Antrenare RoBERT + dataset știri română

O imagine care conține text, captură de ecran, Font, număr

Descriere generată automat

**LR, NB, MF, LSVC 12k intrari**

O imagine care conține text, captură de ecran, diagramă, linie

Descriere generată automat

Model: LogisticRegression

Accuracy: 0.6326530612244898

Precision: 0.6280536891415452

Recall: 0.6326530612244898

F1 Score: 0.5803836506022673

Confusion Matrix:

O imagine care conține text, captură de ecran

Descriere generată automat

precision recall f1-score support

BLACK VOICES 0.60 0.14 0.22 44

BUSINESS 0.00 0.00 0.00 15

COMEDY 0.74 0.30 0.42 88

CRIME 0.40 0.05 0.09 38

CULTURE & ARTS 0.00 0.00 0.00 7

EDUCATION 0.00 0.00 0.00 5

ENTERTAINMENT 0.59 0.82 0.69 371

ENVIRONMENT 1.00 0.04 0.07 28

FOOD & DRINK 0.80 0.31 0.44 26

HOME & LIVING 0.91 0.32 0.48 31

IMPACT 0.00 0.00 0.00 13

LATINO VOICES 0.00 0.00 0.00 2

MEDIA 0.60 0.12 0.21 48

MONEY 0.00 0.00 0.00 9

PARENTING 0.89 0.36 0.52 22

POLITICS 0.66 0.92 0.77 800

QUEER VOICES 0.40 0.05 0.08 43

RELIGION 0.00 0.00 0.00 9

SCIENCE 0.00 0.00 0.00 9

SPORTS 0.50 0.19 0.28 42

STYLE & BEAUTY 1.00 0.46 0.63 37

TECH 0.00 0.00 0.00 10

TRAVEL 0.00 0.00 0.00 6

U.S. NEWS 0.49 0.55 0.52 294

WEDDINGS 0.00 0.00 0.00 1

WEIRD NEWS 1.00 0.04 0.08 47

WELLNESS 1.00 0.17 0.29 29

WOMEN 0.88 0.47 0.61 32

WORLD NEWS 0.70 0.69 0.70 246

accuracy 0.63 2352

macro avg 0.45 0.21 0.24 2352

weighted avg 0.63 0.63 0.58 2352

Model: MultinomialNB

Accuracy: 0.5795068027210885

Precision: 0.5068343200358073

Recall: 0.5795068027210885

F1 Score: 0.49363205374330443

Confusion Matrix:

O imagine care conține text, captură de ecran, model

Descriere generată automat

precision recall f1-score support

BLACK VOICES 0.00 0.00 0.00 44

BUSINESS 0.00 0.00 0.00 15

COMEDY 0.64 0.08 0.14 88

CRIME 0.00 0.00 0.00 38

CULTURE & ARTS 0.00 0.00 0.00 7

EDUCATION 0.00 0.00 0.00 5

ENTERTAINMENT 0.56 0.84 0.67 371

ENVIRONMENT 0.00 0.00 0.00 28

FOOD & DRINK 0.00 0.00 0.00 26

HOME & LIVING 0.00 0.00 0.00 31

IMPACT 0.00 0.00 0.00 13

LATINO VOICES 0.00 0.00 0.00 2

MEDIA 0.00 0.00 0.00 48

MONEY 0.00 0.00 0.00 9

PARENTING 1.00 0.23 0.37 22

POLITICS 0.57 0.95 0.71 800

QUEER VOICES 0.00 0.00 0.00 43

RELIGION 0.00 0.00 0.00 9

SCIENCE 0.00 0.00 0.00 9

SPORTS 0.50 0.02 0.05 42

STYLE & BEAUTY 1.00 0.08 0.15 37

TECH 0.00 0.00 0.00 10

TRAVEL 0.00 0.00 0.00 6

U.S. NEWS 0.51 0.42 0.46 294

WEDDINGS 0.00 0.00 0.00 1

WEIRD NEWS 0.00 0.00 0.00 47

WELLNESS 1.00 0.03 0.07 29

WOMEN 1.00 0.38 0.55 32

WORLD NEWS 0.74 0.58 0.65 246

accuracy 0.58 2352

macro avg 0.26 0.12 0.13 2352

weighted avg 0.51 0.58 0.49 2352

Model: RandomForestClassifier

Accuracy: 0.461734693877551

Precision: 0.46178392217566155

Recall: 0.461734693877551

F1 Score: 0.3575492605075011

Confusion Matrix:

O imagine care conține text, captură de ecran, model

Descriere generată automat

precision recall f1-score support

BLACK VOICES 0.00 0.00 0.00 44

BUSINESS 0.00 0.00 0.00 15

COMEDY 0.68 0.19 0.30 88

CRIME 0.00 0.00 0.00 38

CULTURE & ARTS 0.00 0.00 0.00 7

EDUCATION 0.00 0.00 0.00 5

ENTERTAINMENT 0.65 0.55 0.60 371

ENVIRONMENT 0.00 0.00 0.00 28

FOOD & DRINK 0.00 0.00 0.00 26

HOME & LIVING 1.00 0.03 0.06 31

IMPACT 0.00 0.00 0.00 13

LATINO VOICES 0.00 0.00 0.00 2

MEDIA 0.00 0.00 0.00 48

MONEY 0.00 0.00 0.00 9

PARENTING 1.00 0.23 0.37 22

POLITICS 0.41 0.98 0.58 800

QUEER VOICES 0.00 0.00 0.00 43

RELIGION 0.00 0.00 0.00 9

SCIENCE 0.00 0.00 0.00 9

SPORTS 0.00 0.00 0.00 42

STYLE & BEAUTY 0.00 0.00 0.00 37

TECH 0.00 0.00 0.00 10

TRAVEL 0.00 0.00 0.00 6

U.S. NEWS 0.60 0.05 0.09 294

WEDDINGS 0.00 0.00 0.00 1

WEIRD NEWS 0.00 0.00 0.00 47

WELLNESS 0.00 0.00 0.00 29

WOMEN 1.00 0.41 0.58 32

WORLD NEWS 0.79 0.18 0.30 246

accuracy 0.46 2352

macro avg 0.21 0.09 0.10 2352

weighted avg 0.46 0.46 0.36 2352

Model: LinearSVC

Accuracy: 0.6564625850340136

Precision: 0.6408302492666442

Recall: 0.6564625850340136

F1 Score: 0.6370146325878797

Confusion Matrix:

O imagine care conține text, captură de ecran

Descriere generată automat

precision recall f1-score support

BLACK VOICES 0.47 0.32 0.38 44

BUSINESS 0.40 0.13 0.20 15

COMEDY 0.73 0.52 0.61 88

CRIME 0.31 0.21 0.25 38

CULTURE & ARTS 0.00 0.00 0.00 7

EDUCATION 0.00 0.00 0.00 5

ENTERTAINMENT 0.65 0.76 0.70 371

ENVIRONMENT 0.33 0.14 0.20 28

FOOD & DRINK 0.76 0.73 0.75 26

HOME & LIVING 0.69 0.58 0.63 31

IMPACT 0.50 0.08 0.13 13

LATINO VOICES 0.00 0.00 0.00 2

MEDIA 0.53 0.33 0.41 48

MONEY 1.00 0.11 0.20 9

PARENTING 0.75 0.55 0.63 22

POLITICS 0.75 0.87 0.80 800

QUEER VOICES 0.57 0.28 0.37 43

RELIGION 0.00 0.00 0.00 9

SCIENCE 0.50 0.22 0.31 9

SPORTS 0.46 0.43 0.44 42

STYLE & BEAUTY 0.71 0.59 0.65 37

TECH 0.67 0.40 0.50 10

TRAVEL 0.50 0.33 0.40 6

U.S. NEWS 0.47 0.53 0.50 294

WEDDINGS 0.00 0.00 0.00 1

WEIRD NEWS 0.58 0.15 0.24 47

WELLNESS 0.68 0.45 0.54 29

WOMEN 0.68 0.53 0.60 32

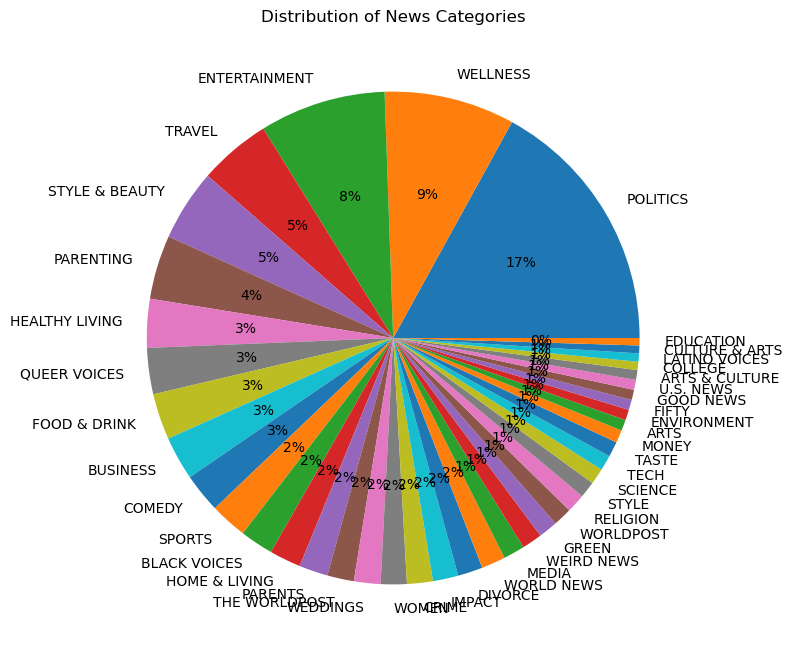
WORLD NEWS 0.66 0.72 0.69 246

accuracy 0.66 2352

macro avg 0.50 0.34 0.38 2352

weighted avg 0.64 0.66 0.64 2352

**200k, linear progression, mlnb, random forest, linear svc**



Model: LogisticRegression

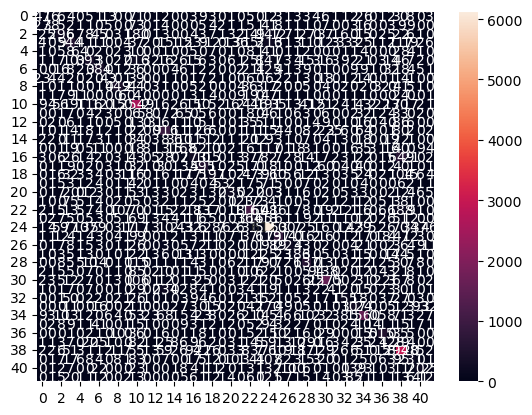
Accuracy: 0.5905598243688255

Precision: 0.5751365905518382

Recall: 0.5905598243688255

F1 Score: 0.5649020437120554

Confusion Matrix:



precision recall f1-score support

ARTS 0.32 0.16 0.21 293

ARTS & CULTURE 0.38 0.14 0.20 275

BLACK VOICES 0.50 0.33 0.40 889

BUSINESS 0.47 0.45 0.46 1216

COLLEGE 0.46 0.32 0.37 202

COMEDY 0.56 0.38 0.46 1022

CRIME 0.55 0.54 0.54 713

CULTURE & ARTS 0.52 0.21 0.30 202

DIVORCE 0.81 0.68 0.74 664

EDUCATION 0.47 0.29 0.36 209

ENTERTAINMENT 0.54 0.75 0.63 3419

ENVIRONMENT 0.68 0.19 0.29 313

FIFTY 0.39 0.12 0.18 263

FOOD & DRINK 0.60 0.72 0.66 1270

GOOD NEWS 0.39 0.13 0.20 270

GREEN 0.39 0.32 0.35 532

HEALTHY LIVING 0.41 0.21 0.28 1302

HOME & LIVING 0.70 0.68 0.69 879

IMPACT 0.41 0.29 0.34 673

LATINO VOICES 0.74 0.22 0.34 238

MEDIA 0.60 0.39 0.47 607

MONEY 0.53 0.35 0.42 355

PARENTING 0.52 0.66 0.58 1768

PARENTS 0.46 0.21 0.29 795

POLITICS 0.67 0.86 0.75 7155

QUEER VOICES 0.74 0.63 0.68 1262

RELIGION 0.62 0.42 0.50 530

SCIENCE 0.57 0.44 0.50 424

SPORTS 0.64 0.63 0.63 1014

STYLE 0.58 0.20 0.30 464

STYLE & BEAUTY 0.72 0.80 0.76 1975

TASTE 0.40 0.10 0.15 427

TECH 0.52 0.39 0.44 398

THE WORLDPOST 0.47 0.41 0.44 741

TRAVEL 0.65 0.77 0.71 2021

U.S. NEWS 0.44 0.05 0.09 269

WEDDINGS 0.75 0.73 0.74 709

WEIRD NEWS 0.38 0.23 0.29 550

WELLNESS 0.55 0.80 0.65 3672

WOMEN 0.45 0.32 0.37 727

WORLD NEWS 0.45 0.30 0.36 665

WORLDPOST 0.45 0.24 0.31 534

accuracy 0.59 41906

macro avg 0.53 0.41 0.44 41906

weighted avg 0.58 0.59 0.56 41906

Model: MultinomialNB

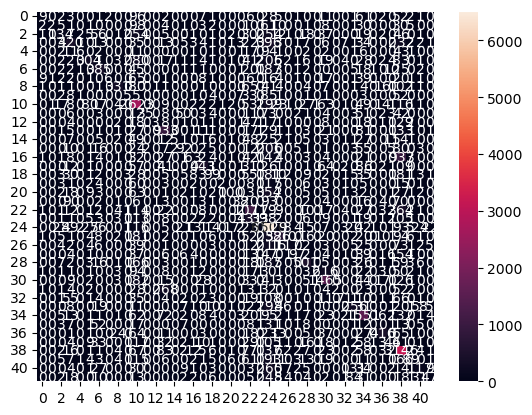
Accuracy: 0.5289218727628502

Precision: 0.563045177551215

Recall: 0.5289218727628502

F1 Score: 0.46833160069368696

Confusion Matrix:



precision recall f1-score support

ARTS 0.33 0.03 0.06 293

ARTS & CULTURE 0.50 0.01 0.01 275

BLACK VOICES 0.54 0.15 0.24 889

BUSINESS 0.48 0.35 0.40 1216

COLLEGE 0.43 0.03 0.06 202

COMEDY 0.64 0.30 0.41 1022

CRIME 0.51 0.54 0.52 713

CULTURE & ARTS 0.75 0.03 0.06 202

DIVORCE 0.84 0.50 0.63 664

EDUCATION 0.50 0.02 0.05 209

ENTERTAINMENT 0.47 0.77 0.58 3419

ENVIRONMENT 0.95 0.11 0.20 313

FIFTY 0.60 0.01 0.02 263

FOOD & DRINK 0.57 0.73 0.64 1270

GOOD NEWS 0.62 0.02 0.04 270

GREEN 0.42 0.17 0.24 532

HEALTHY LIVING 0.50 0.05 0.09 1302

HOME & LIVING 0.78 0.50 0.61 879

IMPACT 0.49 0.15 0.23 673

LATINO VOICES 0.67 0.01 0.02 238

MEDIA 0.75 0.16 0.27 607

MONEY 0.62 0.11 0.18 355

PARENTING 0.43 0.61 0.50 1768

PARENTS 0.62 0.05 0.09 795

POLITICS 0.55 0.91 0.68 7155

QUEER VOICES 0.74 0.46 0.57 1262

RELIGION 0.70 0.22 0.34 530

SCIENCE 0.79 0.23 0.35 424

SPORTS 0.69 0.49 0.57 1014

STYLE 0.75 0.01 0.03 464

STYLE & BEAUTY 0.67 0.74 0.71 1975

TASTE 1.00 0.00 0.01 427

TECH 0.70 0.18 0.28 398

THE WORLDPOST 0.48 0.35 0.40 741

TRAVEL 0.55 0.77 0.64 2021

U.S. NEWS 0.00 0.00 0.00 269

WEDDINGS 0.81 0.59 0.68 709

WEIRD NEWS 0.43 0.08 0.13 550

WELLNESS 0.42 0.86 0.57 3672

WOMEN 0.63 0.12 0.21 727

WORLD NEWS 0.43 0.18 0.25 665

WORLDPOST 0.48 0.09 0.15 534

accuracy 0.53 41906

macro avg 0.59 0.28 0.30 41906

weighted avg 0.56 0.53 0.47 41906

Model: RandomForestClassifier

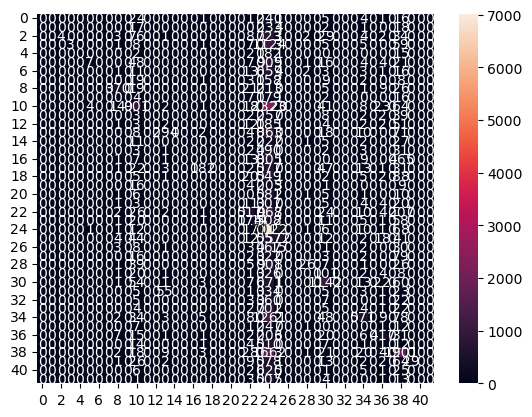
Accuracy: 0.33326969884980673

Precision: 0.4322424521084764

Recall: 0.33326969884980673

F1 Score: 0.2635005877570889

Confusion Matrix:



precision recall f1-score support

ARTS 0.00 0.00 0.00 293

ARTS & CULTURE 0.00 0.00 0.00 275

BLACK VOICES 1.00 0.00 0.01 889

BUSINESS 1.00 0.00 0.00 1216

COLLEGE 0.00 0.00 0.00 202

COMEDY 0.58 0.01 0.01 1022

CRIME 0.00 0.00 0.00 713

CULTURE & ARTS 0.00 0.00 0.00 202

DIVORCE 0.88 0.56 0.68 664

EDUCATION 0.00 0.00 0.00 209

ENTERTAINMENT 0.57 0.26 0.36 3419

ENVIRONMENT 0.00 0.00 0.00 313

FIFTY 0.00 0.00 0.00 263

FOOD & DRINK 0.77 0.23 0.36 1270

GOOD NEWS 0.00 0.00 0.00 270

GREEN 0.00 0.00 0.00 532

HEALTHY LIVING 0.00 0.00 0.00 1302

HOME & LIVING 0.90 0.21 0.34 879

IMPACT 0.00 0.00 0.00 673

LATINO VOICES 0.00 0.00 0.00 238

MEDIA 0.00 0.00 0.00 607

MONEY 0.00 0.00 0.00 355

PARENTING 0.54 0.29 0.38 1768

PARENTS 1.00 0.02 0.03 795

POLITICS 0.23 0.98 0.37 7155

QUEER VOICES 0.81 0.45 0.58 1262

RELIGION 0.00 0.00 0.00 530

SCIENCE 0.00 0.00 0.00 424

SPORTS 0.72 0.03 0.05 1014

STYLE 0.00 0.00 0.00 464

STYLE & BEAUTY 0.69 0.58 0.63 1975

TASTE 0.00 0.00 0.00 427

TECH 0.00 0.00 0.00 398

THE WORLDPOST 0.00 0.00 0.00 741

TRAVEL 0.77 0.28 0.41 2021

U.S. NEWS 0.00 0.00 0.00 269

WEDDINGS 0.77 0.59 0.67 709

WEIRD NEWS 0.00 0.00 0.00 550

WELLNESS 0.49 0.52 0.50 3672

WOMEN 1.00 0.04 0.08 727

WORLD NEWS 0.00 0.00 0.00 665

WORLDPOST 0.00 0.00 0.00 534

accuracy 0.33 41906

macro avg 0.30 0.12 0.13 41906

weighted avg 0.43 0.33 0.26 41906

Model: LinearSVC

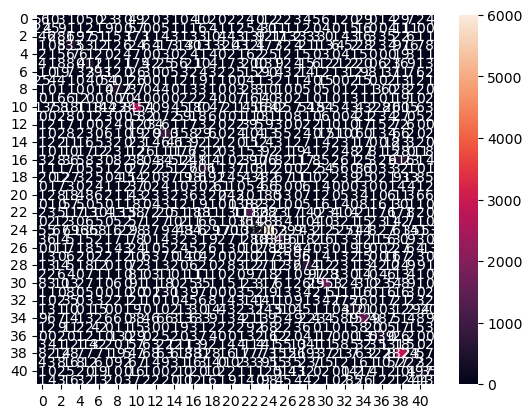
Accuracy: 0.5864792631126807

Precision: 0.5622647311232026

Recall: 0.5864792631126807

F1 Score: 0.565372431611593

Confusion Matrix:



precision recall f1-score support

ARTS 0.29 0.19 0.23 293

ARTS & CULTURE 0.33 0.16 0.22 275

BLACK VOICES 0.44 0.32 0.37 889

BUSINESS 0.49 0.44 0.46 1216

COLLEGE 0.41 0.38 0.39 202

COMEDY 0.56 0.40 0.47 1022

CRIME 0.52 0.55 0.53 713

CULTURE & ARTS 0.36 0.27 0.31 202

DIVORCE 0.75 0.72 0.73 664

EDUCATION 0.41 0.33 0.37 209

ENTERTAINMENT 0.59 0.72 0.65 3419

ENVIRONMENT 0.44 0.26 0.33 313

FIFTY 0.30 0.13 0.18 263

FOOD & DRINK 0.58 0.72 0.64 1270

GOOD NEWS 0.29 0.17 0.21 270

GREEN 0.36 0.33 0.34 532

HEALTHY LIVING 0.37 0.19 0.25 1302

HOME & LIVING 0.65 0.69 0.67 879

IMPACT 0.41 0.28 0.34 673

LATINO VOICES 0.50 0.26 0.34 238

MEDIA 0.53 0.40 0.46 607

MONEY 0.47 0.39 0.42 355

PARENTING 0.52 0.63 0.57 1768

PARENTS 0.39 0.19 0.25 795

POLITICS 0.70 0.84 0.77 7155

QUEER VOICES 0.72 0.67 0.70 1262

RELIGION 0.56 0.45 0.50 530

SCIENCE 0.51 0.46 0.48 424

SPORTS 0.59 0.66 0.63 1014

STYLE 0.46 0.21 0.29 464

STYLE & BEAUTY 0.72 0.79 0.75 1975

TASTE 0.31 0.12 0.17 427

TECH 0.45 0.44 0.45 398

THE WORLDPOST 0.44 0.42 0.43 741

TRAVEL 0.65 0.76 0.70 2021

U.S. NEWS 0.24 0.07 0.11 269

WEDDINGS 0.72 0.76 0.74 709

WEIRD NEWS 0.33 0.24 0.28 550

WELLNESS 0.57 0.77 0.66 3672

WOMEN 0.45 0.31 0.36 727

WORLD NEWS 0.40 0.30 0.34 665

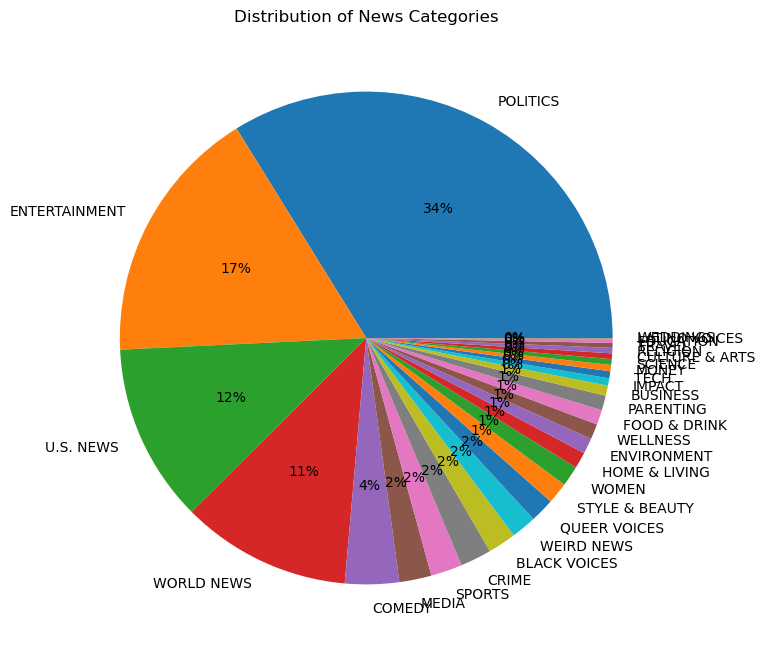
WORLDPOST 0.37 0.25 0.30 534

accuracy 0.59 41906

macro avg 0.48 0.42 0.44 41906

weighted avg 0.56 0.59 0.57 41906

**Undersampling – cu categorii sub 3% unite. – 12k intrări**



Class distribution before undersampling:

category

POLITICS 3978

ENTERTAINMENT 1991

U.S. NEWS 1377

WORLD NEWS 1310

COMEDY 417

MEDIA 248

SPORTS 242

CRIME 241

BLACK VOICES 218

WEIRD NEWS 194

QUEER VOICES 186

STYLE & BEAUTY 165

WOMEN 158

HOME & LIVING 125

ENVIRONMENT 121

WELLNESS 118

FOOD & DRINK 114

PARENTING 114

BUSINESS 78

IMPACT 57

TECH 54

MONEY 49

SCIENCE 44

CULTURE & ARTS 44

RELIGION 42

TRAVEL 37

EDUCATION 27

LATINO VOICES 9

WEDDINGS 2

Name: count, dtype: int64

New Class Distribution:

category\_merged

POLITICS 3978

General 2687

ENTERTAINMENT 1991

U.S. NEWS 1377

WORLD NEWS 1310

COMEDY 417

Name: count, dtype: int64

Class distribution after undersampling:

category\_merged

U.S. NEWS 417

COMEDY 417

General 417

WORLD NEWS 417

ENTERTAINMENT 417

POLITICS 417

Name: count, dtype: int64

Evaluating models on undersampled training data:

Model: LogisticRegression

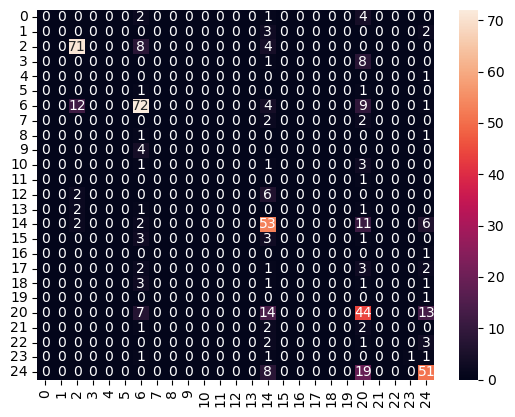
Accuracy: 0.5828343313373253

Precision: 0.4987578752170344

Recall: 0.5828343313373253

F1 Score: 0.533059394931315

Confusion Matrix:



precision recall f1-score support

BLACK VOICES 0.00 0.00 0.00 7

BUSINESS 0.00 0.00 0.00 5

COMEDY 0.80 0.86 0.83 83

CRIME 0.00 0.00 0.00 9

CULTURE & ARTS 0.00 0.00 0.00 1

EDUCATION 0.00 0.00 0.00 2

ENTERTAINMENT 0.66 0.73 0.70 98

ENVIRONMENT 0.00 0.00 0.00 4

FOOD & DRINK 0.00 0.00 0.00 2

HOME & LIVING 0.00 0.00 0.00 4

IMPACT 0.00 0.00 0.00 5

LATINO VOICES 0.00 0.00 0.00 1

MEDIA 0.00 0.00 0.00 8

PARENTING 0.00 0.00 0.00 4

POLITICS 0.50 0.72 0.59 74

QUEER VOICES 0.00 0.00 0.00 7

RELIGION 0.00 0.00 0.00 1

SPORTS 0.00 0.00 0.00 8

STYLE & BEAUTY 0.00 0.00 0.00 6

TRAVEL 0.00 0.00 0.00 1

U.S. NEWS 0.40 0.56 0.47 78

WEIRD NEWS 0.00 0.00 0.00 5

WELLNESS 0.00 0.00 0.00 6

WOMEN 1.00 0.25 0.40 4

WORLD NEWS 0.61 0.65 0.63 78

accuracy 0.58 501

macro avg 0.16 0.15 0.14 501

weighted avg 0.50 0.58 0.53 501

Model: MultinomialNB

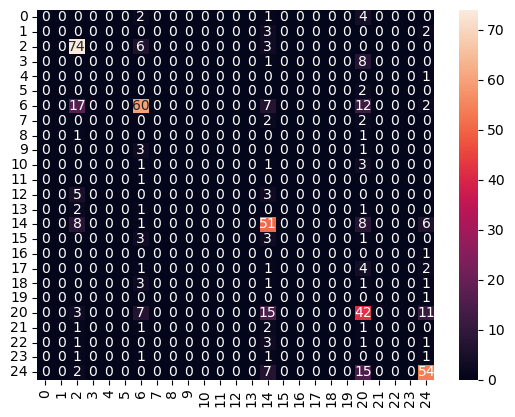
Accuracy: 0.5608782435129741

Precision: 0.4697220514812963

Recall: 0.5608782435129741

F1 Score: 0.5073241948316671

Confusion Matrix:



precision recall f1-score support

BLACK VOICES 0.00 0.00 0.00 7

BUSINESS 0.00 0.00 0.00 5

COMEDY 0.64 0.89 0.75 83

CRIME 0.00 0.00 0.00 9

CULTURE & ARTS 0.00 0.00 0.00 1

EDUCATION 0.00 0.00 0.00 2

ENTERTAINMENT 0.66 0.61 0.63 98

ENVIRONMENT 0.00 0.00 0.00 4

FOOD & DRINK 0.00 0.00 0.00 2

HOME & LIVING 0.00 0.00 0.00 4

IMPACT 0.00 0.00 0.00 5

LATINO VOICES 0.00 0.00 0.00 1

MEDIA 0.00 0.00 0.00 8

PARENTING 0.00 0.00 0.00 4

POLITICS 0.49 0.69 0.57 74

QUEER VOICES 0.00 0.00 0.00 7

RELIGION 0.00 0.00 0.00 1

SPORTS 0.00 0.00 0.00 8

STYLE & BEAUTY 0.00 0.00 0.00 6

TRAVEL 0.00 0.00 0.00 1

U.S. NEWS 0.39 0.54 0.45 78

WEIRD NEWS 0.00 0.00 0.00 5

WELLNESS 0.00 0.00 0.00 6

WOMEN 0.00 0.00 0.00 4

WORLD NEWS 0.65 0.69 0.67 78

accuracy 0.56 501

macro avg 0.11 0.14 0.12 501

weighted avg 0.47 0.56 0.51 501

Model: RandomForestClassifier

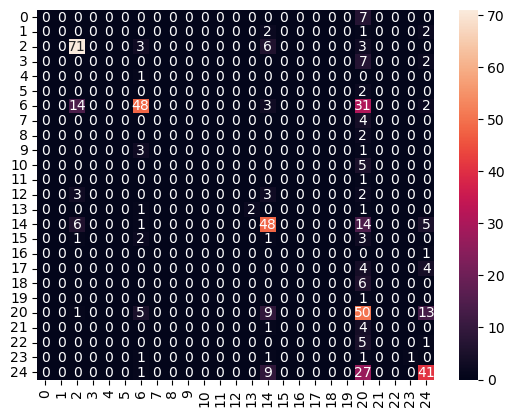
Accuracy: 0.5209580838323353

Precision: 0.4988505980279116

Recall: 0.5209580838323353

F1 Score: 0.4903207143396288

Confusion Matrix:



precision recall f1-score support

BLACK VOICES 0.00 0.00 0.00 7

BUSINESS 0.00 0.00 0.00 5

COMEDY 0.74 0.86 0.79 83

CRIME 0.00 0.00 0.00 9

CULTURE & ARTS 0.00 0.00 0.00 1

EDUCATION 0.00 0.00 0.00 2

ENTERTAINMENT 0.73 0.49 0.59 98

ENVIRONMENT 0.00 0.00 0.00 4

FOOD & DRINK 0.00 0.00 0.00 2

HOME & LIVING 0.00 0.00 0.00 4

IMPACT 0.00 0.00 0.00 5

LATINO VOICES 0.00 0.00 0.00 1

MEDIA 0.00 0.00 0.00 8

PARENTING 1.00 0.50 0.67 4

POLITICS 0.58 0.65 0.61 74

QUEER VOICES 0.00 0.00 0.00 7

RELIGION 0.00 0.00 0.00 1

SPORTS 0.00 0.00 0.00 8

STYLE & BEAUTY 0.00 0.00 0.00 6

TRAVEL 0.00 0.00 0.00 1

U.S. NEWS 0.27 0.64 0.38 78

WEIRD NEWS 0.00 0.00 0.00 5

WELLNESS 0.00 0.00 0.00 6

WOMEN 1.00 0.25 0.40 4

WORLD NEWS 0.58 0.53 0.55 78

accuracy 0.52 501

macro avg 0.20 0.16 0.16 501

weighted avg 0.50 0.52 0.49 501

Model: LinearSVC

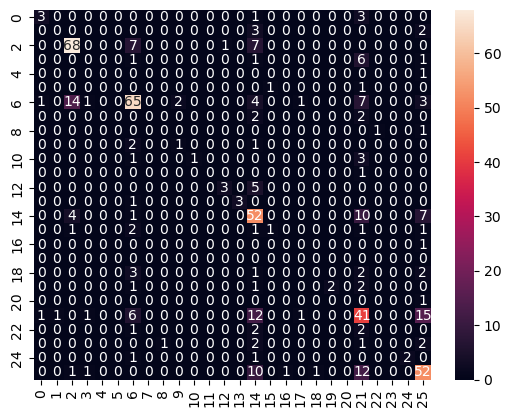
Accuracy: 0.5868263473053892

Precision: 0.5654790908194456

Recall: 0.5868263473053892

F1 Score: 0.5639542860524595

Confusion Matrix:



precision recall f1-score support

BLACK VOICES 0.60 0.43 0.50 7

BUSINESS 0.00 0.00 0.00 5

COMEDY 0.77 0.82 0.80 83

CRIME 0.00 0.00 0.00 9

CULTURE & ARTS 0.00 0.00 0.00 1

EDUCATION 0.00 0.00 0.00 2

ENTERTAINMENT 0.71 0.66 0.68 98

ENVIRONMENT 0.00 0.00 0.00 4

FOOD & DRINK 0.00 0.00 0.00 2

HOME & LIVING 0.33 0.25 0.29 4

IMPACT 1.00 0.20 0.33 5

LATINO VOICES 0.00 0.00 0.00 1

MEDIA 0.75 0.38 0.50 8

PARENTING 1.00 0.75 0.86 4

POLITICS 0.49 0.70 0.58 74

QUEER VOICES 0.50 0.14 0.22 7

RELIGION 0.00 0.00 0.00 1

SCIENCE 0.00 0.00 0.00 0

SPORTS 0.00 0.00 0.00 8

STYLE & BEAUTY 1.00 0.33 0.50 6

TRAVEL 0.00 0.00 0.00 1

U.S. NEWS 0.44 0.53 0.48 78

WEIRD NEWS 0.00 0.00 0.00 5

WELLNESS 0.00 0.00 0.00 6

WOMEN 1.00 0.50 0.67 4

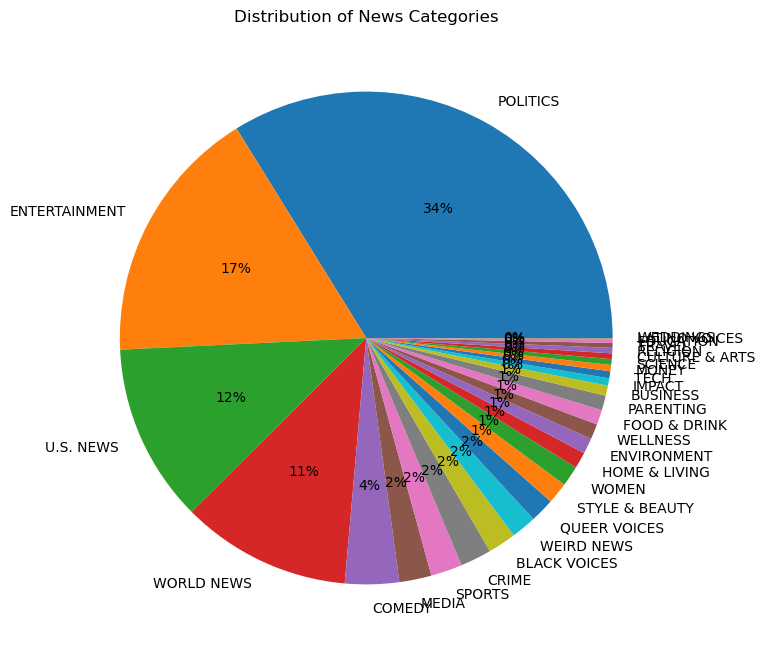
WORLD NEWS 0.58 0.67 0.62 78

accuracy 0.59 501

macro avg 0.35 0.24 0.27 501

weighted avg 0.57 0.59 0.56 501

**Undersampling cu categorii sub 5% unite – 12k intrări**



Class distribution before undersampling:

category

POLITICS 3978

ENTERTAINMENT 1991

U.S. NEWS 1377

WORLD NEWS 1310

COMEDY 417

MEDIA 248

SPORTS 242

CRIME 241

BLACK VOICES 218

WEIRD NEWS 194

QUEER VOICES 186

STYLE & BEAUTY 165

WOMEN 158

HOME & LIVING 125

ENVIRONMENT 121

WELLNESS 118

FOOD & DRINK 114

PARENTING 114

BUSINESS 78

IMPACT 57

TECH 54

MONEY 49

SCIENCE 44

CULTURE & ARTS 44

RELIGION 42

TRAVEL 37

EDUCATION 27

LATINO VOICES 9

WEDDINGS 2

Name: count, dtype: int64

New Class Distribution:

category\_merged

POLITICS 3978

General 3104

ENTERTAINMENT 1991

U.S. NEWS 1377

WORLD NEWS 1310

Name: count, dtype: int64

Class distribution after undersampling:

category\_merged

U.S. NEWS 1310

General 1310

WORLD NEWS 1310

ENTERTAINMENT 1310

POLITICS 1310

Name: count, dtype: int64

Evaluating models on undersampled training data:

Model: LogisticRegression

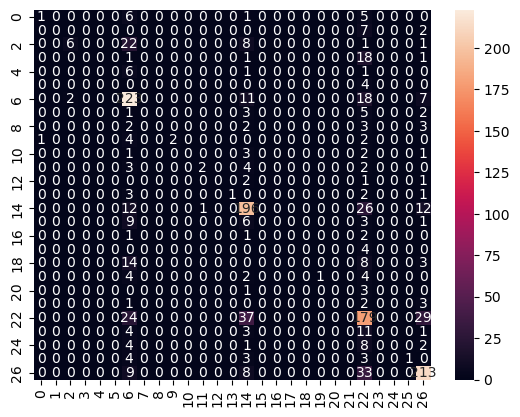
Accuracy: 0.6297709923664122

Precision: 0.5651923354635222

Recall: 0.6297709923664122

F1 Score: 0.5721684669191532

Confusion Matrix:



precision recall f1-score support

BLACK VOICES 0.50 0.08 0.13 13

BUSINESS 0.00 0.00 0.00 9

COMEDY 0.75 0.16 0.26 38

CRIME 0.00 0.00 0.00 21

CULTURE & ARTS 0.00 0.00 0.00 8

EDUCATION 0.00 0.00 0.00 4

ENTERTAINMENT 0.62 0.85 0.72 261

ENVIRONMENT 0.00 0.00 0.00 11

FOOD & DRINK 0.00 0.00 0.00 10

HOME & LIVING 1.00 0.22 0.36 9

IMPACT 0.00 0.00 0.00 7

MEDIA 0.67 0.18 0.29 11

MONEY 0.00 0.00 0.00 4

PARENTING 1.00 0.14 0.25 7

POLITICS 0.67 0.79 0.72 247

QUEER VOICES 0.00 0.00 0.00 19

RELIGION 0.00 0.00 0.00 4

SCIENCE 0.00 0.00 0.00 4

SPORTS 0.00 0.00 0.00 25

STYLE & BEAUTY 1.00 0.09 0.17 11

TECH 0.00 0.00 0.00 4

TRAVEL 0.00 0.00 0.00 6

U.S. NEWS 0.50 0.67 0.57 269

WEIRD NEWS 0.00 0.00 0.00 19

WELLNESS 0.00 0.00 0.00 15

WOMEN 1.00 0.09 0.17 11

WORLD NEWS 0.75 0.81 0.78 263

accuracy 0.63 1310

macro avg 0.31 0.15 0.16 1310

weighted avg 0.57 0.63 0.57 1310

Model: MultinomialNB

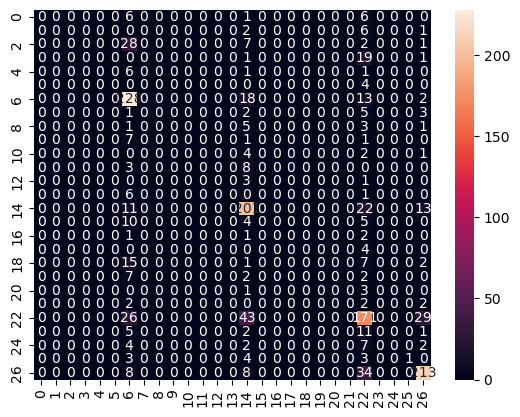
Accuracy: 0.6213740458015267

Precision: 0.5076790504473384

Recall: 0.6213740458015267

F1 Score: 0.552535655893492

Confusion Matrix:



precision recall f1-score support

BLACK VOICES 0.00 0.00 0.00 13

BUSINESS 0.00 0.00 0.00 9

COMEDY 0.00 0.00 0.00 38

CRIME 0.00 0.00 0.00 21

CULTURE & ARTS 0.00 0.00 0.00 8

EDUCATION 0.00 0.00 0.00 4

ENTERTAINMENT 0.60 0.87 0.71 261

ENVIRONMENT 0.00 0.00 0.00 11

FOOD & DRINK 0.00 0.00 0.00 10

HOME & LIVING 0.00 0.00 0.00 9

IMPACT 0.00 0.00 0.00 7

MEDIA 0.00 0.00 0.00 11

MONEY 0.00 0.00 0.00 4

PARENTING 0.00 0.00 0.00 7

POLITICS 0.62 0.81 0.71 247

QUEER VOICES 0.00 0.00 0.00 19

RELIGION 0.00 0.00 0.00 4

SCIENCE 0.00 0.00 0.00 4

SPORTS 0.00 0.00 0.00 25

STYLE & BEAUTY 0.00 0.00 0.00 11

TECH 0.00 0.00 0.00 4

TRAVEL 0.00 0.00 0.00 6

U.S. NEWS 0.51 0.64 0.56 269

WEIRD NEWS 0.00 0.00 0.00 19

WELLNESS 0.00 0.00 0.00 15

WOMEN 1.00 0.09 0.17 11

WORLD NEWS 0.78 0.81 0.80 263

accuracy 0.62 1310

macro avg 0.13 0.12 0.11 1310

weighted avg 0.51 0.62 0.55 1310

Model: RandomForestClassifier

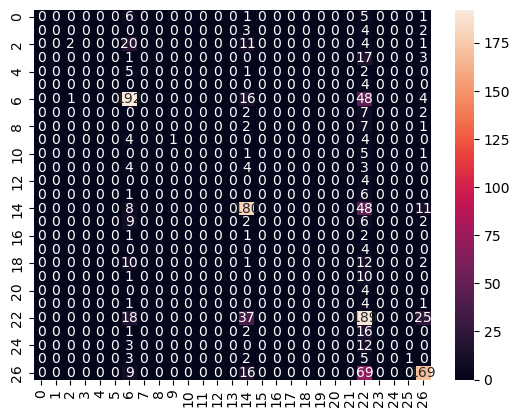
Accuracy: 0.5603053435114503

Precision: 0.512016369974871

Recall: 0.5603053435114503

F1 Score: 0.5108844827423705

Confusion Matrix:



precision recall f1-score support

BLACK VOICES 0.00 0.00 0.00 13

BUSINESS 0.00 0.00 0.00 9

COMEDY 0.67 0.05 0.10 38

CRIME 0.00 0.00 0.00 21

CULTURE & ARTS 0.00 0.00 0.00 8

EDUCATION 0.00 0.00 0.00 4

ENTERTAINMENT 0.65 0.74 0.69 261

ENVIRONMENT 0.00 0.00 0.00 11

FOOD & DRINK 0.00 0.00 0.00 10

HOME & LIVING 1.00 0.11 0.20 9

IMPACT 0.00 0.00 0.00 7

MEDIA 0.00 0.00 0.00 11

MONEY 0.00 0.00 0.00 4

PARENTING 0.00 0.00 0.00 7

POLITICS 0.64 0.73 0.68 247

QUEER VOICES 0.00 0.00 0.00 19

RELIGION 0.00 0.00 0.00 4

SCIENCE 0.00 0.00 0.00 4

SPORTS 0.00 0.00 0.00 25

STYLE & BEAUTY 0.00 0.00 0.00 11

TECH 0.00 0.00 0.00 4

TRAVEL 0.00 0.00 0.00 6

U.S. NEWS 0.38 0.70 0.49 269

WEIRD NEWS 0.00 0.00 0.00 19

WELLNESS 0.00 0.00 0.00 15

WOMEN 1.00 0.09 0.17 11

WORLD NEWS 0.75 0.64 0.69 263

accuracy 0.56 1310

macro avg 0.19 0.11 0.11 1310

weighted avg 0.51 0.56 0.51 1310

Model: LinearSVC

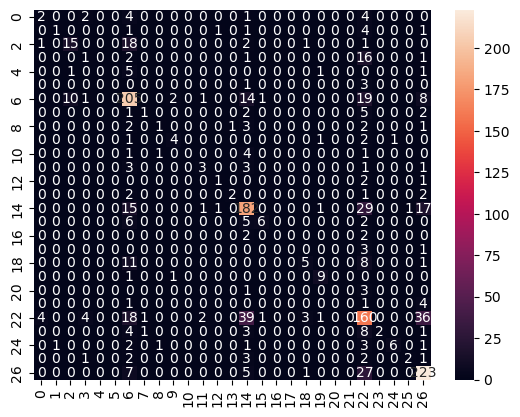
Accuracy: 0.6343511450381679

Precision: 0.6090762083292589

Recall: 0.6343511450381679

F1 Score: 0.605550264972537

Confusion Matrix:



precision recall f1-score support

BLACK VOICES 0.29 0.15 0.20 13

BUSINESS 0.50 0.11 0.18 9

COMEDY 0.58 0.39 0.47 38

CRIME 0.11 0.05 0.07 21

CULTURE & ARTS 0.00 0.00 0.00 8

EDUCATION 0.00 0.00 0.00 4

ENTERTAINMENT 0.66 0.79 0.72 261

ENVIRONMENT 0.33 0.09 0.14 11

FOOD & DRINK 0.33 0.10 0.15 10

HOME & LIVING 0.57 0.44 0.50 9

IMPACT 0.00 0.00 0.00 7

MEDIA 0.43 0.27 0.33 11

MONEY 0.33 0.25 0.29 4

PARENTING 0.67 0.29 0.40 7

POLITICS 0.67 0.74 0.70 247

QUEER VOICES 0.75 0.32 0.44 19

RELIGION 0.00 0.00 0.00 4

SCIENCE 0.00 0.00 0.00 4

SPORTS 0.50 0.20 0.29 25

STYLE & BEAUTY 0.69 0.82 0.75 11

TECH 0.00 0.00 0.00 4

TRAVEL 0.00 0.00 0.00 6

U.S. NEWS 0.52 0.59 0.55 269

WEIRD NEWS 1.00 0.11 0.19 19

WELLNESS 0.86 0.40 0.55 15

WOMEN 0.67 0.18 0.29 11

WORLD NEWS 0.74 0.85 0.79 263

accuracy 0.63 1310

macro avg 0.41 0.26 0.30 1310

weighted avg 0.61 0.63 0.61 1310

**Undersampling, merge la categorii sub 5%, 200k intrări**

O imagine care conține text, captură de ecran, Color, cerc

Descriere generată automat

Class distribution before undersampling:

category

POLITICS 35602

WELLNESS 17945

ENTERTAINMENT 17362

TRAVEL 9900

STYLE & BEAUTY 9814

PARENTING 8791

HEALTHY LIVING 6694

QUEER VOICES 6347

FOOD & DRINK 6340

BUSINESS 5992

COMEDY 5400

SPORTS 5077

BLACK VOICES 4583

HOME & LIVING 4320

PARENTS 3955

THE WORLDPOST 3664

WEDDINGS 3653

WOMEN 3572

CRIME 3562

IMPACT 3484

DIVORCE 3426

WORLD NEWS 3299

MEDIA 2944

WEIRD NEWS 2777

GREEN 2622

WORLDPOST 2579

RELIGION 2577

STYLE 2254

SCIENCE 2206

TECH 2104

TASTE 2096

MONEY 1756

ARTS 1509

ENVIRONMENT 1444

FIFTY 1401

GOOD NEWS 1398

U.S. NEWS 1377

ARTS & CULTURE 1339

COLLEGE 1144

LATINO VOICES 1130

CULTURE & ARTS 1074

EDUCATION 1014

Name: count, dtype: int64

New Class Distribution:

category\_merged

General 138618

POLITICS 35602

WELLNESS 17945

ENTERTAINMENT 17362

Name: count, dtype: int64

Class distribution after undersampling:

category\_merged

General 17362

ENTERTAINMENT 17362

POLITICS 17362

WELLNESS 17362

Name: count, dtype: int64

Model: LogisticRegression

Accuracy: 0.7436285097192225

Precision: 0.683336588547986

Recall: 0.7436285097192225

F1 Score: 0.6867231042287619

Confusion Matrix:

O imagine care conține text, captură de ecran, model, petic

Descriere generată automat

precision recall f1-score support

ARTS 0.33 0.03 0.05 36

ARTS & CULTURE 0.00 0.00 0.00 29

BLACK VOICES 0.47 0.08 0.13 89

BUSINESS 0.50 0.11 0.18 149

COLLEGE 0.44 0.15 0.23 26

COMEDY 0.75 0.05 0.09 127

CRIME 0.62 0.20 0.30 81

CULTURE & ARTS 0.00 0.00 0.00 23

DIVORCE 0.90 0.36 0.51 72

EDUCATION 0.00 0.00 0.00 22

ENTERTAINMENT 0.74 0.91 0.81 3463

ENVIRONMENT 1.00 0.03 0.06 34

FIFTY 1.00 0.05 0.10 37

FOOD & DRINK 0.66 0.41 0.51 156

GOOD NEWS 0.00 0.00 0.00 28

GREEN 0.75 0.04 0.08 67

HEALTHY LIVING 0.00 0.00 0.00 168

HOME & LIVING 0.78 0.36 0.49 100

IMPACT 0.17 0.01 0.02 80

LATINO VOICES 0.00 0.00 0.00 35

MEDIA 0.00 0.00 0.00 84

MONEY 0.00 0.00 0.00 39

PARENTING 0.43 0.31 0.36 235

PARENTS 0.20 0.01 0.02 102

POLITICS 0.78 0.92 0.85 3554

QUEER VOICES 0.83 0.37 0.51 194

RELIGION 0.53 0.12 0.20 66

SCIENCE 0.40 0.04 0.08 45

SPORTS 0.44 0.16 0.24 137

STYLE 0.00 0.00 0.00 64

STYLE & BEAUTY 0.72 0.59 0.65 234

TASTE 0.00 0.00 0.00 48

TECH 0.91 0.16 0.28 61

THE WORLDPOST 0.36 0.04 0.07 100

TRAVEL 0.72 0.46 0.56 262

U.S. NEWS 0.00 0.00 0.00 29

WEDDINGS 0.74 0.45 0.56 78

WEIRD NEWS 0.00 0.00 0.00 73

WELLNESS 0.74 0.95 0.83 3429

WOMEN 0.45 0.06 0.11 81

WORLD NEWS 0.25 0.01 0.02 82

WORLDPOST 0.00 0.00 0.00 71

accuracy 0.74 13890

macro avg 0.42 0.18 0.21 13890

weighted avg 0.68 0.74 0.69 13890

Model: MultinomialNB

Accuracy: 0.7061915046796257

Precision: 0.6189774339119444

Recall: 0.7061915046796257

F1 Score: 0.6182057111736654

Confusion Matrix:

O imagine care conține text, captură de ecran, model, Dreptunghi

Descriere generată automat

precision recall f1-score support

ARTS 0.00 0.00 0.00 36

ARTS & CULTURE 0.00 0.00 0.00 29

BLACK VOICES 0.00 0.00 0.00 89

BUSINESS 0.00 0.00 0.00 149

COLLEGE 0.00 0.00 0.00 26

COMEDY 0.00 0.00 0.00 127

CRIME 1.00 0.01 0.02 81

CULTURE & ARTS 0.00 0.00 0.00 23

DIVORCE 0.00 0.00 0.00 72

EDUCATION 0.00 0.00 0.00 22

ENTERTAINMENT 0.70 0.90 0.79 3463

ENVIRONMENT 0.00 0.00 0.00 34

FIFTY 0.00 0.00 0.00 37

FOOD & DRINK 0.67 0.13 0.22 156

GOOD NEWS 0.00 0.00 0.00 28

GREEN 0.00 0.00 0.00 67

HEALTHY LIVING 0.00 0.00 0.00 168

HOME & LIVING 1.00 0.04 0.08 100

IMPACT 0.00 0.00 0.00 80

LATINO VOICES 0.00 0.00 0.00 35

MEDIA 0.00 0.00 0.00 84

MONEY 0.00 0.00 0.00 39

PARENTING 0.50 0.01 0.02 235

PARENTS 0.00 0.00 0.00 102

POLITICS 0.75 0.92 0.83 3554

QUEER VOICES 1.00 0.01 0.01 194

RELIGION 1.00 0.02 0.03 66

SCIENCE 0.00 0.00 0.00 45

SPORTS 1.00 0.01 0.01 137

STYLE 0.00 0.00 0.00 64

STYLE & BEAUTY 0.82 0.26 0.40 234

TASTE 0.00 0.00 0.00 48

TECH 0.00 0.00 0.00 61

THE WORLDPOST 0.00 0.00 0.00 100

TRAVEL 0.83 0.16 0.27 262

U.S. NEWS 0.00 0.00 0.00 29

WEDDINGS 0.00 0.00 0.00 78

WEIRD NEWS 0.00 0.00 0.00 73

WELLNESS 0.67 0.96 0.79 3429

WOMEN 0.00 0.00 0.00 81

WORLD NEWS 0.00 0.00 0.00 82

WORLDPOST 0.00 0.00 0.00 71

accuracy 0.71 13890

macro avg 0.24 0.08 0.08 13890

weighted avg 0.62 0.71 0.62 13890

Model: RandomForestClassifier

Accuracy: 0.6177105831533477

Precision: 0.5178912283714744

Recall: 0.6177105831533477

F1 Score: 0.538635506109906

Confusion Matrix:

O imagine care conține text, captură de ecran, model, petic

Descriere generată automat

precision recall f1-score support

ARTS 0.00 0.00 0.00 36

ARTS & CULTURE 0.00 0.00 0.00 29

BLACK VOICES 0.00 0.00 0.00 89

BUSINESS 0.00 0.00 0.00 149

COLLEGE 0.00 0.00 0.00 26

COMEDY 0.00 0.00 0.00 127

CRIME 0.00 0.00 0.00 81

CULTURE & ARTS 0.00 0.00 0.00 23

DIVORCE 0.00 0.00 0.00 72

EDUCATION 0.00 0.00 0.00 22

ENTERTAINMENT 0.49 0.86 0.62 3463

ENVIRONMENT 0.00 0.00 0.00 34

FIFTY 0.00 0.00 0.00 37

FOOD & DRINK 0.00 0.00 0.00 156

GOOD NEWS 0.00 0.00 0.00 28

GREEN 0.00 0.00 0.00 67

HEALTHY LIVING 0.00 0.00 0.00 168

HOME & LIVING 1.00 0.01 0.02 100

IMPACT 0.00 0.00 0.00 80

LATINO VOICES 0.00 0.00 0.00 35

MEDIA 0.00 0.00 0.00 84

MONEY 0.00 0.00 0.00 39

PARENTING 0.00 0.00 0.00 235

PARENTS 0.00 0.00 0.00 102

POLITICS 0.80 0.75 0.78 3554

QUEER VOICES 0.00 0.00 0.00 194

RELIGION 1.00 0.02 0.03 66

SCIENCE 0.00 0.00 0.00 45

SPORTS 0.00 0.00 0.00 137

STYLE 0.00 0.00 0.00 64

STYLE & BEAUTY 1.00 0.04 0.08 234

TASTE 0.00 0.00 0.00 48

TECH 0.00 0.00 0.00 61

THE WORLDPOST 0.00 0.00 0.00 100

TRAVEL 0.00 0.00 0.00 262

U.S. NEWS 0.00 0.00 0.00 29

WEDDINGS 0.00 0.00 0.00 78

WEIRD NEWS 0.00 0.00 0.00 73

WELLNESS 0.66 0.85 0.74 3429

WOMEN 0.00 0.00 0.00 81

WORLD NEWS 0.00 0.00 0.00 82

WORLDPOST 0.00 0.00 0.00 71

accuracy 0.62 13890

macro avg 0.12 0.06 0.05 13890

weighted avg 0.52 0.62 0.54 13890

Model: LinearSVC

Accuracy: 0.7476601871850252

Precision: 0.700393112091144

Recall: 0.7476601871850252

F1 Score: 0.7127202676882683

Confusion Matrix:

O imagine care conține text, captură de ecran, model, petic

Descriere generată automat

precision recall f1-score support

ARTS 0.00 0.00 0.00 36

ARTS & CULTURE 0.14 0.03 0.06 29

BLACK VOICES 0.31 0.11 0.17 89

BUSINESS 0.33 0.19 0.24 149

COLLEGE 0.33 0.23 0.27 26

COMEDY 0.36 0.11 0.17 127

CRIME 0.43 0.35 0.38 81

CULTURE & ARTS 0.00 0.00 0.00 23

DIVORCE 0.64 0.50 0.56 72

EDUCATION 0.20 0.09 0.13 22

ENTERTAINMENT 0.79 0.88 0.83 3463

ENVIRONMENT 0.29 0.06 0.10 34

FIFTY 0.57 0.11 0.18 37

FOOD & DRINK 0.56 0.51 0.53 156

GOOD NEWS 0.00 0.00 0.00 28

GREEN 0.30 0.13 0.19 67

HEALTHY LIVING 0.29 0.03 0.05 168

HOME & LIVING 0.69 0.55 0.61 100

IMPACT 0.21 0.07 0.11 80

LATINO VOICES 1.00 0.09 0.16 35

MEDIA 0.47 0.10 0.16 84

MONEY 0.50 0.13 0.20 39

PARENTING 0.39 0.32 0.35 235

PARENTS 0.22 0.08 0.12 102

POLITICS 0.80 0.90 0.85 3554

QUEER VOICES 0.70 0.49 0.58 194

RELIGION 0.42 0.20 0.27 66

SCIENCE 0.48 0.24 0.32 45

SPORTS 0.42 0.31 0.36 137

STYLE 0.21 0.05 0.08 64

STYLE & BEAUTY 0.66 0.62 0.64 234

TASTE 0.16 0.06 0.09 48

TECH 0.50 0.28 0.36 61

THE WORLDPOST 0.42 0.16 0.23 100

TRAVEL 0.66 0.60 0.63 262

U.S. NEWS 0.20 0.03 0.06 29

WEDDINGS 0.64 0.50 0.56 78

WEIRD NEWS 0.18 0.07 0.10 73

WELLNESS 0.78 0.93 0.85 3429

WOMEN 0.27 0.10 0.14 81

WORLD NEWS 0.06 0.01 0.02 82

WORLDPOST 0.23 0.08 0.12 71

accuracy 0.75 13890

macro avg 0.40 0.25 0.28 13890

weighted avg 0.70 0.75 0.71 13890