AI-written vs Human-written abstract: text mining classification

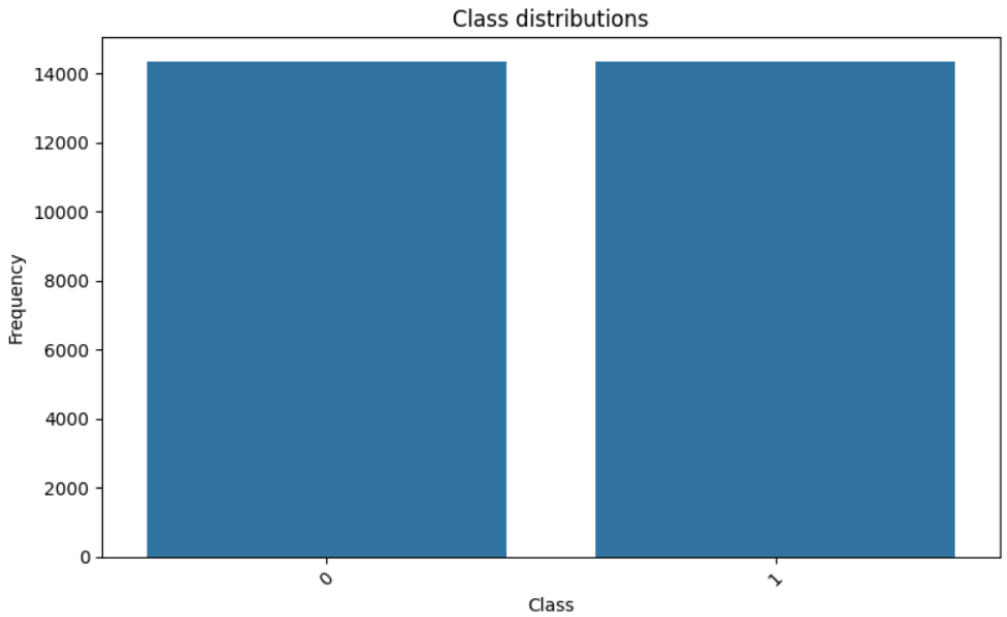
Francesco Galardi

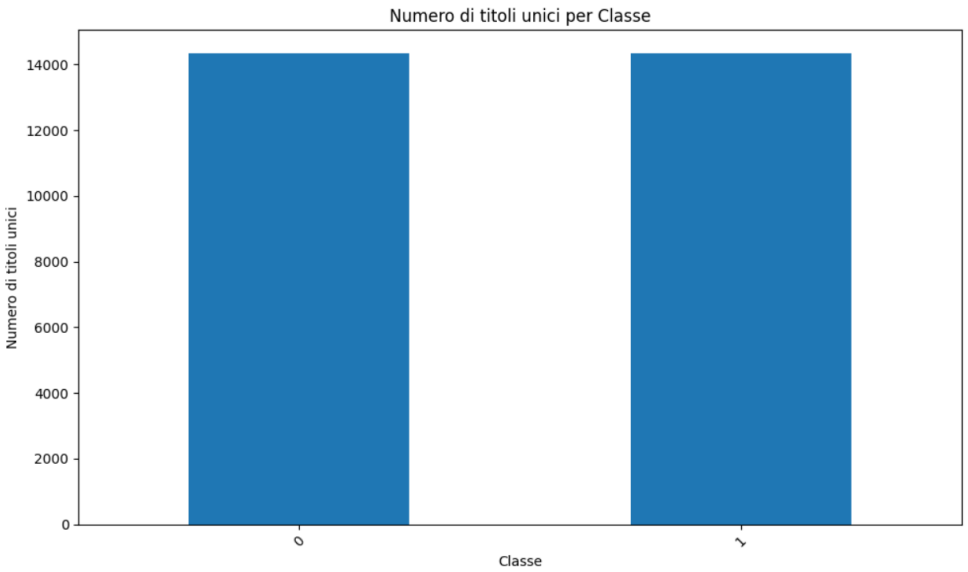
Introduction

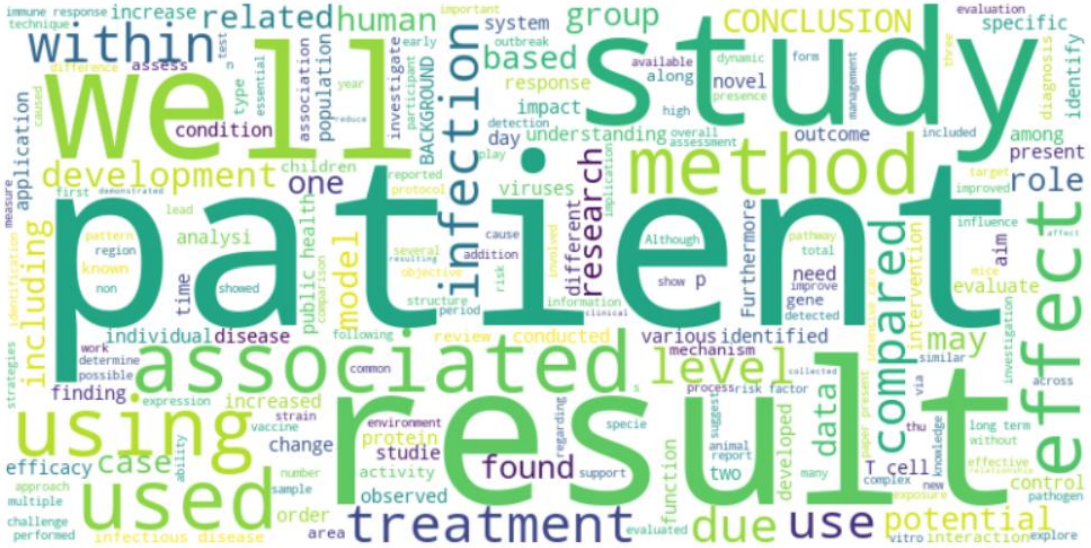
- Since AI is becoming increasingly present in everyone's daily life, it is becoming difficult to distinguish whether a text is written by a human or generated by artificial intelligence. Here we will consider as texts some abstracts associated to medical-scientific articles.
- The idea is to train a classification model in order to classify scientific abstract into human written or ai-written taking into account how the abstract is written

Dataset

- Features: title,abstract,label
- 28662 observations
- 14331 different titles

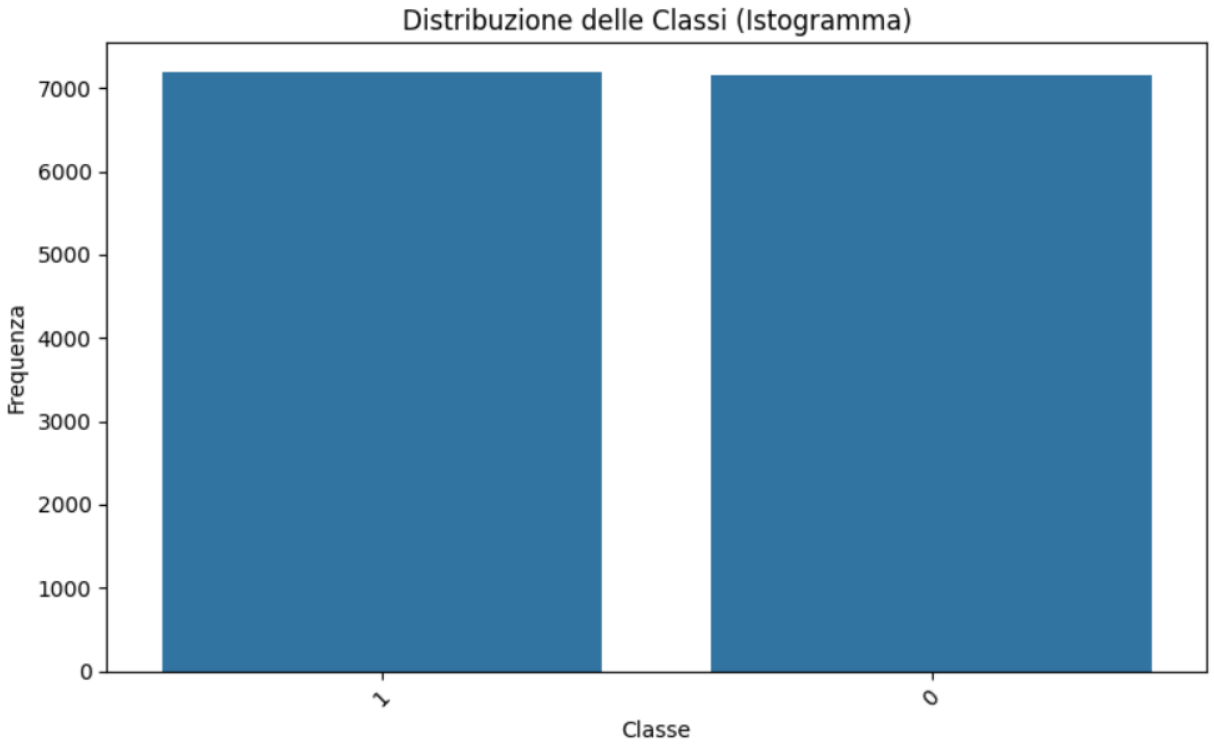






Preprocessing

- For each title, randomly selection of one of the two observations (this still ensure balance)
- Elimination of feature title



Pipeline

- Tokenization
- Stopwords remover
- Text representation: tf-idf and word2Vec models are taken into account
- Classification model: logistic regression, naive-bayes,decision tree,svc,random forest and xgboost are taken into account

Model selection phase

- Each possible combination (text representation model+classification model) is validated via nested loop cross validation on accuracy score
- Best 3 classification models for both tf-idf and word2Vec are chosen
- The resultant 6 combinations (alogside with their hyperparams founded during first step) are tested 2 by 2 against wilcoxon test or t-test based on wheter the differences of the accuracies fold-by-fold of the two models are gaussian or not

Tf-idf+logistic regression:

```
Accuracy fold-by-fold: [0.95466434 0.95466434 0.95989538 0.96338274 0.95636998 0.95549738
 0.96335079 0.94589878 0.96858639 0.95724258]
Media accuracy: 0.9580
Deviazione standard: 0.0059
{'clf__C': 1}
```

Tf-idf+nayve-bayes:

```
Accuracy fold-by-fold: [0.88666085 0.85265911 0.87183958 0.87619878 0.86823735 0.86910995
 0.88394415 0.87870855 0.89179756 0.86038394]
Media accuracy: 0.8740
Deviazione standard: 0.0114
{'clf__alpha': 0.1}
```

Tf-idf+decision tree:

```
Accuracy fold-by-fold: [0.87968614 0.88404534 0.88404534 0.90671316 0.87958115 0.89092496
 0.88917976 0.88045375 0.89703316 0.88394415]
Media accuracy: 0.8876
Deviazione standard: 0.0083
{'clf  max depth': 10}
```

Tf-idf+svc:

```
Accuracy fold-by-fold: [0.95640802 0.95640802 0.9625109  0.96163906 0.96335079 0.95375218
 0.96596859 0.94851658 0.97033159 0.96073298]
Media accuracy: 0.9600
Deviazione standard: 0.0060
{'clf__C': 1}
```

Tf-idf+random forest:

```
Accuracy fold-by-fold: [0.9537925  0.95030514 0.93374019 0.95466434 0.94415358 0.95113438
 0.95811518 0.94589878 0.96247818 0.94851658]
Media accuracy: 0.9503
Deviazione standard: 0.0076
{'clf  max depth': 20, 'clf  n estimators': 200}
```

Tf-idf+xgboost:

```
Accuracy fold-by-fold: [0.9625109  0.96774194 0.95989538 0.97297297 0.96945899 0.96422339
 0.96684119 0.95636998 0.97382199 0.96596859]
Media accuracy: 0.9660
Deviazione standard: 0.0052
{'clf__max_depth': 3, 'clf__n_estimators': 200}
```

W2Vec+logistic regression:

```
Accuracy fold-by-fold: [0.95292066 0.94768963 0.95902354 0.95117698 0.94240838 0.95636998
 0.94938918 0.94153578 0.95811518 0.95200698]
Media accuracy: 0.9511
Deviazione standard: 0.0057
{'clf__C': 10}
```

W2Vec+decision tree:

```
Accuracy fold-by-fold: [0.83958152 0.81778553 0.83522232 0.82476024 0.84991274 0.81326353
 0.83769634 0.82984293 0.84991274 0.84904014]
Media accuracy: 0.8347
Deviazione standard: 0.0125
{'clf__max_depth': 10}
```

W2Vec+svc:

```
Accuracy fold-by-fold: [0.96512642 0.95640802 0.9625109  0.95989538 0.95200698 0.96247818
 0.95200698 0.95462478 0.96247818 0.95200698]
Media accuracy: 0.9580
Deviazione standard: 0.0049
{'clf  C': 100}
```

W2Vec+random forest:

```
Accuracy fold-by-fold: [0.90932868 0.89712293 0.9163034  0.91281604 0.91361257 0.90750436
 0.90837696 0.91012216 0.90924956 0.90139616]
Media accuracy: 0.9086
Deviazione standard: 0.0054
{'clf__max_depth': 20, 'clf__n_estimators': 200}
```

W2Vec+xgboost:

```
Accuracy fold-by-fold: [0.95117698 0.93548387 0.95466434 0.95292066 0.94764398 0.94851658
 0.95811518 0.95287958 0.95200698 0.94764398]
Media accuracy: 0.9501
Deviazione standard: 0.0058
{'clf__max_depth': 3, 'clf__n_estimators': 200}
```

Final model construction

We now select the 6 best models (with hyperparameters founded during nested cv) basing our decision on the accuracies founded during the validation:

- Tf-idf: logistic regression, svc and xgboost
- word2Vec: logistic regression, svc and xgboost

After comparing the models exploiting statistical tests, the idea is to create an ensemble with the statistically different configurations in order to maximize the accuracy and to generalyze better on unseen data

Final model construction

	vett1logreg	vett2logreg	vett1svc	vett2svc	vett1xgb	vett2xgb
vett1logreg						
vett2logreg	gaussian					
vett1svc	gaussian	gaussian				
vett2svc	gaussian	gaussian	gaussian			
vett1xgb	gaussian	gaussian	gaussian	gaussian		
vett2xgb	gaussian	gaussian	gaussian	gaussian	gaussian	

	vett1logreg	vett2logreg	vett1svc	vett2svc	vett1xgb	vett2xgb
vett1logreg						
vett2logreg	Different					
vett1svc	Different	Different				
vett2svc	Equal	Different	Equal			
vett1xgb	Different	Different	Different	Different		
vett2xgb	Different	Different	Different	Different	Different	

Final model construction

- Taking into account only the statistically different models to create the ensemble, we have two ways:

1. Create an ensemble with all models without counting tf-idf+logistic regression and tf-idf+svc
2. Create an ensemble with all models without counting word2Vec+svc

We are going to validate both configurations in order to choose the best one

Final model construction

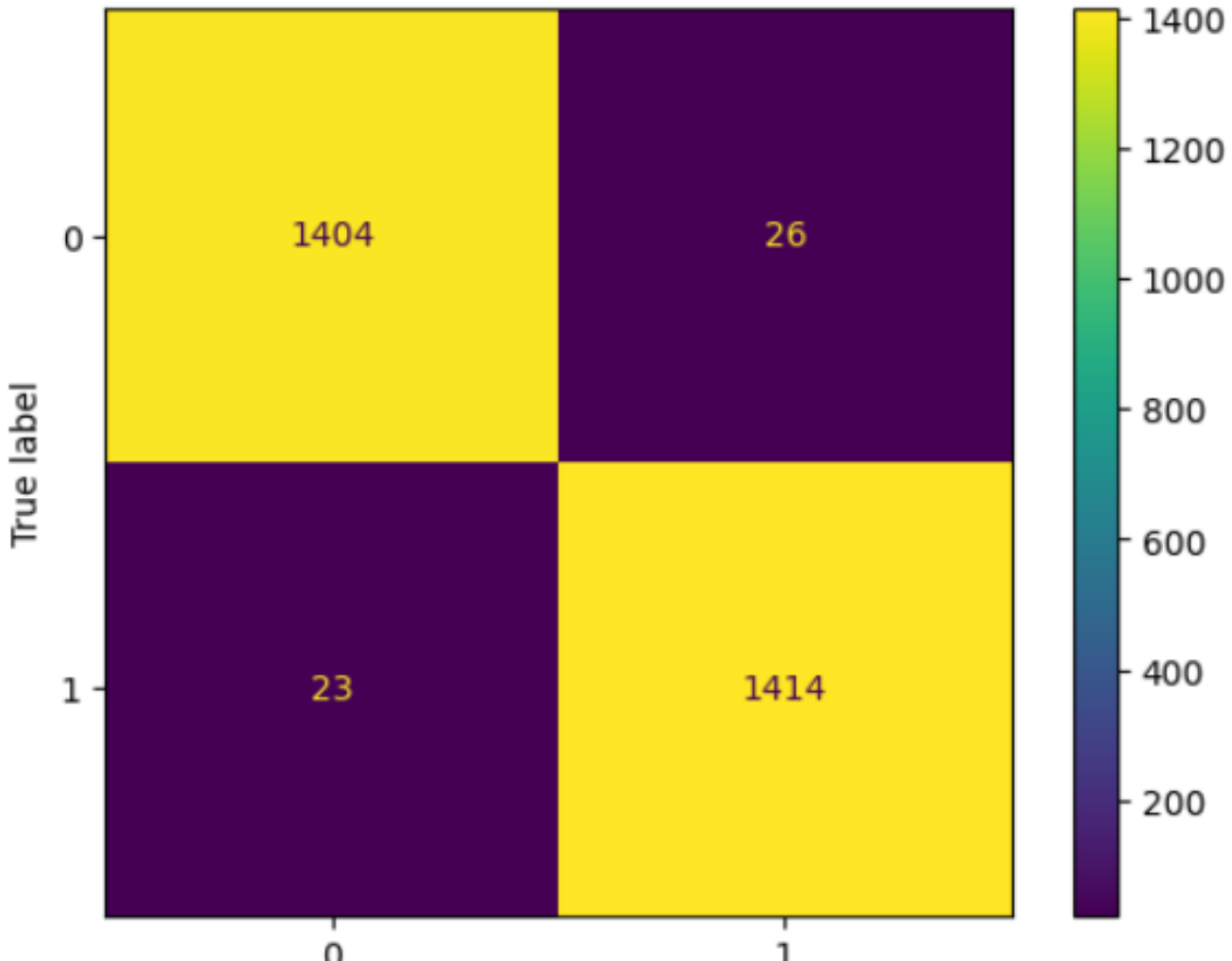
```
Accuracy fold-by-fold: [0.97384481 0.9668701  0.98169137 0.97646033 0.96858639 0.96945899
 0.97382199 0.96684119 0.97294939 0.97294939]
Media accuracy: 0.9723
Deviazione standard: 0.0044
Accuracy fold-by-fold: [0.97907585 0.97297297 0.97994769 0.97646033 0.97120419 0.96771379
 0.97643979 0.96335079 0.98080279 0.97731239]
Media accuracy: 0.9745
Deviazione standard: 0.0054
```

We see from the statistical test that the two ensembles are different, so we choose the second one as our final model because it ensure slitghly better accuracy

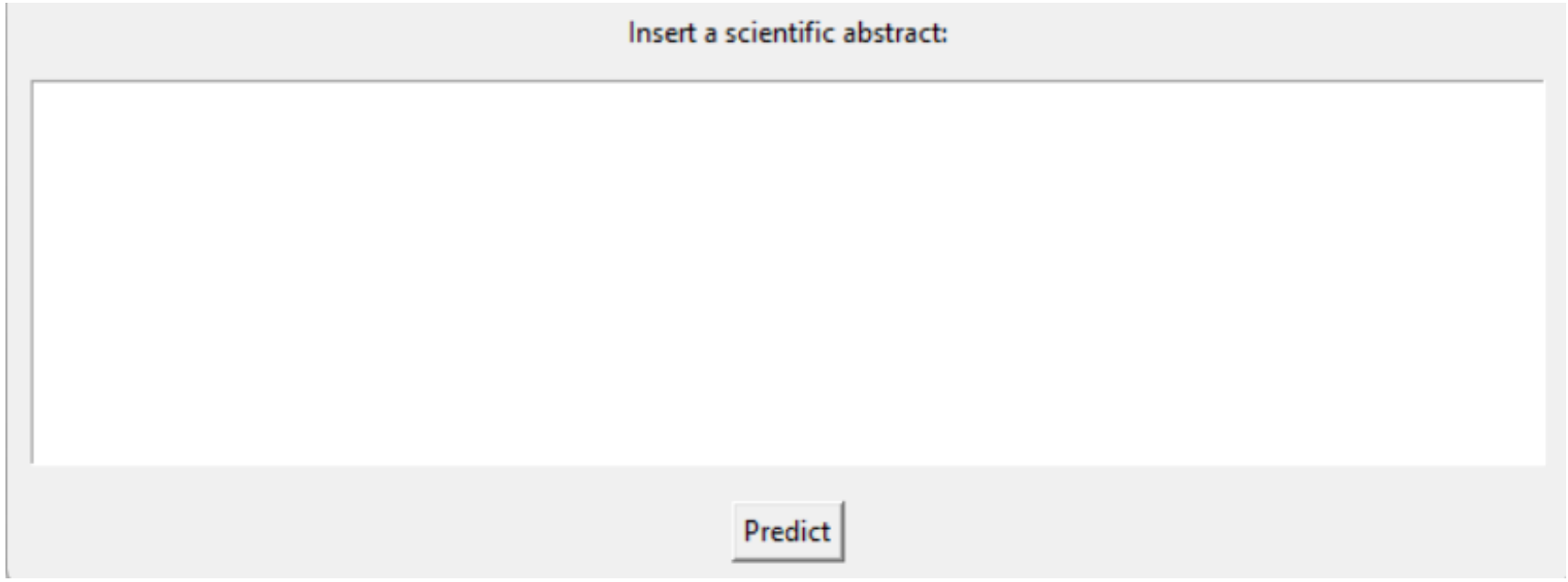
```
               vett_ensemble1 vett_ensemble2
vett_ensemble1
vett_ensemble2       gaussian
               vett_ensemble1 vett_ensemble2
vett_ensemble1
vett_ensemble2      Different
```

Final model performance evaluation

	precision	recall	f1-score	support
0	0.98	0.98	0.98	1430
1	0.98	0.98	0.98	1437
accuracy			0.98	2867
macro avg	0.98	0.98	0.98	2867
weighted avg	0.98	0.98	0.98	2867



Interface



Related work

- [1] Theocharopoulos, P. C., Anagnostou, P., Tsoukala, A., Georgakopoulos, S. V., Tasoulis, S. K., & Plagianakos, V. P. (2023, April 12). *Detection of fake generated scientific abstract*