CSE 256 Demo Report

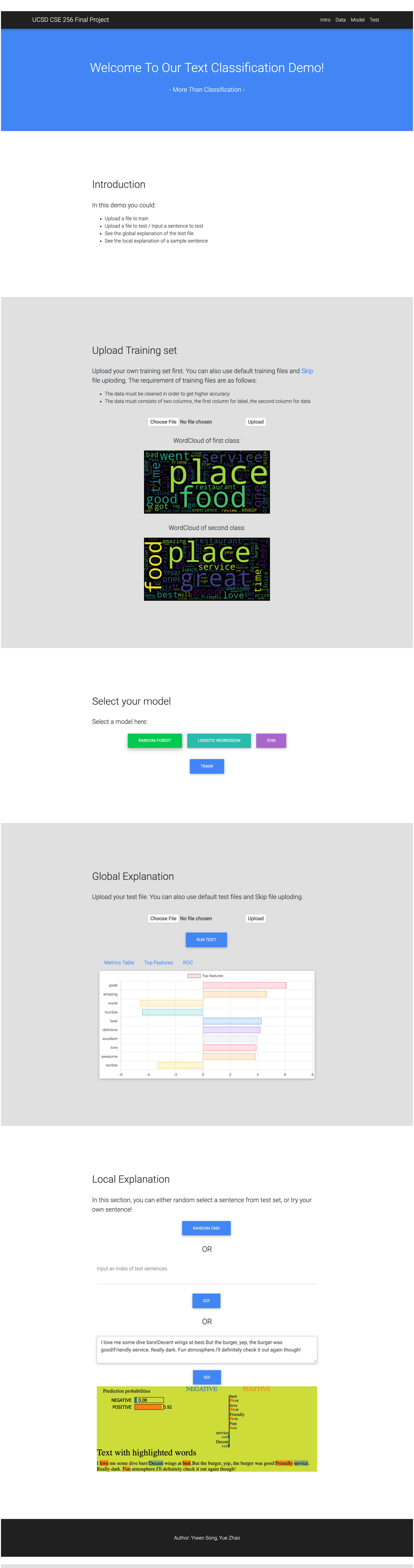
Yiwen Song, Yue Zhao

Overview

In this project, we've developed a pipeline for text classification. The steps to take are as follows:

- Firstly, you may **upload** a **train file**.
    - This train file must be of the following format: It consists of 2 columns, the first column is the label and the second column is the related sentence. Note that all the text in the train file must have been cleaned before uploading.
    - After uploading the train file, you will see n **WordCloud** images of the high-frequency words for each class, which will be a preview of the training data.
- Secondly, you may **select a model** from `RandomForestClassifier`, `LogisticRegressionClassifier` and `SVMClassifier`.
- Secondly, you may **upload** a **test file**.
    - This test file must be of the same format as the train file.
- Next, you may click the `run test` button AND see the **global explanation** results. The results are divided into the following three parts:
    - The metrics table of the **precision**, **recall** and **F1-score** of each class.
    - The **top features** of the model.
    - The **ROC curve** on this test set.
- Finally, you could go to the last section to see the **local explanation** of each sentence. You may
    - See the explanation of the prediction of a **random sentence** from the test set.
    - See the explanation of the prediction of a **self-indexed sentence** from the test set.
    - See the explanation of the prediction of a **self-input sentence**.

☐ The figure below shows how this project works:



UCSD CSE 256 Final Project
Intro
Data
Model
Test

Welcome To Our Text Classification Demo!

- More Than Classification -

Introduction

In this demo you could:

- Upload a file to train
- Upload a file to test / Input a sentence to test
- See the global explanation of the test file
- See the local explanation of a sample sentence

Upload Training set

Upload your own training set first. You can also use default training files and Skip file uploding. The requirement of training files are as follows:

- The data must be cleaned in order to get higher accuracy
- The data must consists of two columns, the first column for label, the second column for data

WordCloud of first class:

WordCloud of second class:

Select your model

Select a model here:

Choose File
No file chosen
Upload
RANDOM FOREST
LOGISTIC REGRESSION
SVM
TRAIN!

Global Explanation

Upload your test file. You can also use default test files and Skip file uploding.

Choose File
No file chosen
Upload
RUN TEST!
Metrics Table
Top Features
ROC
Top features
great
amazing
worst
horrible
best
delicious
excellent
love
awesome
terrible
-6
-4
-2
0
2
4
6
8

Local Explanation

In this section, you can either random select a sentence from test set, or try your own sentence!

RANDOM ONE!
OR
Input an index of test sentences.
GO!
OR
GO!

I love me some dive bars!Decent wings at best.But the burger, yep, the burger was good!Friendly service. Really dark. Fun atmosphere.I'll definitely check it out again though!

Prediction probabilities

NEGATIVE 0.08

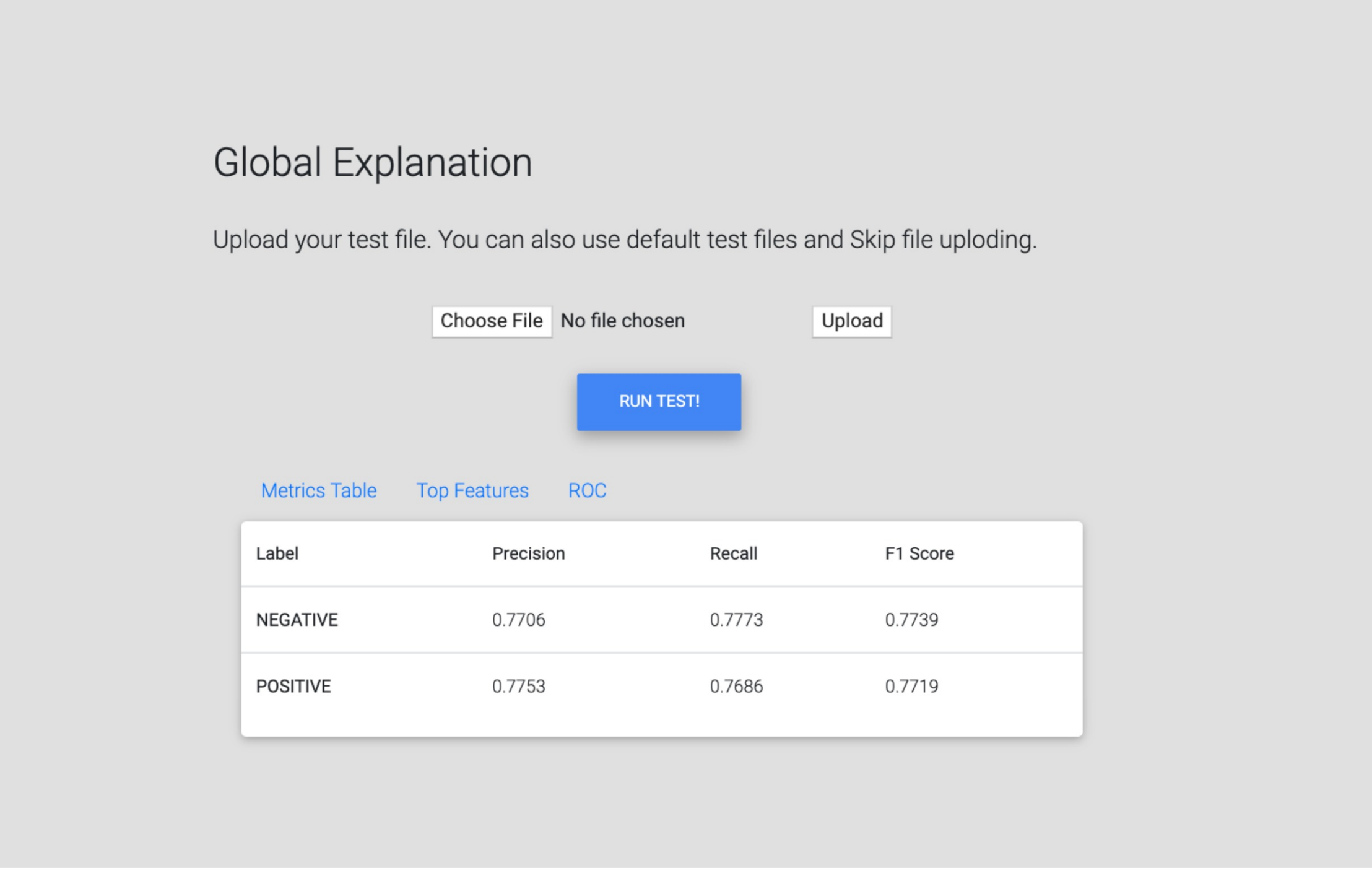
POSITIVE 0.92

NEGATIVE
POSITIVE
best
love
Friendly
Fun
service
Decent

Text with highlighted words

I love me some dive bars!Decent wings at best.But the burger, yep, the burger was good!Friendly service. Really dark. Fun atmosphere.I'll definitely check it out again though!

Author: Yiwen Song, Yue Zhao

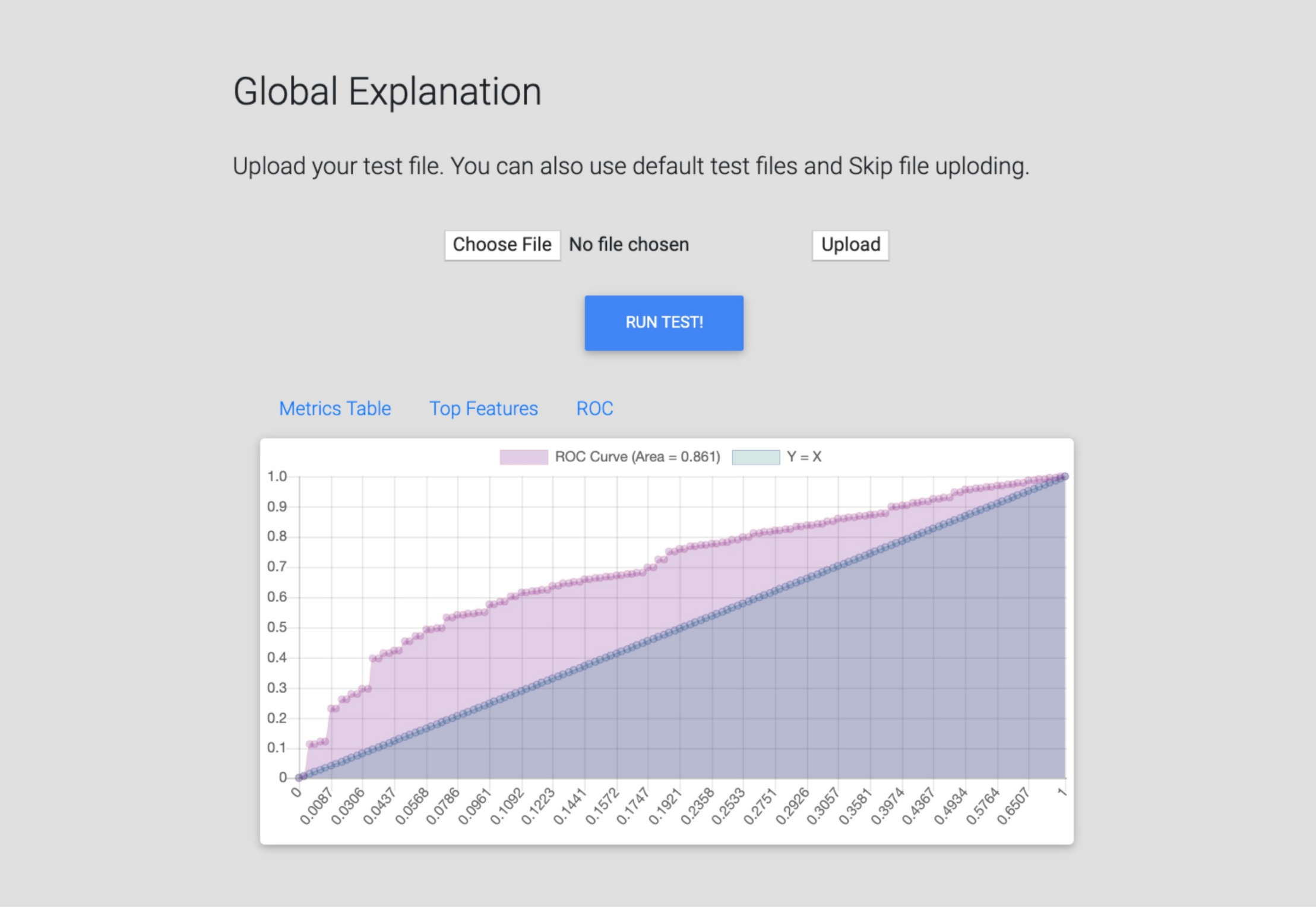


Global Explanation

Upload your test file. You can also use default test files and Skip file uploding.

Choose File
No file chosen
Upload
RUN TEST!
Metrics Table
Top Features
ROC

Label	Precision	Recall	F1 Score
NEGATIVE	0.7706	0.7773	0.7739
POSITIVE	0.7753	0.7686	0.7719



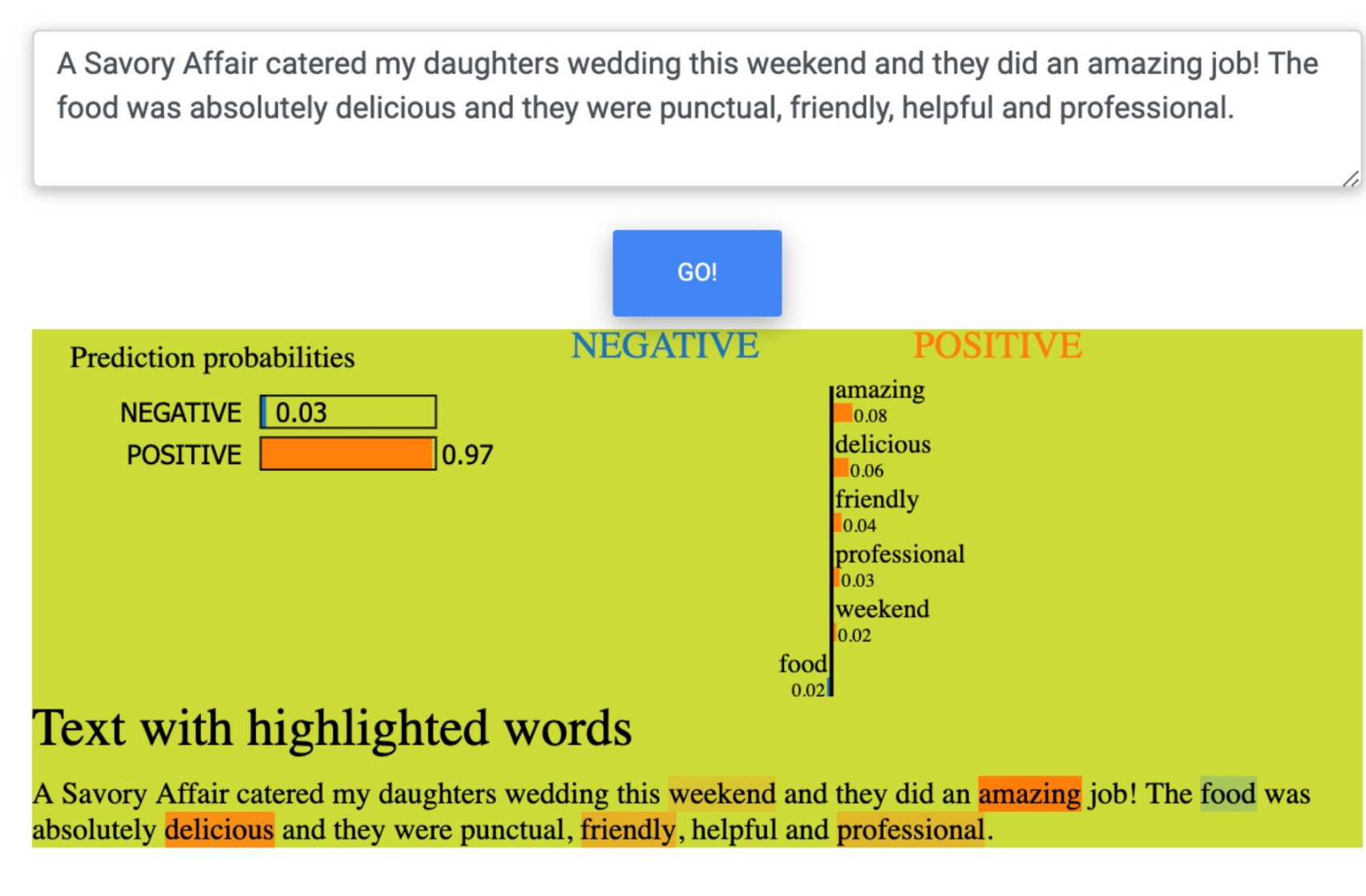
Global Explanation

Upload your test file. You can also use default test files and Skip file uploding.

Choose File
No file chosen
Upload
RUN TEST!
Metrics Table
Top Features
ROC
ROC Curve (Area = 0.861)
Y = X
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

Part 1

1.1 My favirote review



A Savory Affair catered my daughters wedding this weekend and they did an amazing job! The food was absolutely delicious and they were punctual, friendly, helpful and professional.

GO!
NEGATIVE
POSITIVE
amazing
delicious
friendly
professional
weekend
food

Prediction probabilities

NEGATIVE 0.03

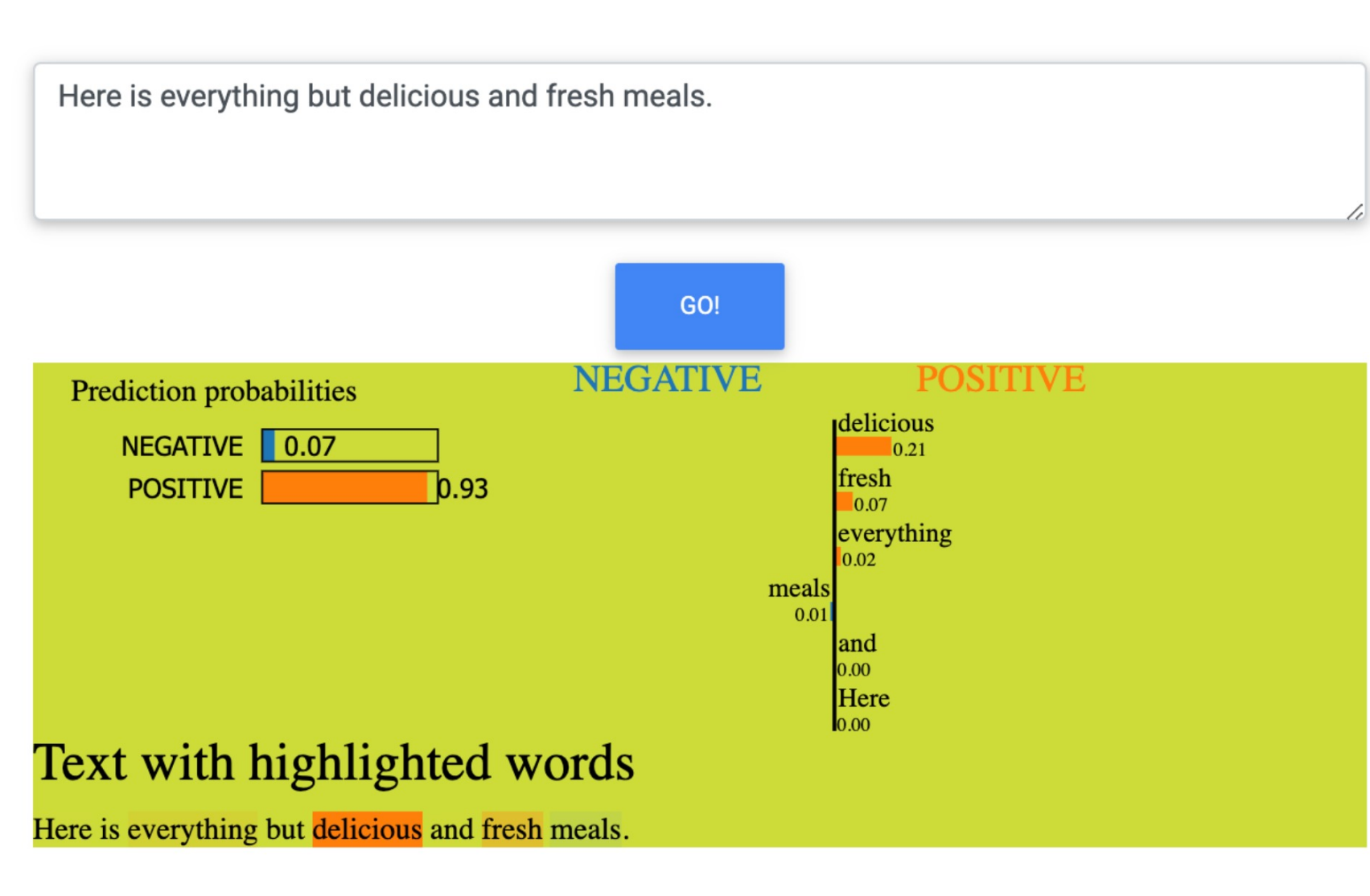
POSITIVE 0.97

Text with highlighted words

A Savory Affair catered my daughters wedding this weekend and they did an amazing job! The food was absolutely delicious and they were punctual, friendly, helpful and professional.

1.2 Over-Confident Review

☐ This sentence should be negative, but is classified as positive with confidence 0.93:



Here is everything but delicious and fresh meals.

GO!
NEGATIVE
POSITIVE
delicious
fresh
everything
meals
and
Here

Prediction probabilities

NEGATIVE 0.07

POSITIVE 0.93

Text with highlighted words

Here is everything but delicious and fresh meals.

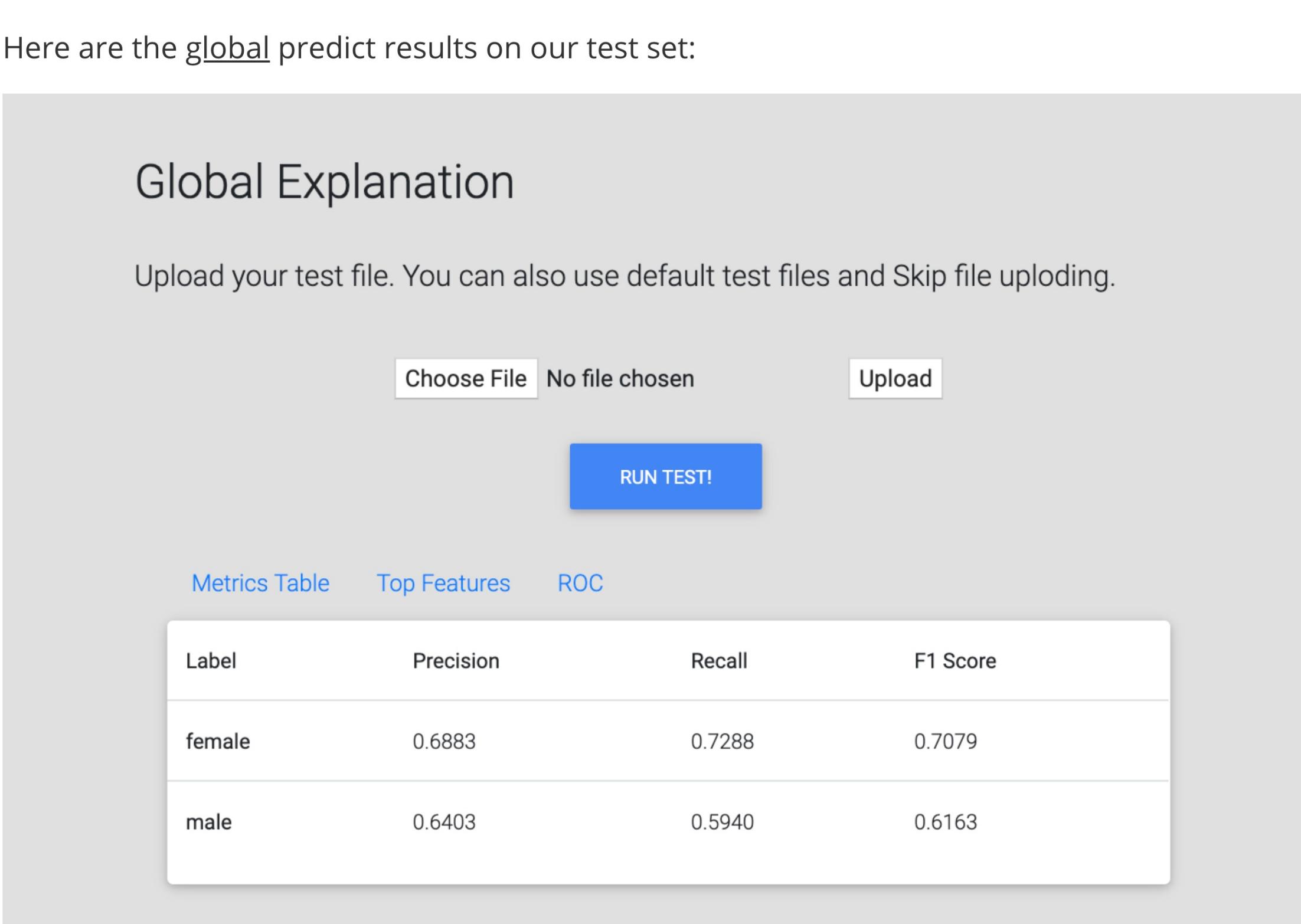
Part 2

In this part, we use the data set download from https://www.kaggle.com/crowdflower/twitter-user-gender-classification, the task is to classify the gender of the twitter users.

- The data we used are the twitter text and the users' description text
- The labels are `male` and `female`
- When cleaning the dataset, we remove all the samples with a label confidence value < 0.9

2.1 Global Explanations

Here are the global predict results on our test set:

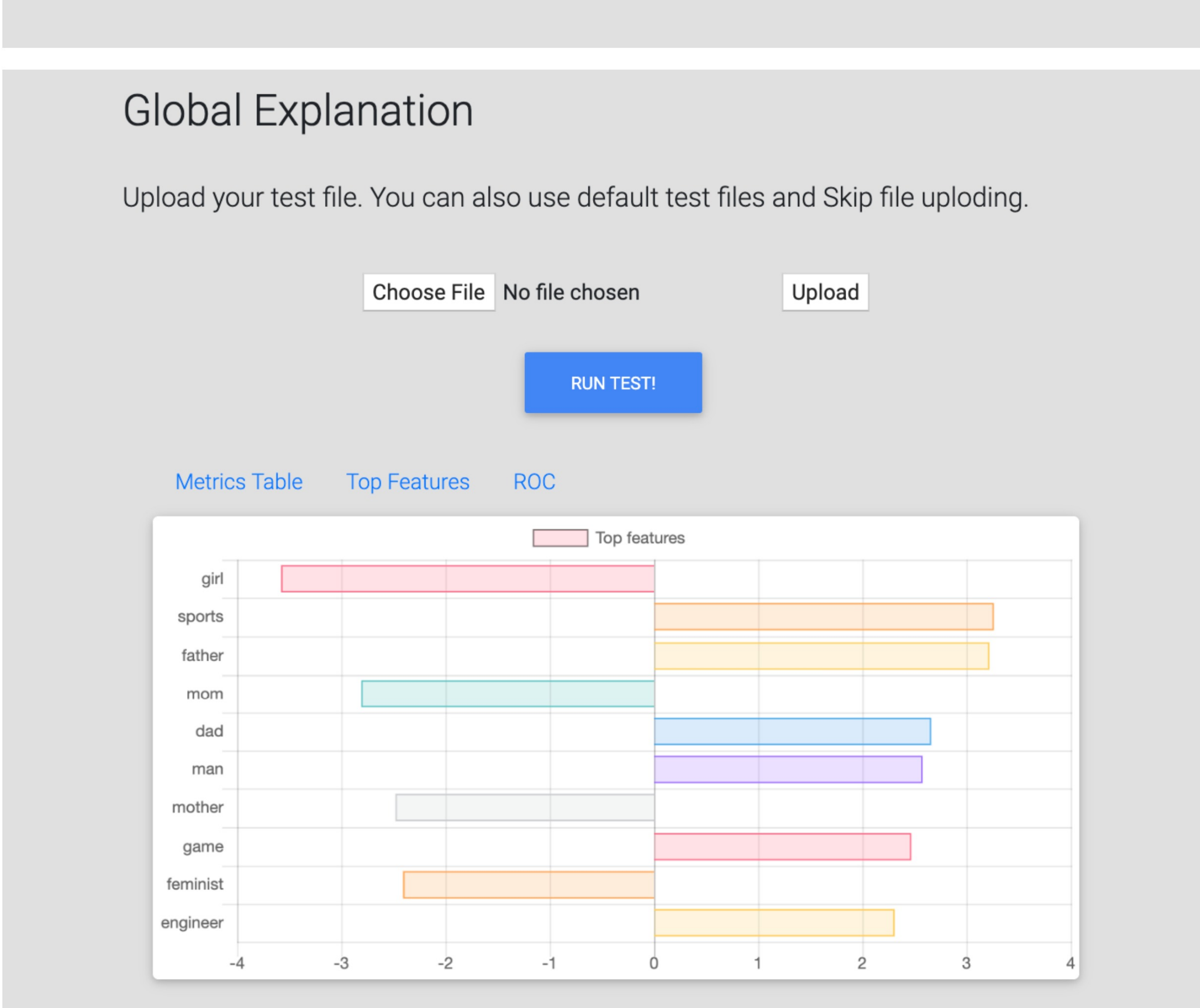


Global Explanation

Upload your test file. You can also use default test files and Skip file uploding.

Choose File
No file chosen
Upload
RUN TEST!
Metrics Table
Top Features
ROC

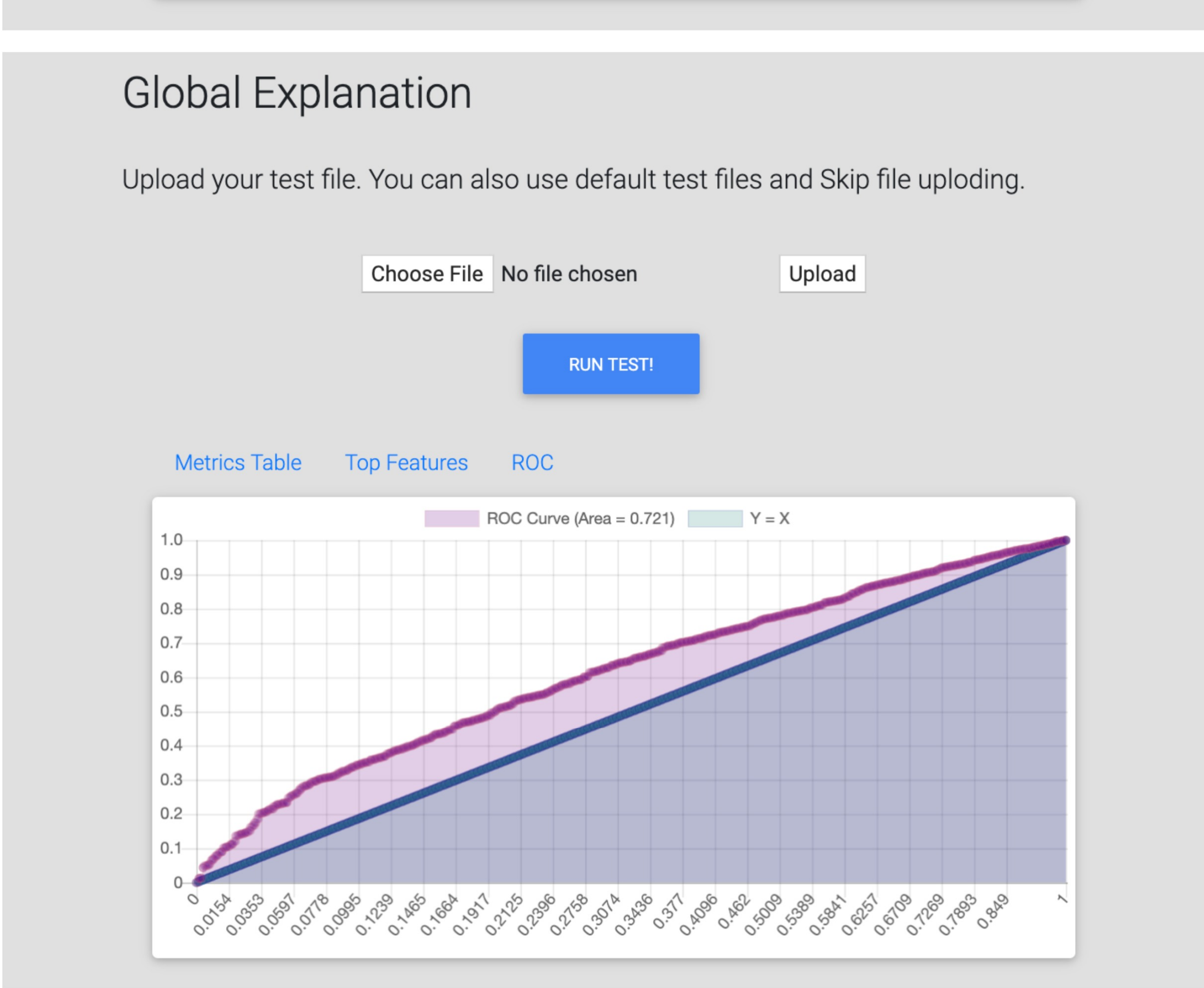
Label	Precision	Recall	F1 Score
female	0.6883	0.7288	0.7079
male	0.6403	0.5940	0.6163



Global Explanation

Upload your test file. You can also use default test files and Skip file uploding.

Choose File
No file chosen
Upload
RUN TEST!
Metrics Table
Top Features
ROC
Top features
girl
sports
father
mom
dad
man
mother
game
feminist
engineer
-4
-3
-2
-1
0
1
2
3
4



Global Explanation

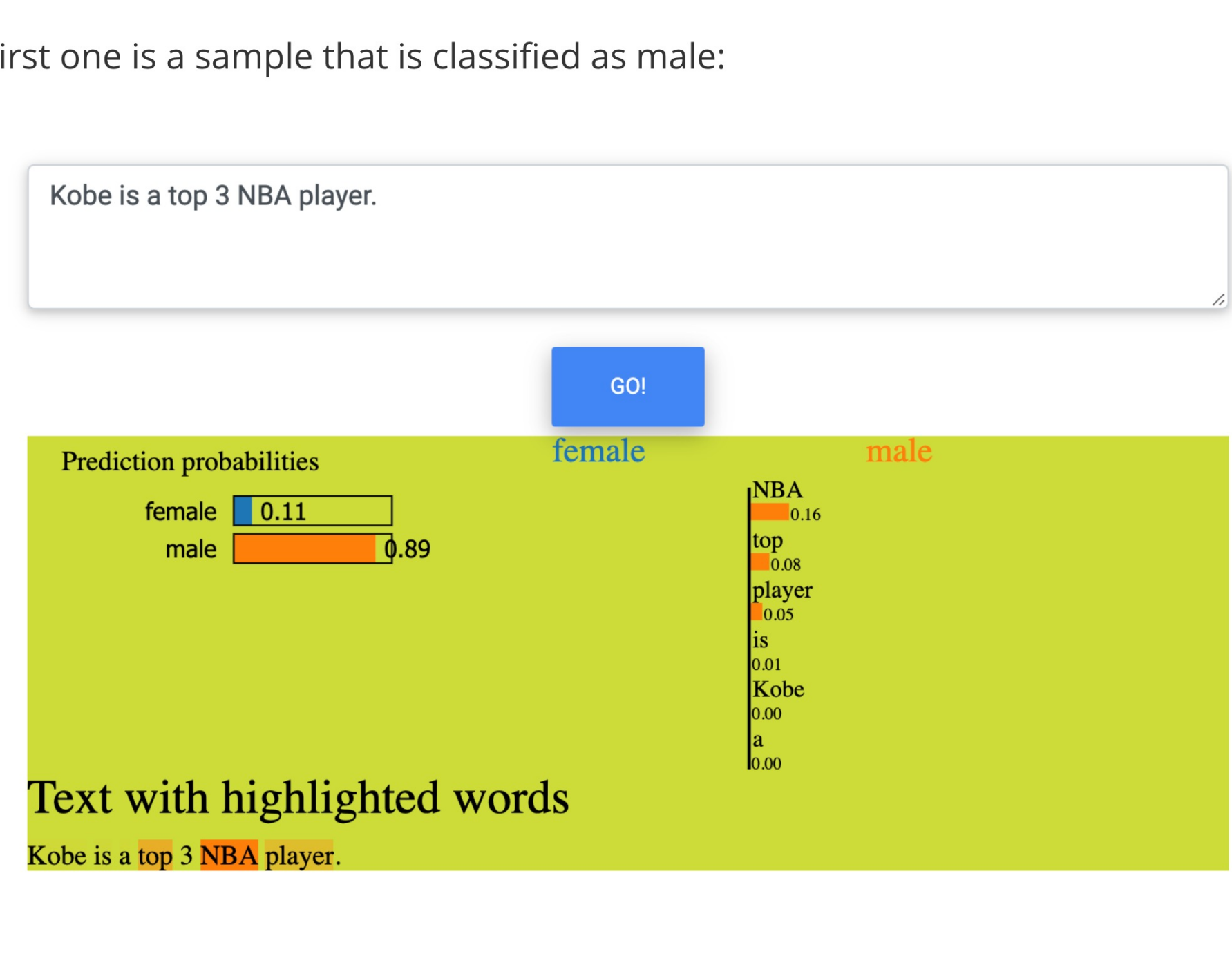
Upload your test file. You can also use default test files and Skip file uploding.

Choose File
No file chosen
Upload
RUN TEST!
Metrics Table
Top Features
ROC
ROC Curve (Area = 0.721)
Y = X
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

2.2 Locla Explanations

Here are two samples of our local explanation:

1. The first one is a sample that is classified as male:



Kobe is a top 3 NBA player.

GO!
female
male
NBA
top
player
is
Kobe
a

Prediction probabilities

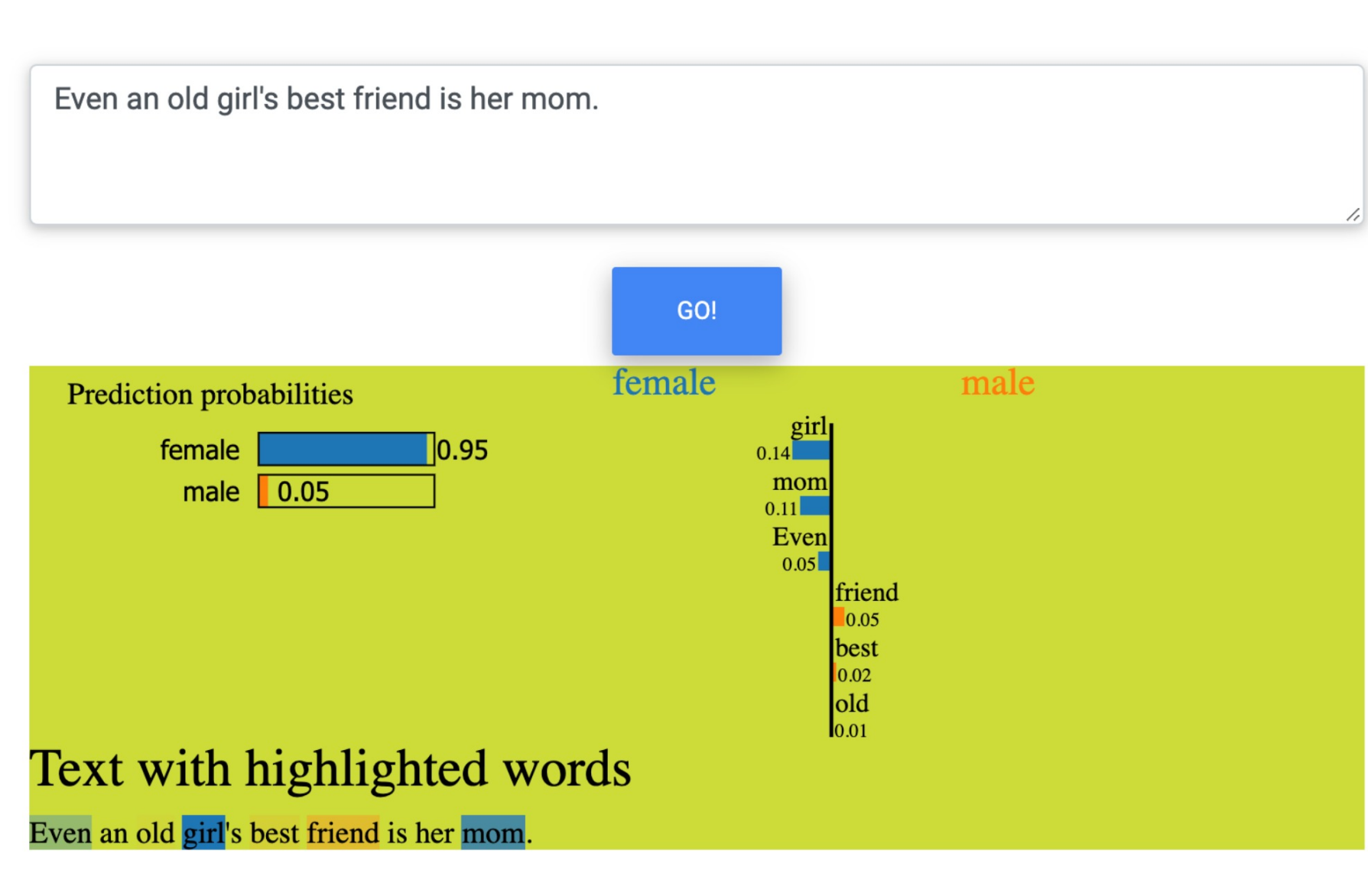
female 0.11

male 0.89

Text with highlighted words

Kobe is a top 3 NBA player.

2. The second one is a sample that is classified as female:



Even an old girl's best friend is her mom.

GO!
female
male
girl
mom
Even
friend
best
old

Prediction probabilities

female 0.95

male 0.05

Text with highlighted words

Even an old girl's best friend is her mom.

Part 3

The creative points are as follows:

- We use a unified pipeline to do **part1** and **part2**, which means, this text classification system can not only be applied to the sentiment review classification and gender classsification, but also can be applied to all the binary classification tasks.
- Three different models `RandomForest`, `LogisticRegression`, and `SVM` are embedded in our system for users to choose.
- Detailed global explanation are provided, so users can have an understanding about the overall performance.

Tech stack

- **Flask** is used for backend and **Bootstrap** is used for frontend.
- **Jquery** is used for dynamically rendering the html.
- **CSS** is used for animation and style design.
- **Sklearn, NLTK** are used for feature extraction and model training.
- **Lime** is used for local explanation.