Hogeschool van Amsterdam

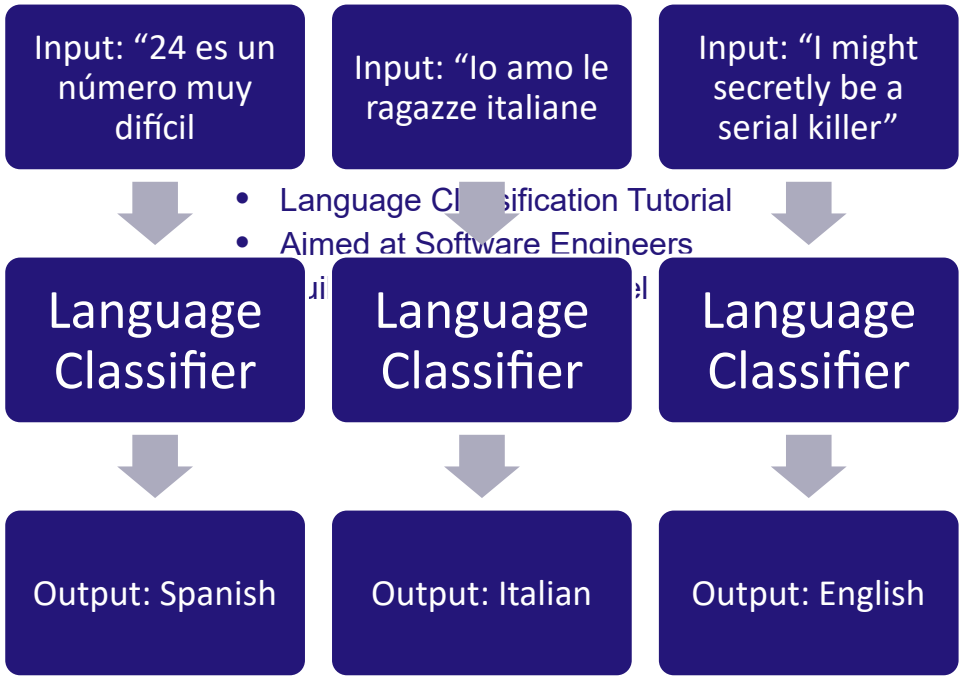
A Tutorial on Langauge Classification

Naam: Kaan Selcuk Gögcay
Datum: 16-01-2026

Creating Tomorrow

Table of content

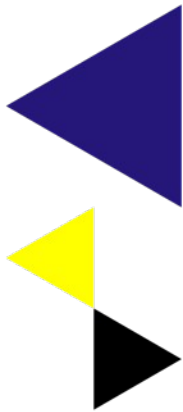
Context



- Language Classification Tutorial
- Aimed at Software Engineers

Methodology

- Exploratory Data Analysis (EDA)
- Modelling
- Evaluation (including results)
- Writing Tutorial

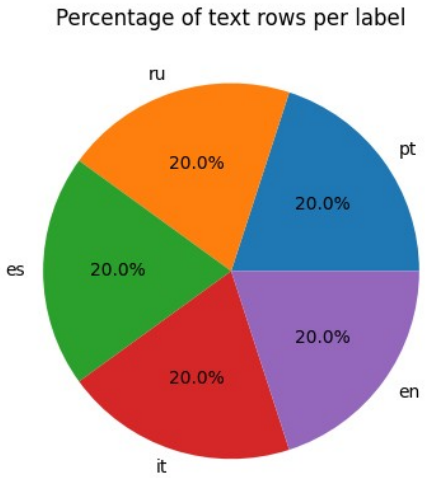


Methodology: EDA

- Language Identification dataset by Papluca
- Amazon reviews
- 20 languages
- Limited to 5 languages (because basic review)
-

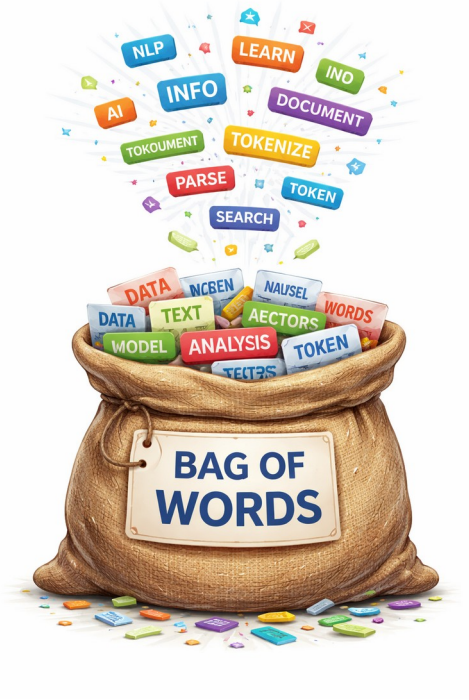
Methodology: EDA

- 5 Languages
- 3500 samples per language
- 17500 samples total
- Evenly split
- Sample: 'A little tight but ok..good quality',
- Sample: 'I cried, oh god it was ugly cries the likes of any I have ever had reading a book. Della and Ren will never be forgotten but their son..... absolutely devastated me. These books will forever be my favorites. They shall never be forgotten!',
    -
-
-



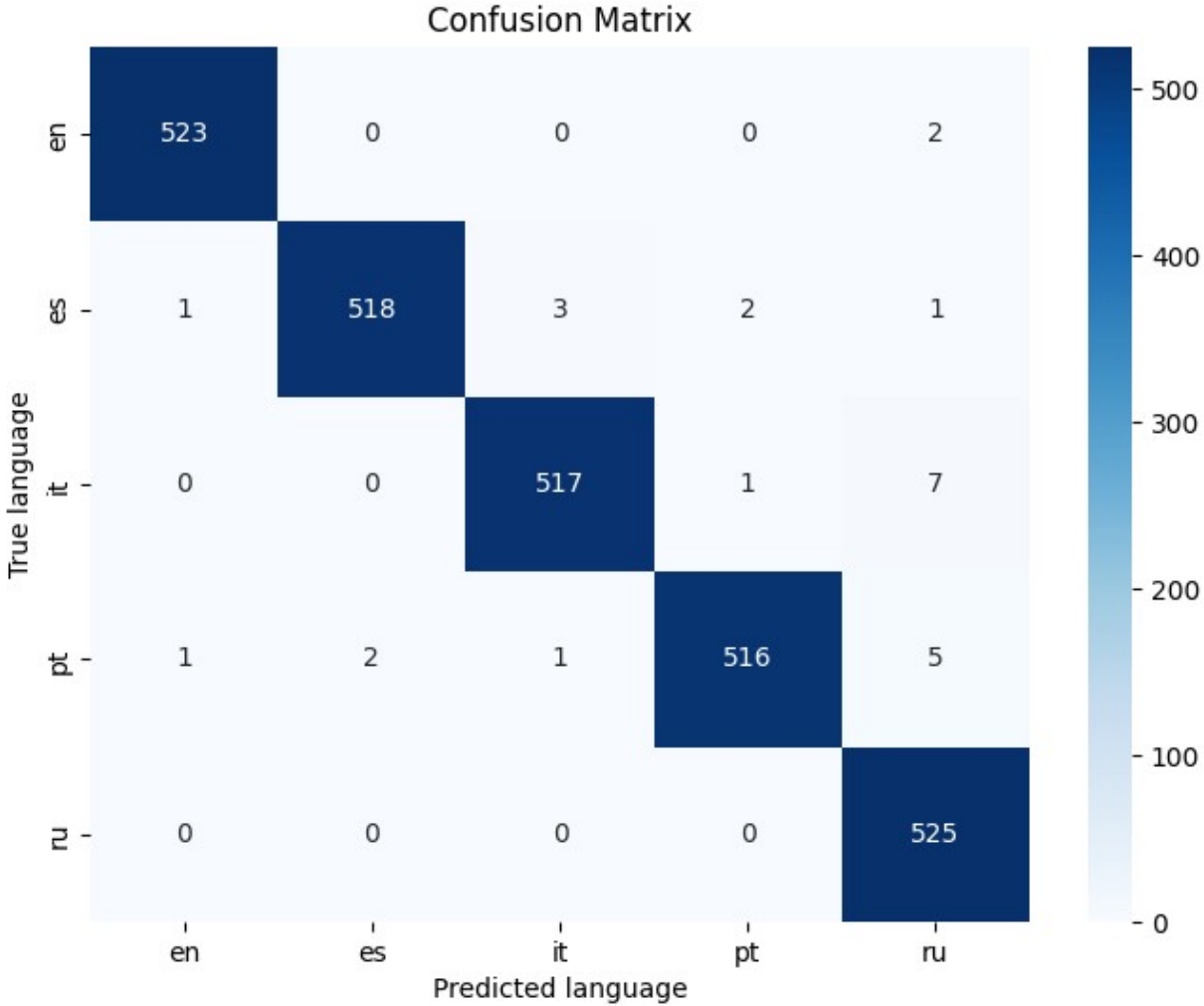
Methodology: Modelling

- Deep Learning with Python Chapter 14 Text Classification
- Bag-of-Words model
- Word level tokenization (easiest to understand)
- Vocabulary of 20000 words
-
-



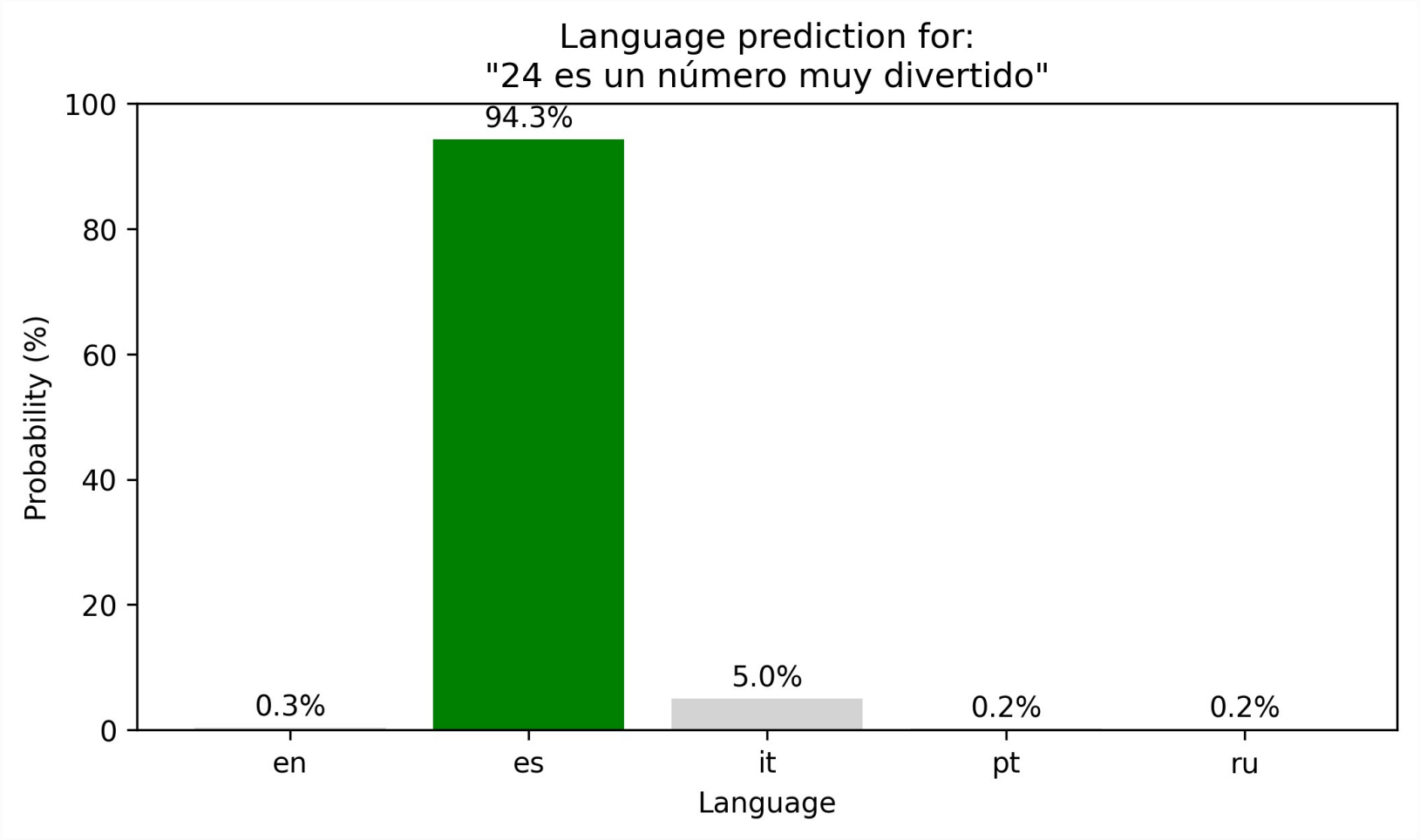
Methodology: Evaluation

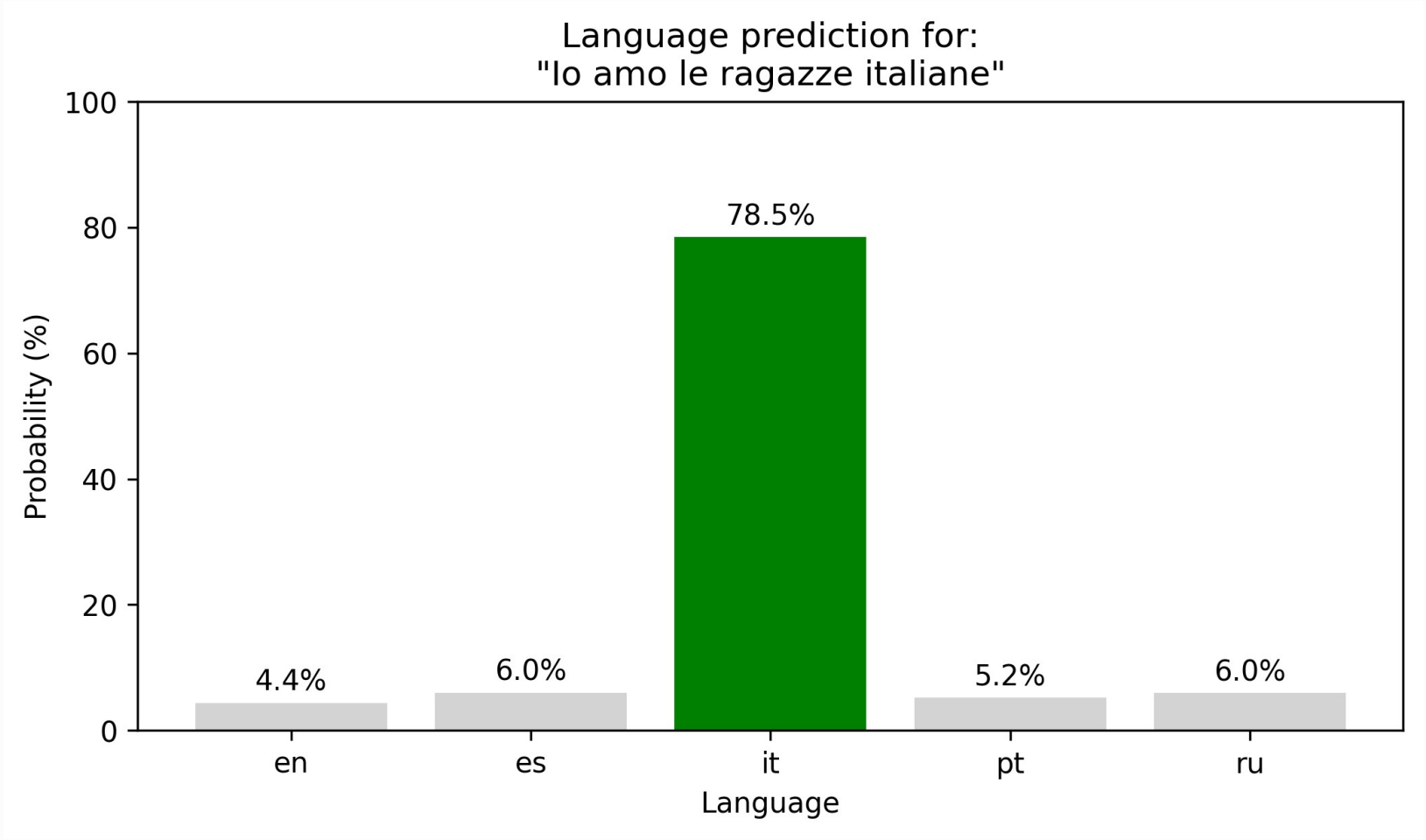
- Train accuracy: 99.94%
- Validation accuracy: 98.78%
- Test accuracy: 99.00%
- Confusion Matrix
-

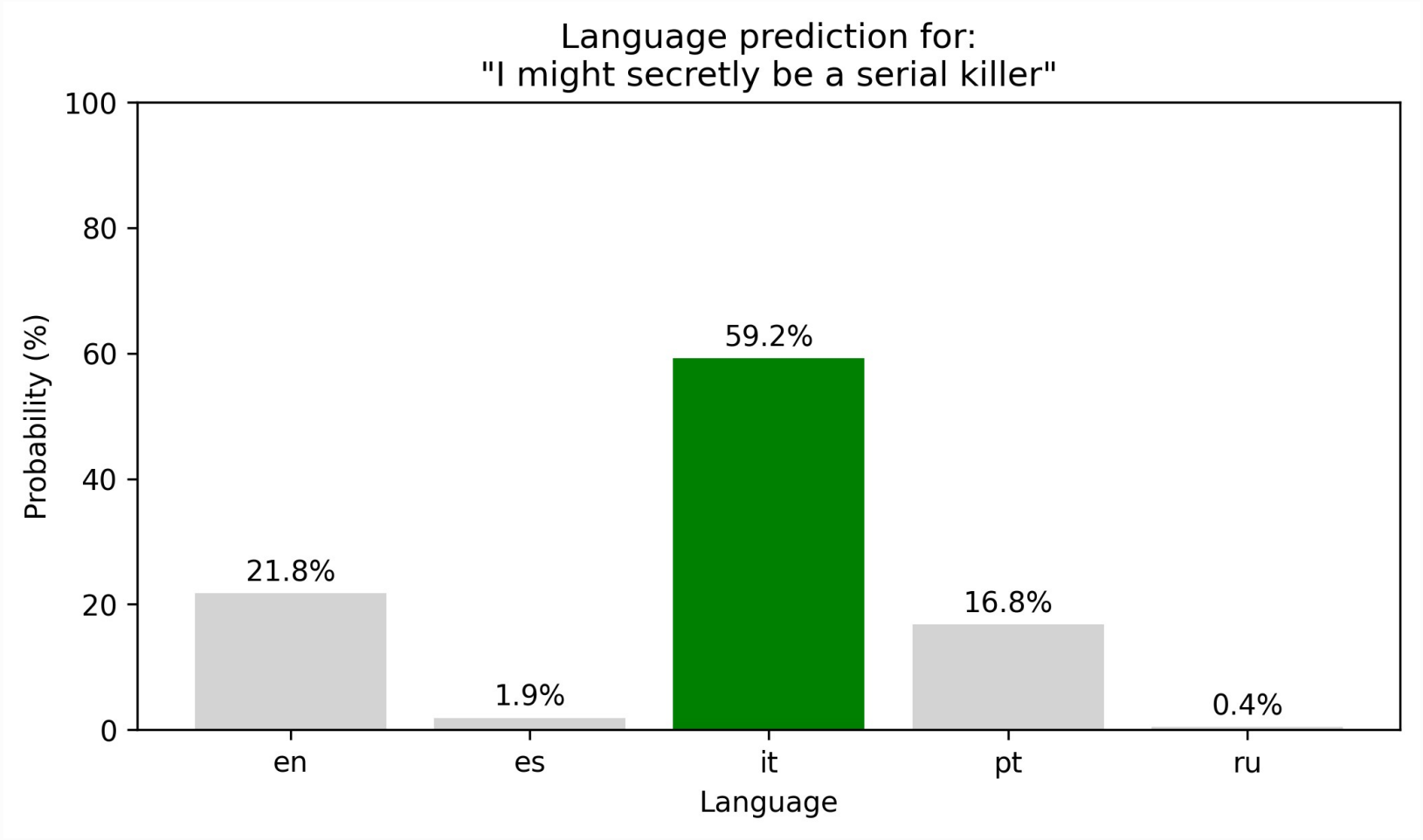

Confusion Matrix
True language
Predicted language
en
es
it
pt
ru
523 0 0 0 2
1 518 3 2 1
0 0 517 1 7
1 2 1 516 5
0 0 0 0 525
500
400
300
200
100
0

Methodology: Evaluation

- Train accuracy: 99.94%
- Validation accuracy: 98.78%
- Test accuracy: 99.00%
- Confusion Matrix
- Classifying sentences
-

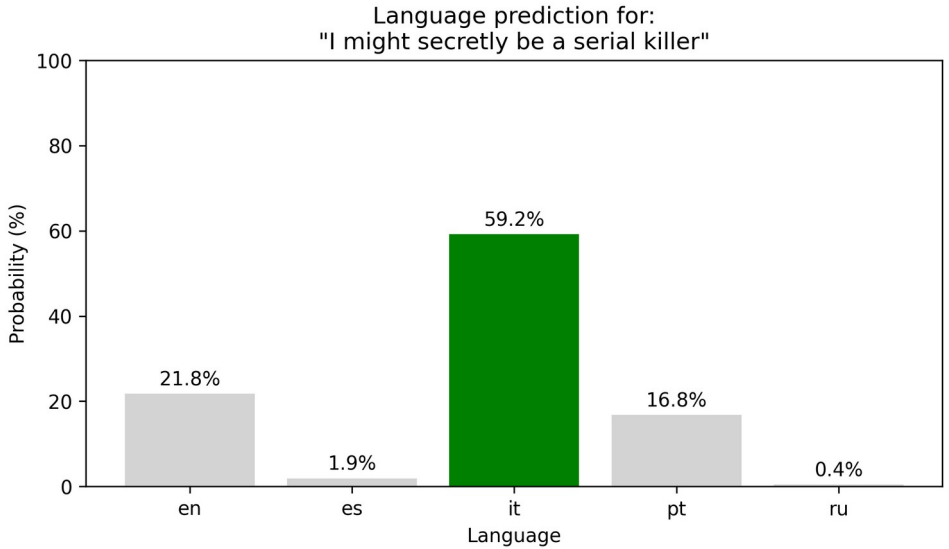

Language prediction for:
"24 es un número muy divertido"
Probability (%)
100
80
60
40
20
0
94.3%
5.0%
0.3%
0.2%
0.2%
en
es
it
pt
ru
Language


Language prediction for:
"Io amo le ragazze italiane"
Probability (%)
100
80
60
40
20
0
78.5%
4.4%
6.0%
5.2%
6.0%
en
es
it
pt
ru
Language


Language prediction for:
"I might secretly be a serial killer"
Probability (%)
100
80
60
40
20
0
21.8%
1.9%
59.2%
16.8%
0.4%
en
es
it
pt
ru
Language

Methodology: Evaluation

- Vocabulary
  - i - 6
  - might - 1223
  - secretly - 0
  - be - 77
  - a - 3
  - serial - 13297
  - killer - 0

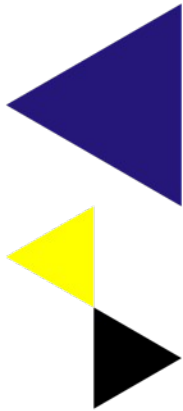


Methodology: Writing Tutorial

- What you will learn
- Introduction
- Approach Overview
- Scope and Limitations
- Coding
  - Step 1: Setting up the environment
  - Step 2: Loading the data
  - Step 3: Filtering languages
  - Step 4: Data Preperation
  - Step 5: Tokenization
  - Step 6: Bag-of-Words model
  - Step 7: Training the Classifier
  - Step 8: Evaluating the model
  -

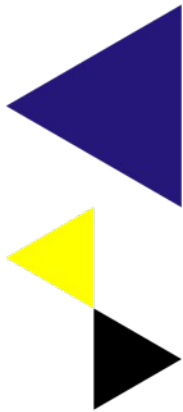
Methodology: Writing Tutorial

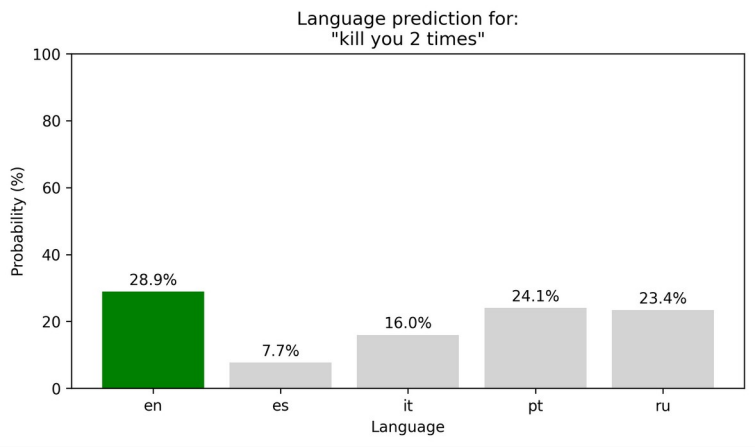
- What you will learn
- Introduction
- Approach Overview
- Scope and Limitations
- Coding
  - Step 1: Setting up the environment
  - Step 2: Loading the data
  - Step 3: Filtering languages
  - Step 4: Data Preperation
  - Step 5: Tokenization
  - Step 6: Bag-of-Words model
  - Step 7: Training the Classifier
  - Step 8: Evaluating the model
  -

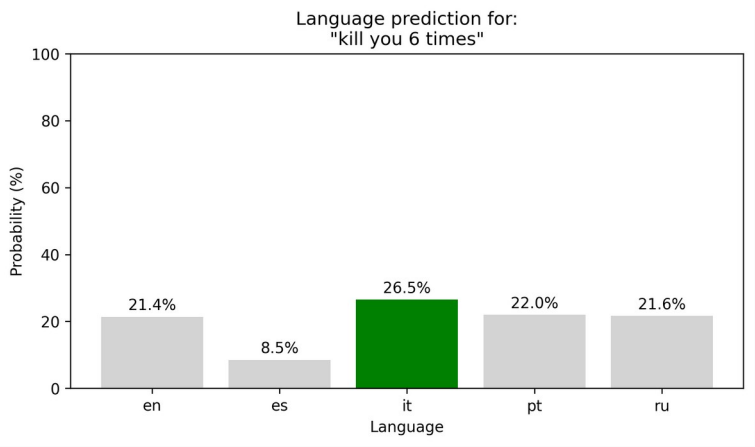


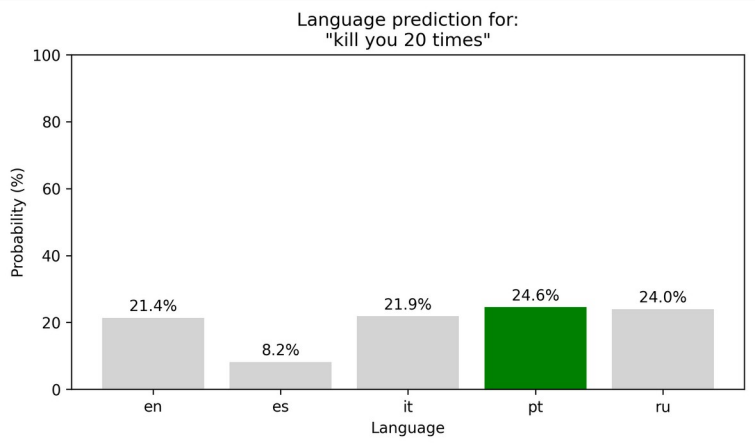
Ideas for improvement

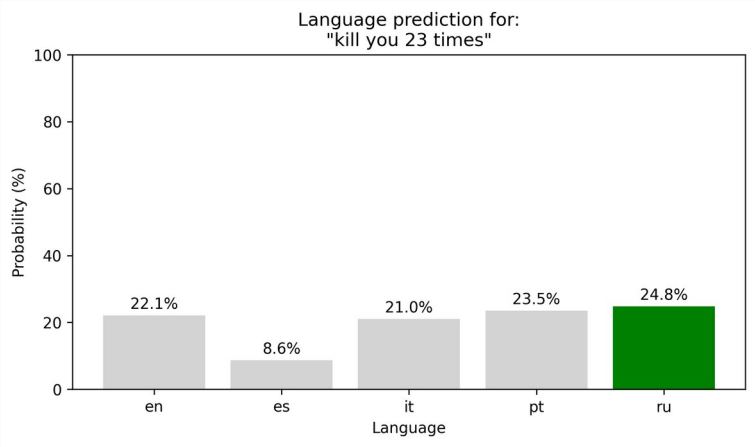
- Remove numbers from data
  -




Language prediction for:
"kill you 2 times"
Probability (%)
100
80
60
40
20
0
28.9%
7.7%
16.0%
24.1%
23.4%
en
es
it
pt
ru
Language

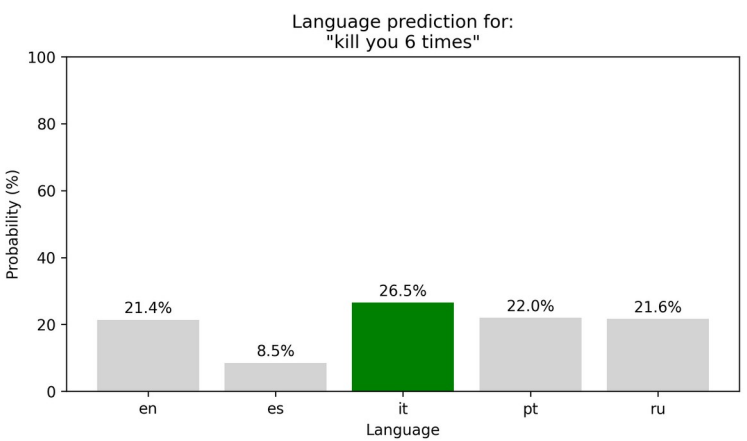

Language prediction for:
"kill you 6 times"
Probability (%)
100
80
60
40
20
0
21.4%
8.5%
26.5%
22.0%
21.6%
en
es
it
pt
ru
Language

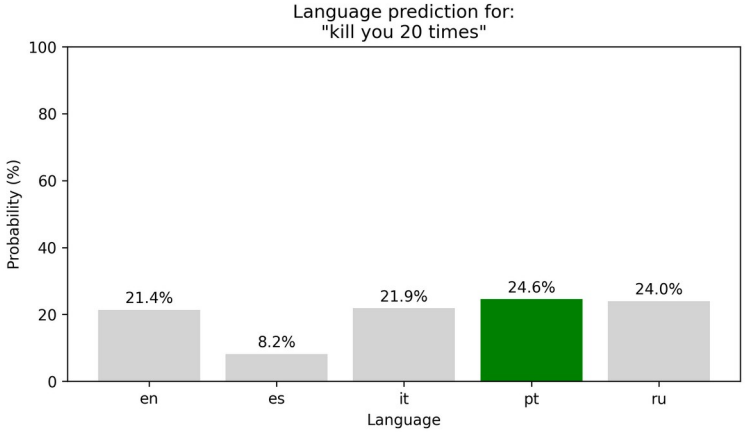

Language prediction for:
"kill you 20 times"
Probability (%)
100
80
60
40
20
0
21.4%
8.2%
21.9%
24.6%
24.0%
en
es
it
pt
ru
Language

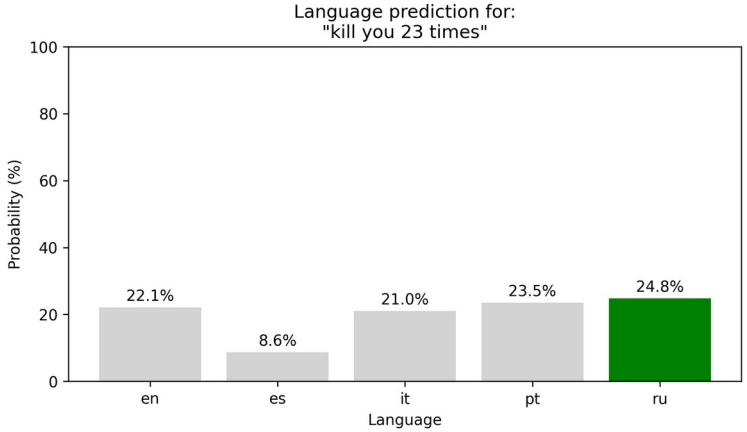

Language prediction for:
"kill you 23 times"
Probability (%)
100
80
60
40
20
0
22.1%
8.6%
21.0%
23.5%
24.8%
en
es
it
pt
ru
Language

Vocabulary

- kill - 10661
- you – 69
- 2 - 100

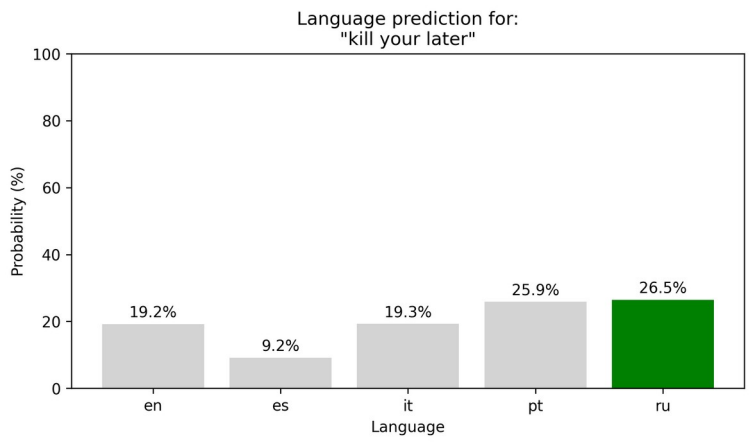


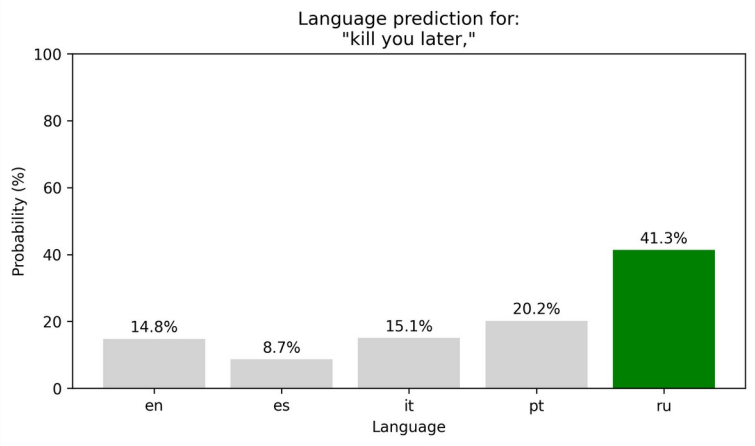


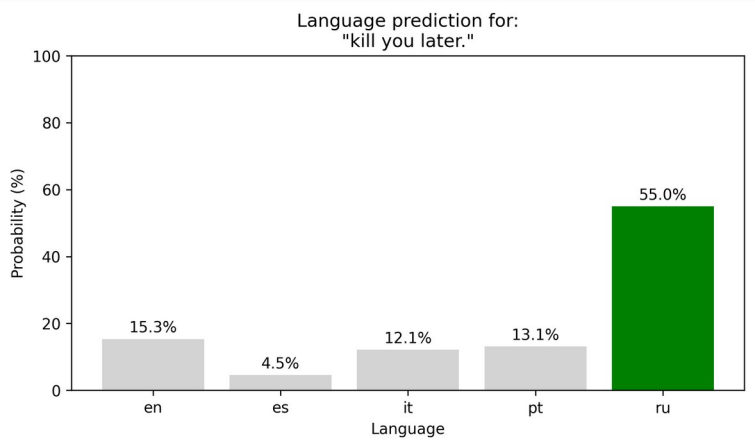


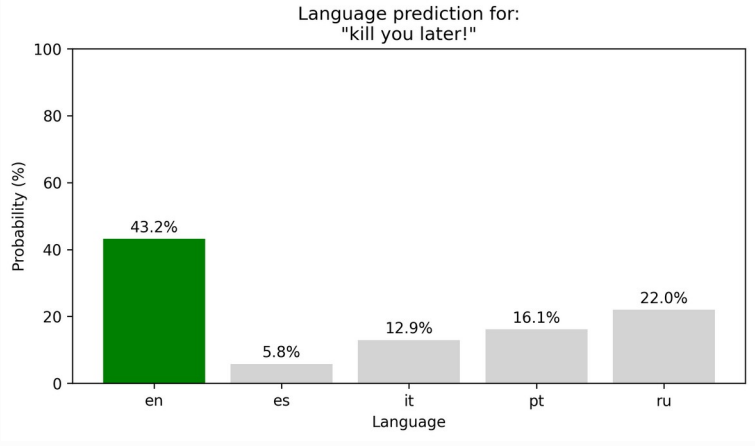
Ideas for improvement

- Remove numbers from data
- Consider removing punctuation
  -


Language prediction for:
"kill your later"
Probability (%)
100
80
60
40
20
0
19.2%
9.2%
19.3%
25.9%
26.5%
en
es
it
pt
ru
Language

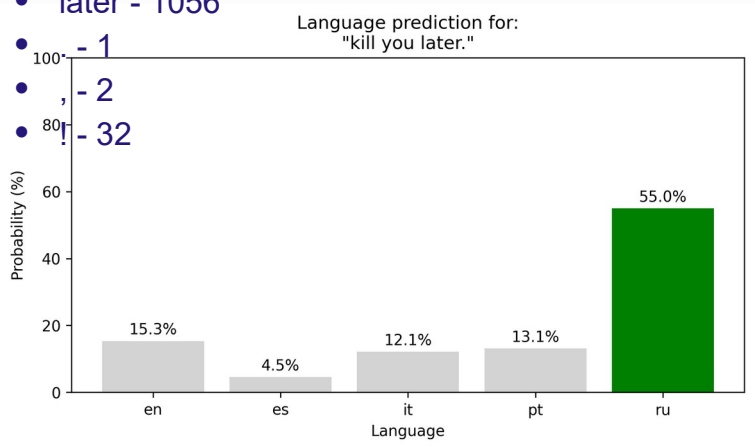

Language prediction for:
"kill you later,"
Probability (%)
100
80
60
40
20
0
14.8%
8.7%
15.1%
20.2%
41.3%
en
es
it
pt
ru
Language

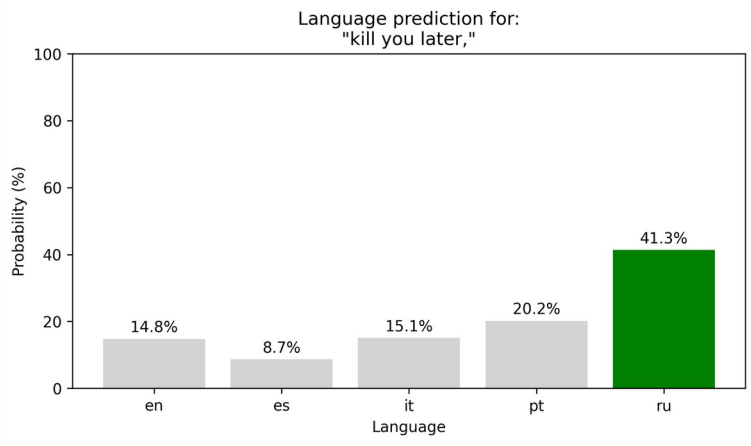

Language prediction for:
"kill you later."
Probability (%)
100
80
60
40
20
0
15.3%
4.5%
12.1%
13.1%
55.0%
en
es
it
pt
ru
Language

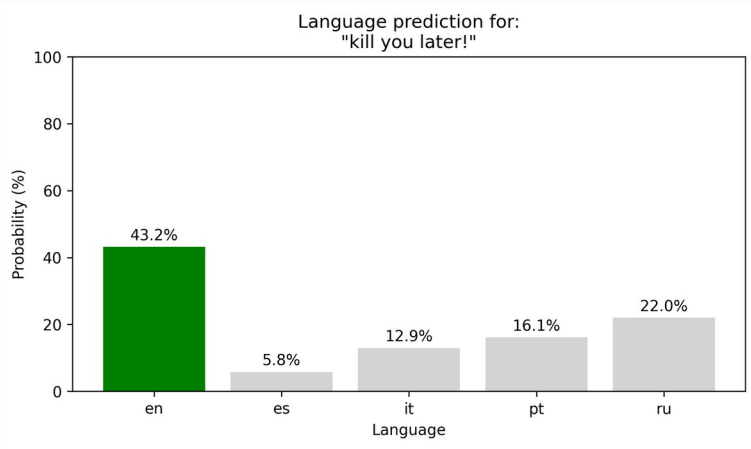

Language prediction for:
"kill you later!"
Probability (%)
100
80
60
40
20
0
43.2%
5.8%
12.9%
16.1%
22.0%
en
es
it
pt
ru
Language

Vocabulary

- kill - 10661
- you – 69
- later - 1056
- . - 1
- , - 2
- ! - 32







Ideas for improvement

- Remove numbers from data
- Consider removing punctuation
- Add more languages
  -

Discussion Points

- Punctuation. Keep? Remove? Partially?
- Every language contains some english words
  - В блант резюме (у меня мало времени) , возможно , просить нас (в массы) **Shell** 20 баксов в год…
  - На веб-сайте агентства по вопросам международного развития , на сайте " **Care** " или…
  - Обсуждаем статью в журнале " **The New York Times** " в журнале " **New York Times** автор…
  - Lo utilice con unos guantes de boxeo **everlast** de 12 onzas y también…
  - Per l'anno in corso, si prevede un fatturato di 94 milioni di dollari e un utile di 26 centesimi per azione da operazioni **continue**.