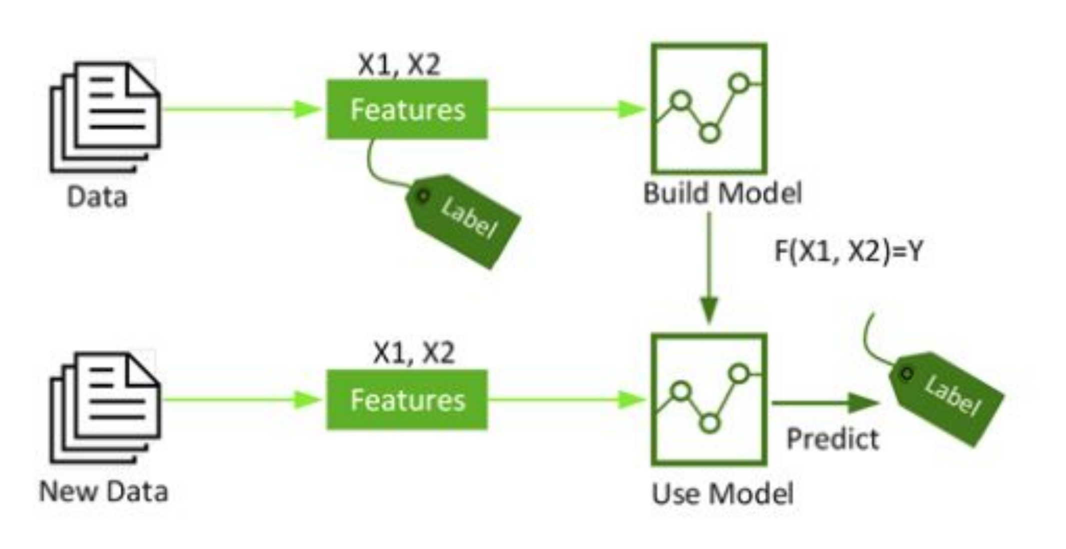
Summary



TEXT
Pre-processing
Token
Lemma
Stem
POS
NER
Stop word
...
Feature Extraction
One-hot Encode
Count
TF-IDF
...
Modeling
Naive Bayes
SVM
logistic regression
...
Output
Sentiment
POSITIVE
NEGATIVE

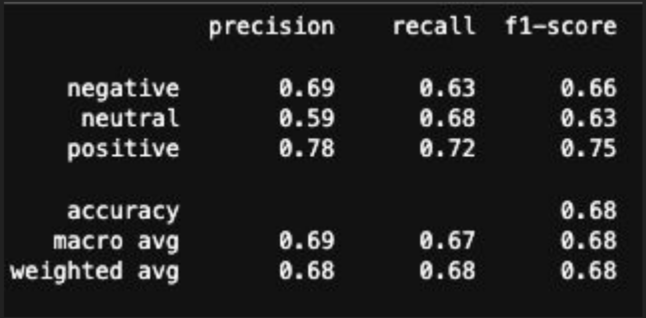
Modeling: **Logistic Regression**

It works well for classification tasks, and sentiment analysis is a classic case of **multi-class classification** (with classes being positive, neutral, and negative).

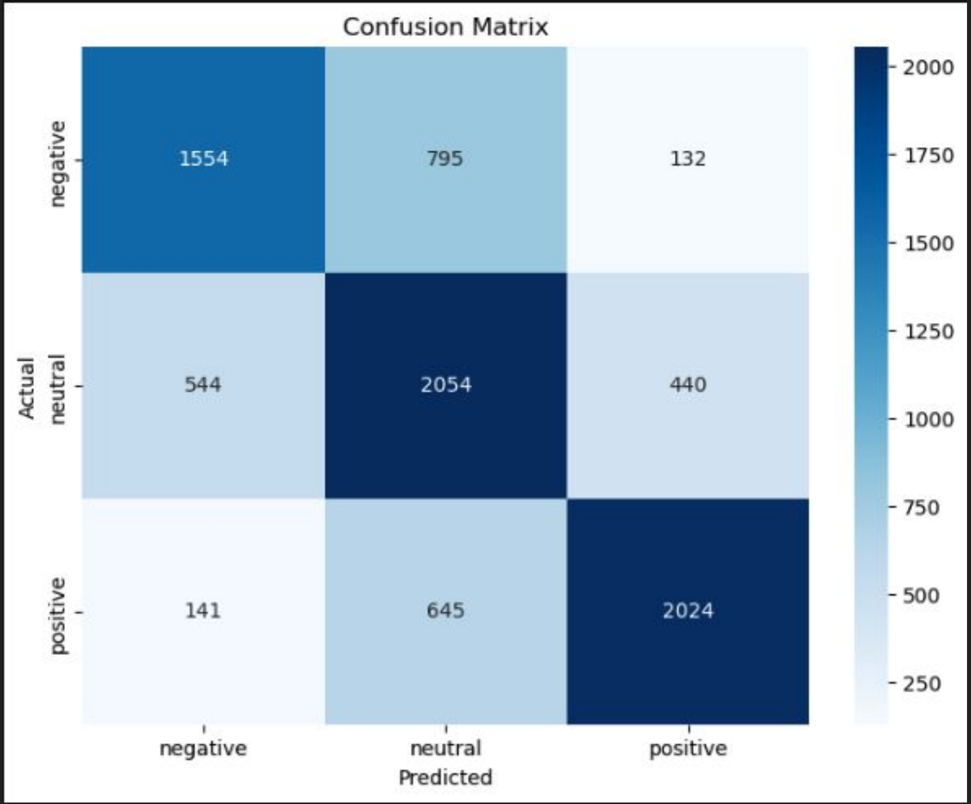


Predictive model in Python

Evaluation



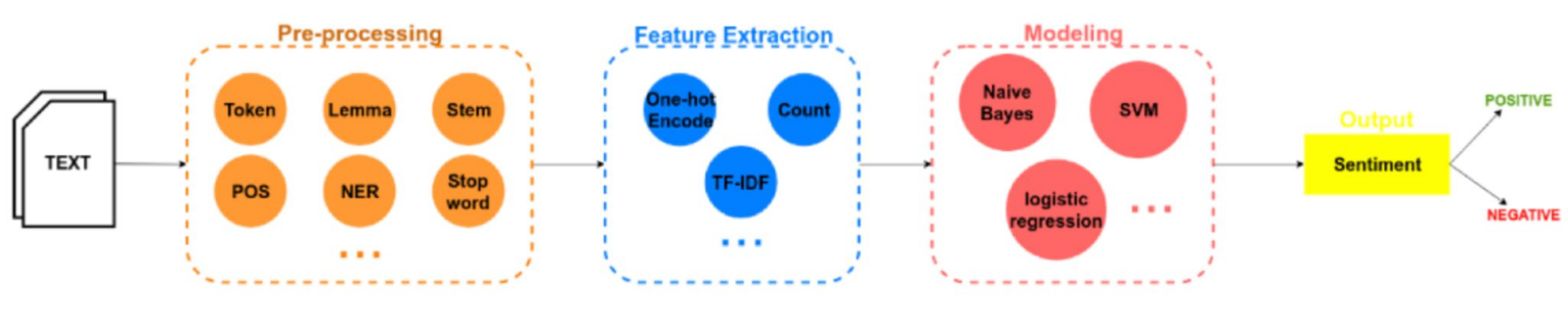
```
              precision    recall  f1-score

    negative       0.69      0.63      0.66
     neutral       0.59      0.68      0.63
    positive       0.78      0.72      0.75

    accuracy                           0.68
   macro avg       0.69      0.67      0.68
weighted avg       0.68      0.68      0.68
```



Confusion Matrix

Actual \ Predicted	negative	neutral	positive
negative	1554	795	132
neutral	544	2054	440
positive	141	645	2024

Conclusion


TEXT
Pre-processing
Token
Lemma
Stem
POS
NER
Stop word
...
Feature Extraction
One-hot Encode
Count
TF-IDF
...
Modeling
Naive Bayes
SVM
logistic regression
...
Output
Sentiment
POSITIVE
NEGATIVE

Visualization

Demo: http://localhost:8501/