

# Scrapy Project

*Multi-Class Text Classification  
with Scikit-Learn*

Felipe da Silva Santos

# Introduction

A screenshot of a web browser window titled "Multi-Class Text Classification". The URL is <https://towardsdatascience.com/multi-class-text-classification->. The page is from the website "Towards Data Sci...", which has a logo with a large white "M". The navigation bar includes links for HOME, DATA SCIENCE, MACHINE LEARNING, PROGRAMMING, VISUALIZATION, and PICKS. A "Get started" button is visible. On the left, there's a profile picture of Susan Li and her bio: "Susan Li  
Changing the world, one story at a time. Data. @WaveHQ. Opinions = my own.  
Feb 19 · 11 min read". The main content title is "Multi-Class Text Classification with Scikit-Learn". Below the title is a large image of a camera lens with the word "Frequently" overlaid. At the bottom, there's a "Never miss a story from Towards Data Science" button and a "GET UPDATES" button.

## Susan's article

Classifying customer finance complaints into 12 pre-defined classes.

Does it work in different situations?

# Introduction

This article is about learning costumer finance complaints using pre-defined classes.

it work in different situations?

Scrapy  
Source of data.

**JOKE CATEGORIES**

ANIMAL	BLONDE	BLUE COLLAR	CROSS THE ROAD	DARK HUMOR
DIRTY	DOCTOR	FAT	FOOD	GOD
GROSS	INSULTS	KIDS	LAWYER	LITTLE JOHNNY
LOOKIN' GOOD	MARRIAGE	MEN/WOMEN	MISCELLANEOUS	MONEY
NSFW	NATIONALITY	NEWS & POLITICS	PARTYING & BAD BEHAVIOR	PICK-UP LINES
POLICE & MILITARY	POP CULTURE & CELEBRITY	SCHOOL	SPORTS & ATHLETES	TECHNOLOGY
TRAVEL & CAR	WALKS INTO A BAR	WORK	YO' MAMA	

# Steps

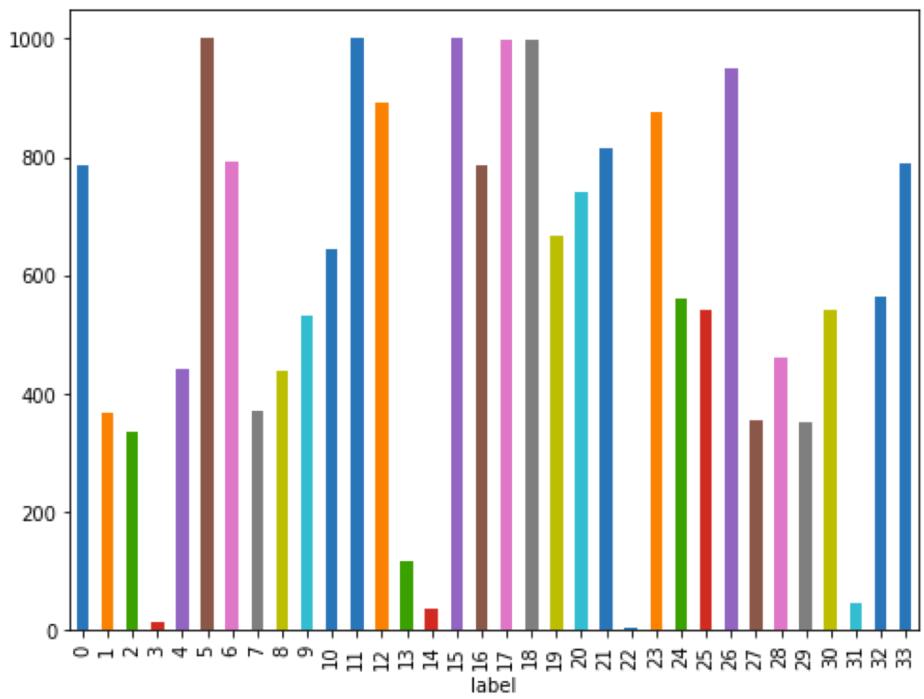


- Tf–idf term weighting
- Naive Bayes Classifier
- Random Forest
- Logistic Regression



Accuracy

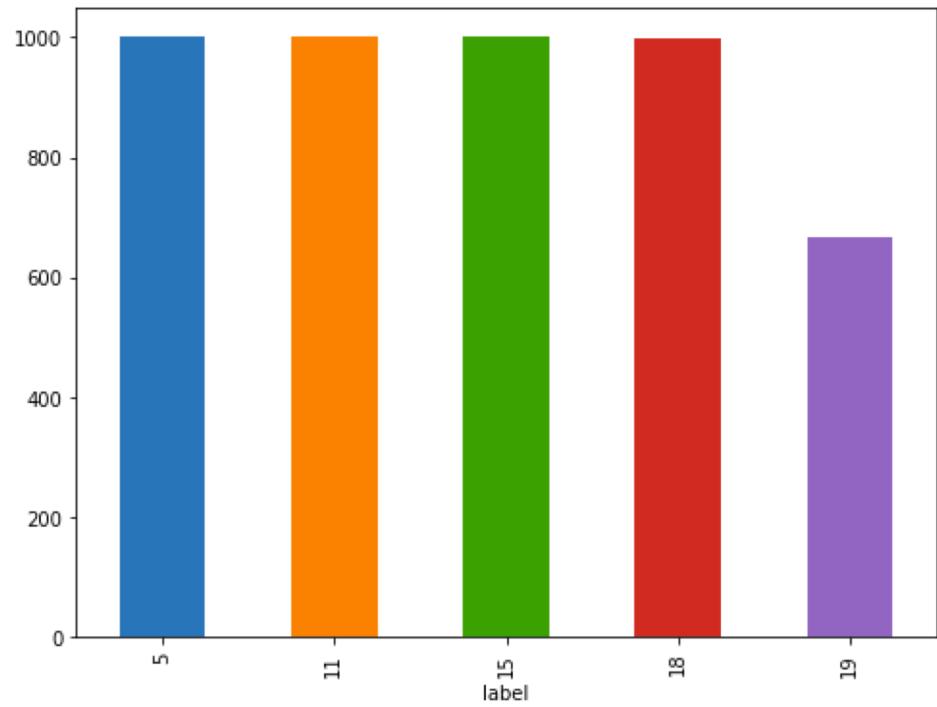
# 34 Imbalanced Classes



Model	Accuracy	
	Train	Test*
Naive Bayes Classifier	39%	17%
Random Forest	30%	15%
Logistic Regression	43%	18%

\* Test size: 20%

# 5 Balanced Classes

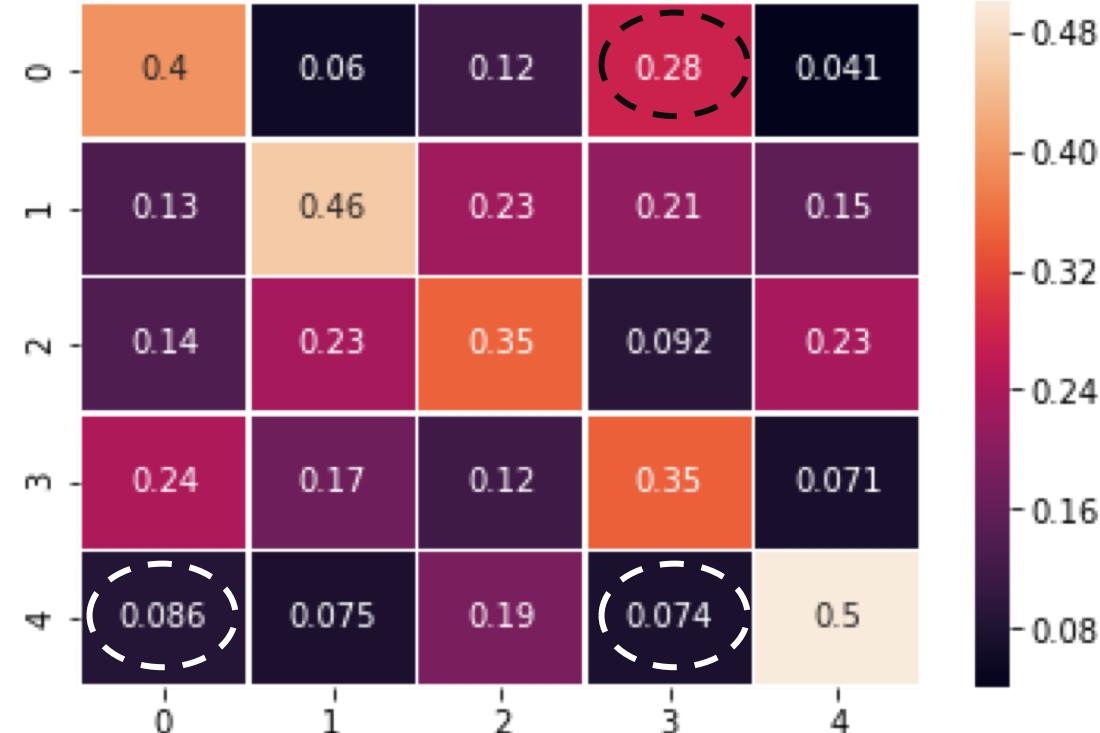


Model	Train	Test*
Naive Bayes Classifier	60%	38%
Random Forest	61%	35%
Logistic Regression	65%	39%

\* Test size: 20%

# 5 Balanced Classes

Test sample size: 934



0	dirty
1	insults
2	looking good
3	miscellaneous
4	money

Rows: Real Class

Columns: Predicted Class

Accuracy

Model	Train	Test*
Logistic Regression	65%	39%

\* Test size: 20%

# Improvements

- Tuning the models parameters properly (GridSearchCV);
- Finding the appropriate metrics to compare the model's performance.