
This sf.net email is sponsored by:ThinkGeek
Welcome to geek heaven.
http://thinkgeek.com/sf

_____

Spamassassin-talk mailing list
Spamassassin-talk@lists.sourceforge.net
https://lists.sourceforge.net/lists/listinfo/spamassassin-talk

hi my name is madeline and i ' m 20 years old .
i have fit body , blonde hair and blue eyes .
i would like to get to know men . i am a bit shy but
open for everything .
you can contact me now at : http : / / muck . folosko
. com / 575 r . html
(registration is free of charge)
see you soon

hi cindy and sally ,
before i forget here are the phone numbers
where we can be reached for the next several
days : june 8 th and 10 th la quinta in alexandria ,
la 318 442 3700 june 9 th monmouth plantation
nachez 1 - 800 - 828 - 4531
we will be in natchez until later in the afternoon
on the 10 th

TM4SPAM

Sara Borello - 882793

Keita Jacopo Viganò - 870980

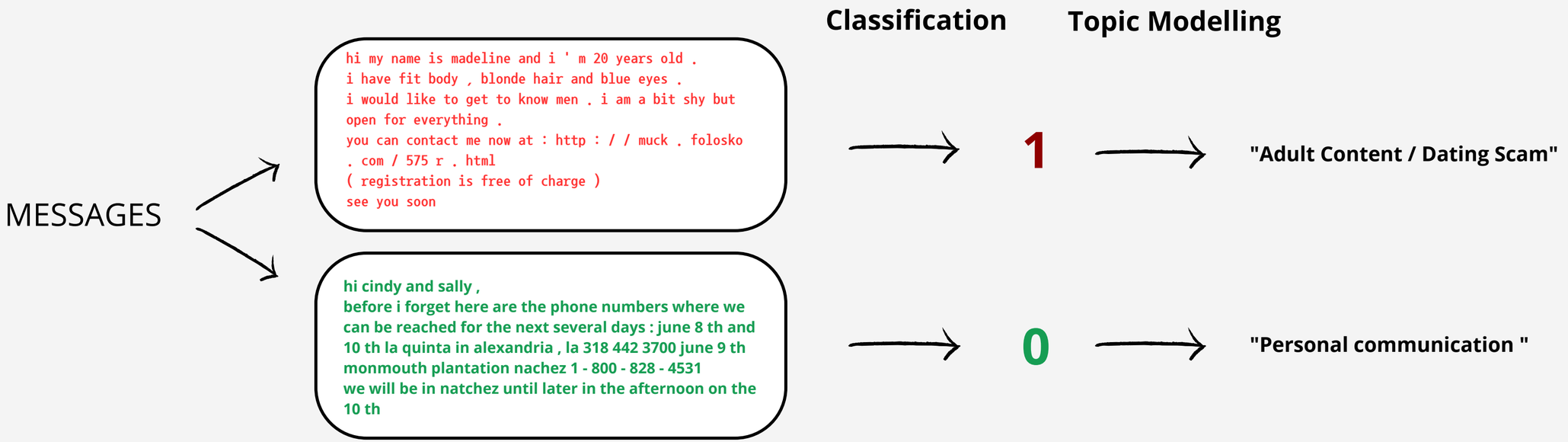
The Problem of Spam Classification

The proliferation of digital communication channels has intensified the presence of spam messages, posing challenges in terms of:

- User experience degradation
- Cybersecurity threats (scams, phishing, malware)

The goal of this project is to develop and evaluate an integrated approach to spam detection in text messages, by combining:

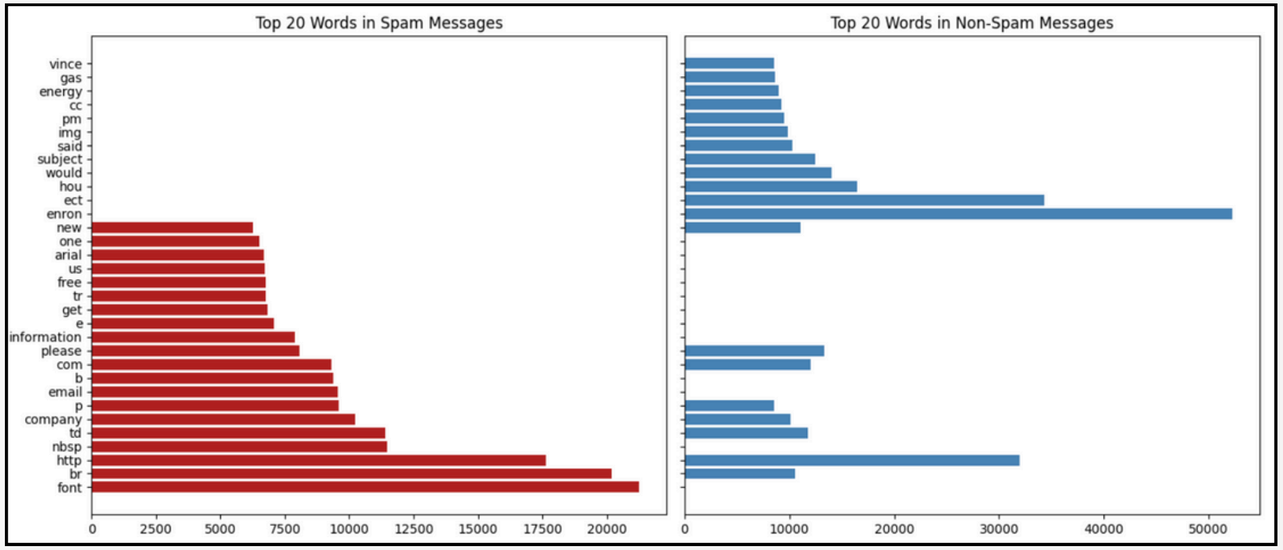
- **Supervised machine learning techniques** for classification (Logistic Regression and DistilBERT)
- **Unsupervised topic modeling** (via BERTopic)

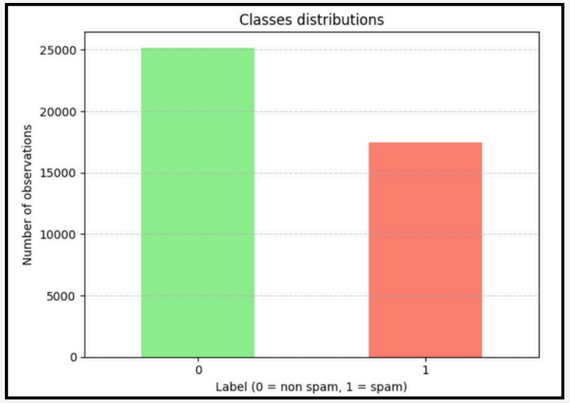


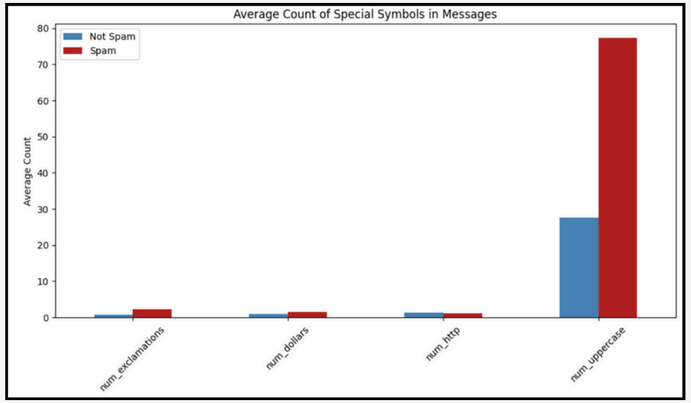
Dataset Overview & PreProcessing

Dataset

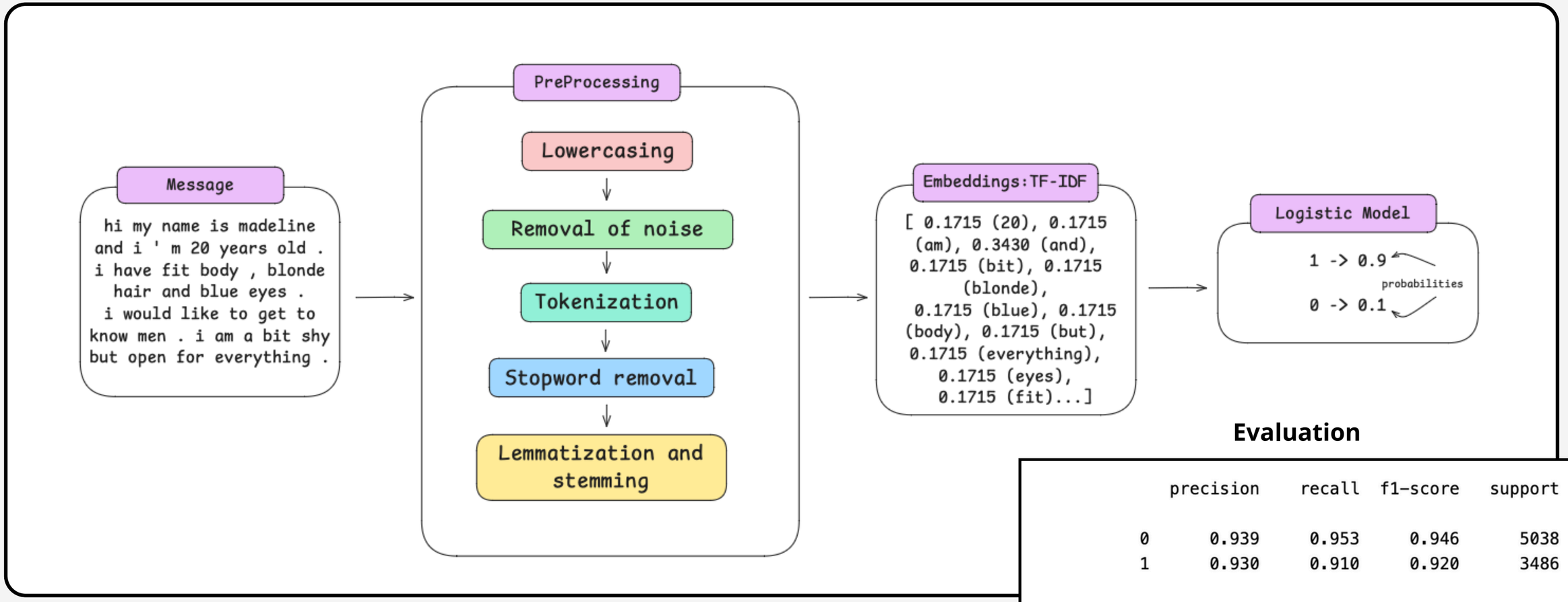
- **Source**: https://huggingface.co/datasets/FredZhang7/all-scam-spam
- **Content**: a total of 42,619 preprocessed text messages and emails
- **Columns**: text (string), is_spam (binary)
- **Language Filtering**: more than 40,000 out of approximately 42,000 total entries are in English, while other languages such as Spanish, French, and German appear only marginally
- **Train & Test:** 80% Train and 20% Test







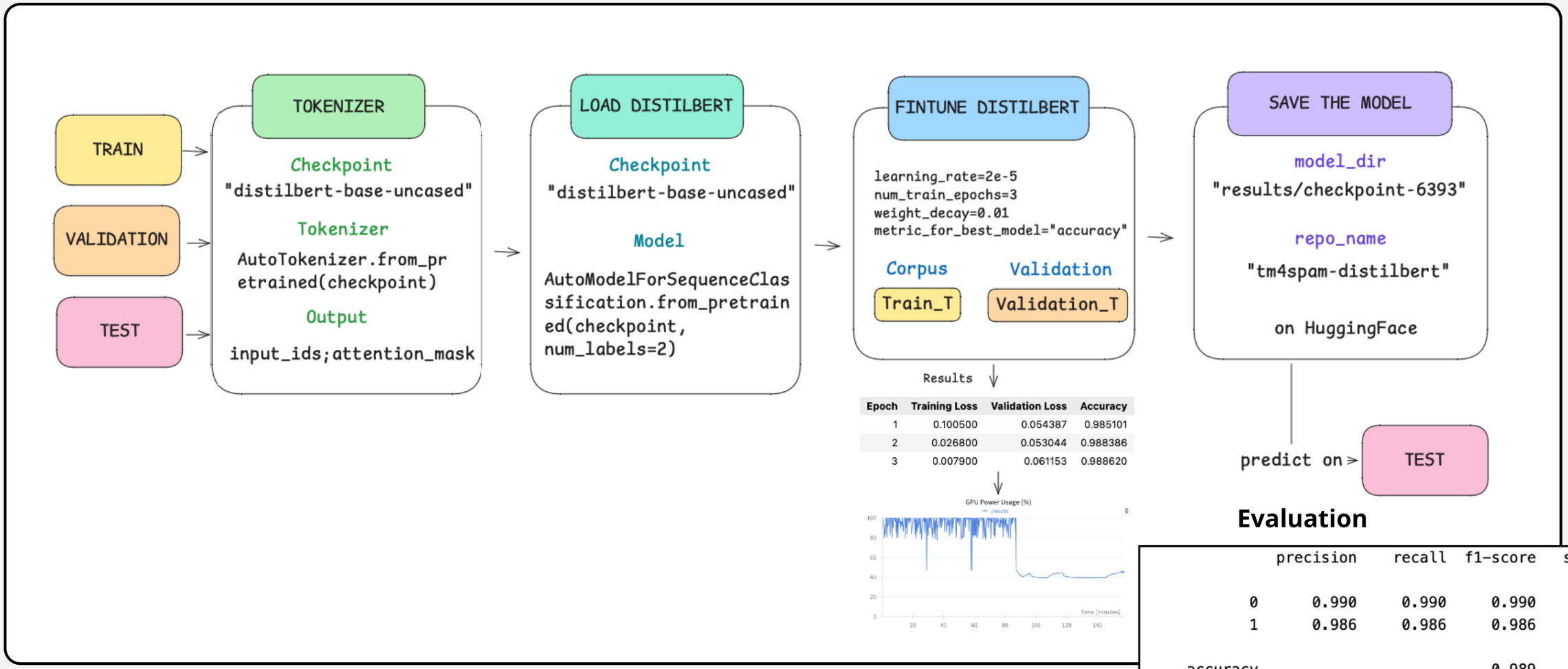
Classification: Logistic Model



Evaluation

	precision	recall	f1-score	support
0	0.939	0.953	0.946	5038
1	0.930	0.910	0.920	3486
accuracy			0.935	8524
macro avg	0.935	0.932	0.933	8524
weighted avg	0.935	0.935	0.935	8524

Classification: Fine-Tuned DistilBert



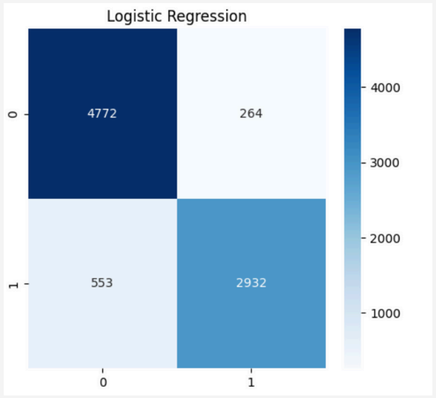
Evaluation

	precision	recall	f1-score	support
0	0.990	0.990	0.990	5038
1	0.986	0.986	0.986	3486
accuracy			0.989	8524
macro avg	0.988	0.988	0.988	8524
weighted avg	0.989	0.989	0.989	8524

Classification Results Comparison

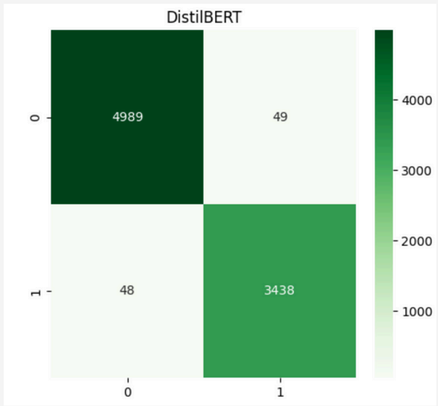
Logistic regression

	precision	recall	f1-score	support
0	0.896	0.948	0.921	5036
1	0.917	0.841	0.878	3485
accuracy			0.904	8521
macro avg	0.907	0.894	0.899	8521
weighted avg	0.905	0.904	0.903	8521

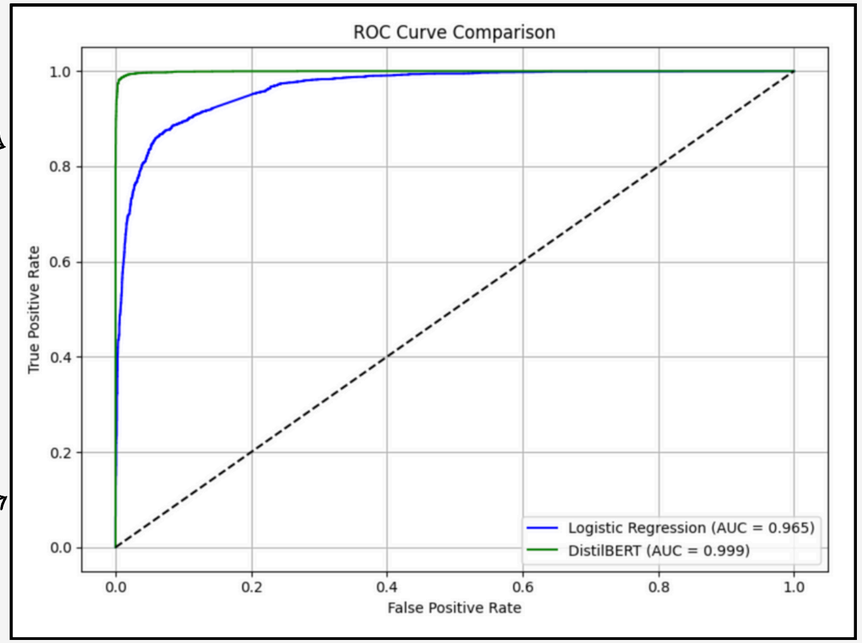


Finetuned DistilBert

	precision	recall	f1-score	support
0	0.990	0.990	0.990	5038
1	0.986	0.986	0.986	3486
accuracy			0.989	8524
macro avg	0.988	0.988	0.988	8524
weighted avg	0.989	0.989	0.989	8524

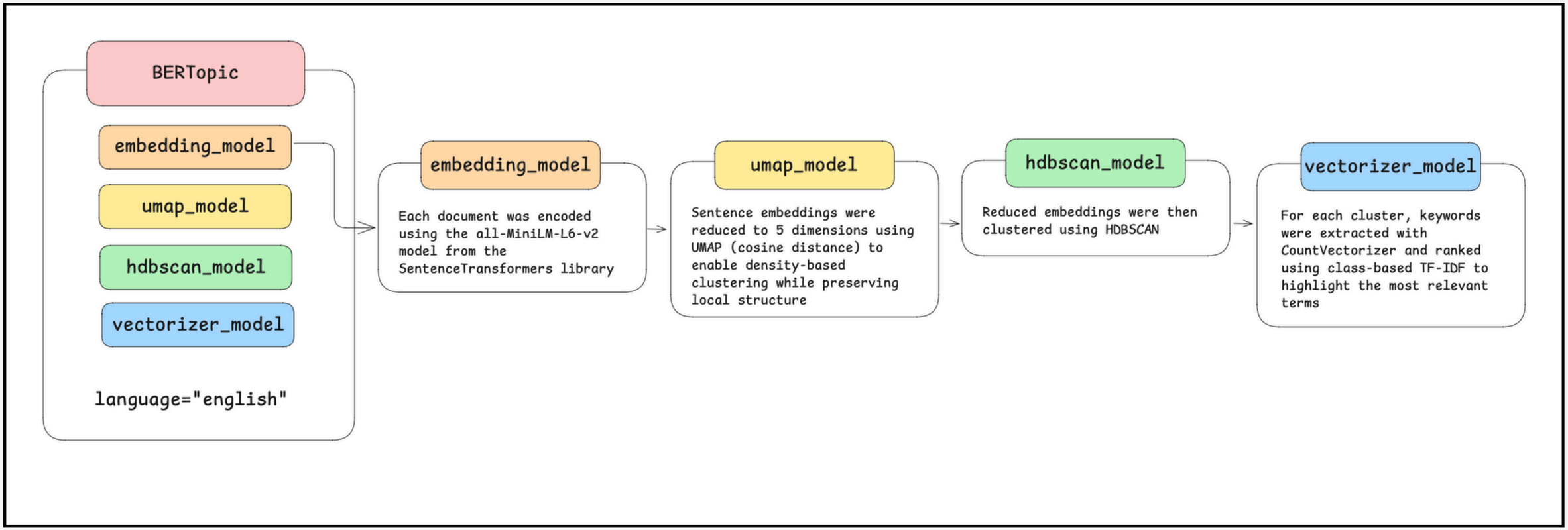


Comparison



1 = SPAM 0 = NON SPAM

Topic Modeling with BERTopic



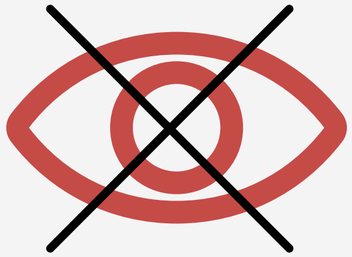
Spam messages Topics

Topic	Term	Weight
0	girl	0.0372
	adult	0.0290
	sex	0.0230
	video	0.0225
	site	0.0216
	free	0.0207
	membership	0.0205
	date	0.0192
	movi	0.0190
	http	0.0189

Topic	Term	Weight
5	statement	0.0146
	compani	0.0144
	stock	0.0142
	invest	0.0129
	trade	0.0105
	expect	0.0095
	industri	0.0094
	investor	0.0092
	materi	0.0092
	market	0.0091

Topic	Term	Weight
10	prize	0.0972
	urgent	0.0746
	txt	0.0730
	claim	0.0712
	award	0.0702
	ur	0.0701
	cash	0.0481
	holiday	0.0446
	draw	0.0436
	mobil	0.0411

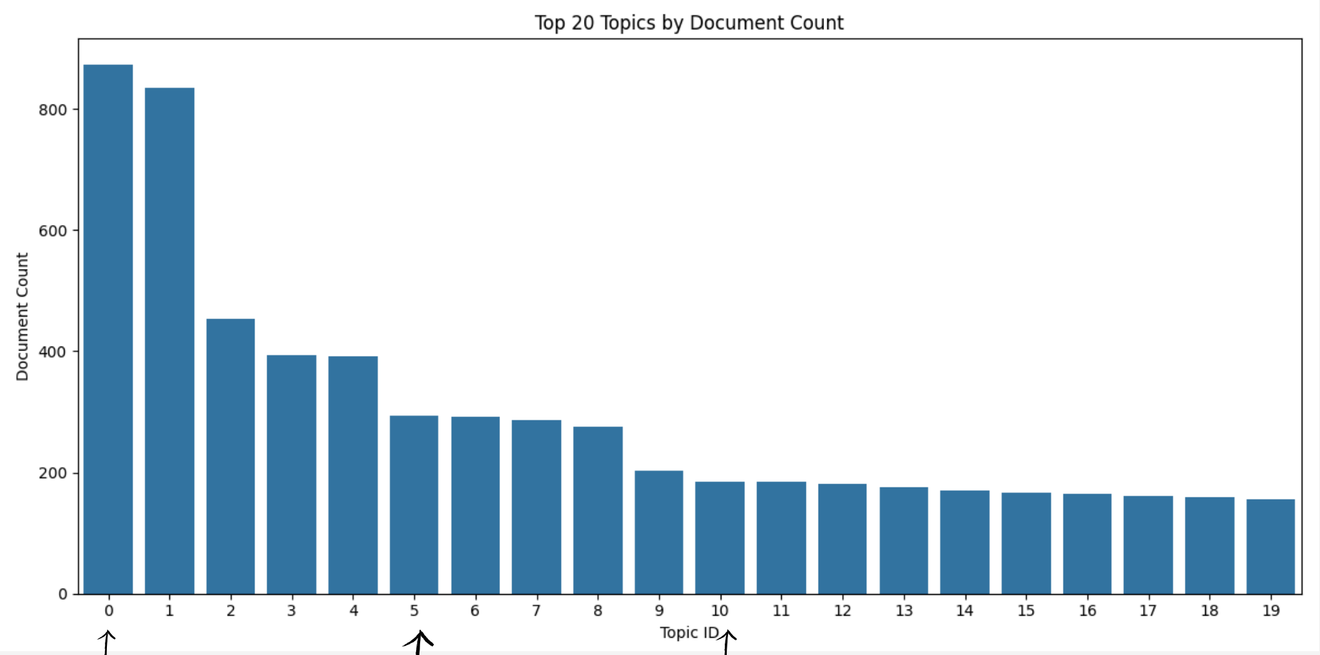
Adult Content



Investment & Financial Markets

Prize & Scam Alerts

Spam messages: Evaluation



Metrics

Jaccard Diversity Score **0.9941**

measures the distinctiveness of topics by quantifying the overlap between their top keywords

Coherence score **0.5938**

reflects the semantic consistency of the keywords within each topic, with higher values indicating that the words tend to co-occur in similar contexts.

Explanation

Topic 1	
width:	0.0361
font:	0.036
td:	0.0349
size:	0.0339
tr:	0.0335
height:	0.0285
br:	0.0281
tabl:	0.027
href:	0.0247
helvetica:	0.0235

Topic 3	
la:	0.1594
en:	0.1296
para:	0.1135
el:	0.1077
lo:	0.0853
del:	0.0758
da:	0.0752
su:	0.067
le:	0.065
est:	0.0592

? ?

Non Spam messages Topics

Topic	Term	Weight
2	interview	0.0690
	vinc	0.0293
	resum	0.0264
	kaminski	0.0247
	research	0.0183
	thank	0.0173
	pm	0.0148
	shirley	0.0137
	pleas	0.0135
	forward	0.0128

Recruitment Interview

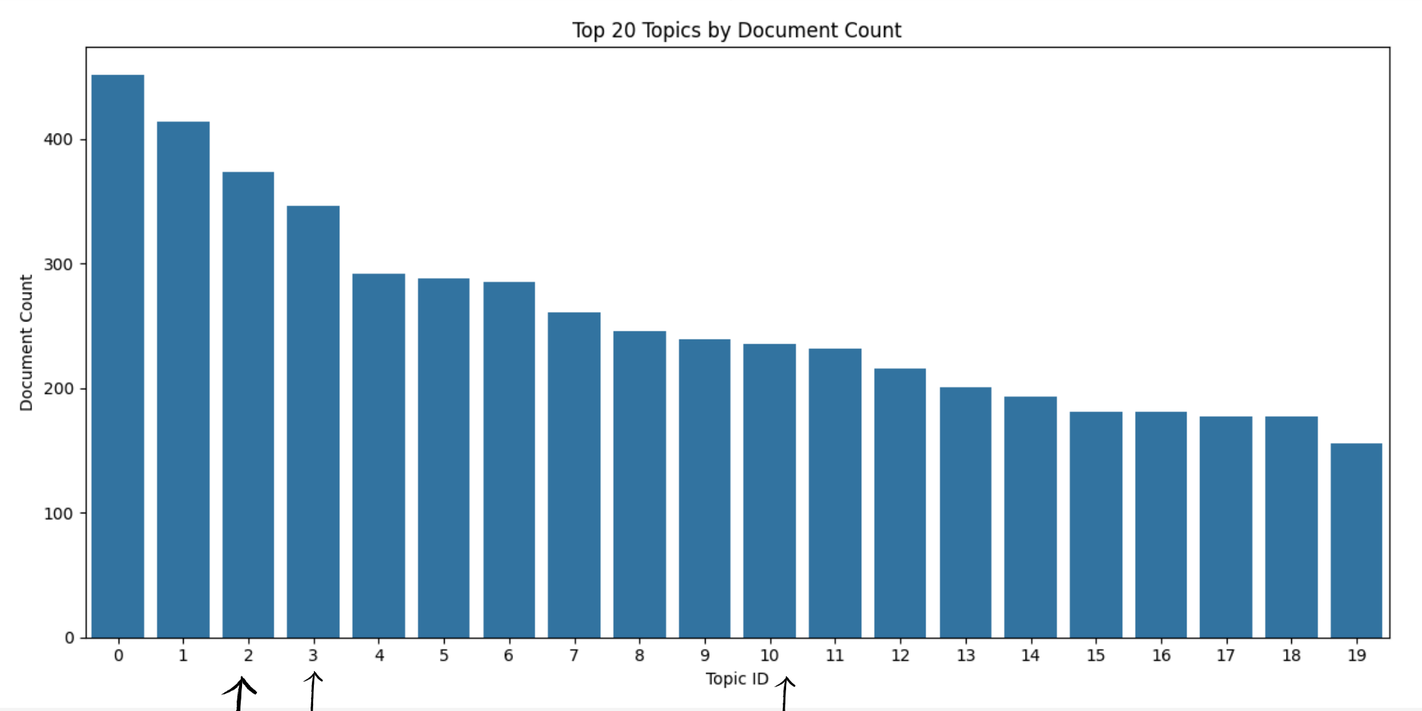
Topic	Term	Weight
3	meet	0.0679
	prc	0.0261
	attend	0.0230
	pleas	0.0188
	eb	0.0169
	thank	0.0163
	dinner	0.0163
	pm	0.0158
	vinc	0.0157
	th	0.0153

Corporate Meeting

Topic	Term	Weight
10	nation	0.0221
	state	0.0182
	peopl	0.0177
	world	0.0169
	govern	0.0167
	war	0.0164
	polit	0.0159
	american	0.0148
	unit	0.0135
	attack	0.0119

Political or international affairs

Non Spam messages: Evaluation



Metrics

Jaccard Diversity Score **0.9906**

measures the distinctiveness of topics by quantifying the overlap between their top keywords

Coherence score **0.5714**

reflects the semantic consistency of the keywords within each topic, with higher values indicating that the words tend to co-occur in similar contexts.

Explanation

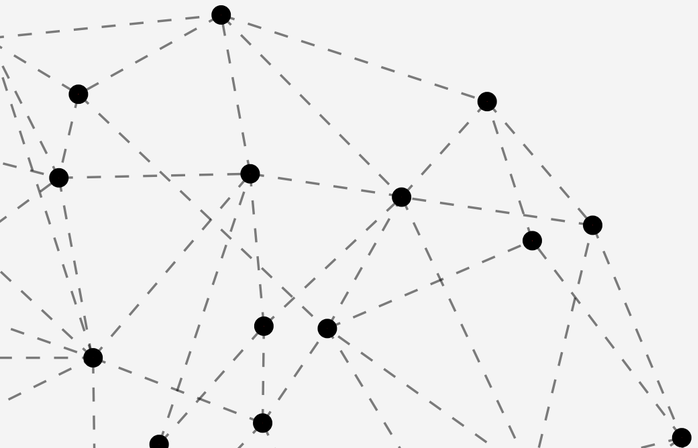
```
 Topic 0
schedul: 0.2274
hour: 0.2132
start: 0.1845
pars: 0.1733
detect: 0.1691
award: 0.1623
date: 0.1576
portland: 0.1516
log: 0.1433
iso: 0.1418
```

```
 Topic 1
spam: 0.0697
email: 0.0287
mail: 0.0262
header: 0.0259
list: 0.0219
score: 0.0217
test: 0.0216
sponsor: 0.0215
dont: 0.0169
train: 0.0161
```

? ?

Classification Results

FineTuned DistilBERT > Logistic Regression



Topic Modeling Insights

BERTopic identified clear patterns

- **Spam**: adult content, financial scams, fake prize notifications
- **Non-spam**: job interviews, meeting arrangements, political discussions

Further developments

- Extend to multilingual spam detection
- Use topic distributions as features in supervised models
- Explore few-shot / zero-shot learning with the annotated subset