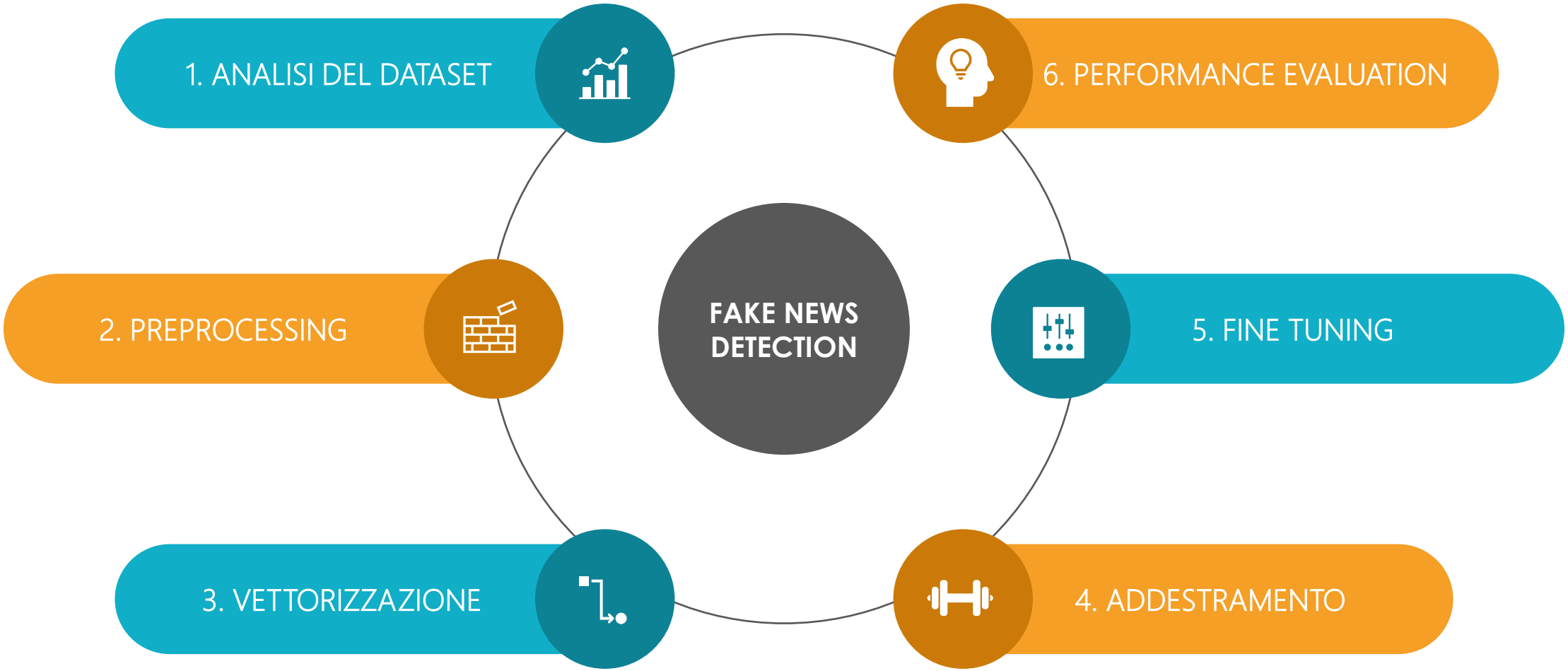
Fake News Detection

Pulerà Francesca & Zarantonello Massimo

Progetto di Machine Learning

Pipeline del Progetto



Analisi del Dataset

DATASET SELEZIONATO

Il dataset scelto include sia fake news che real news provenienti da **articoli online inglesi**, principalmente di natura **politica**.

UNIONE DEI DUE DATASET

Due **dataset separati**, uno per notizie reali e uno per fake news, sono stati uniti e mescolati per creare un **dataset di 44898 istanze** con una colonna "**class**" per indicarne la veridicità.

ANALISI VARIE

È stata identificata una prevalenza di notizie politiche. L'analisi della **lunghezza** degli articoli ha consentito di escludere quelli troppo brevi o carenti di contenuto, garantendo l'inclusione solo di articoli rilevanti nel processo di valutazione delle fake news.

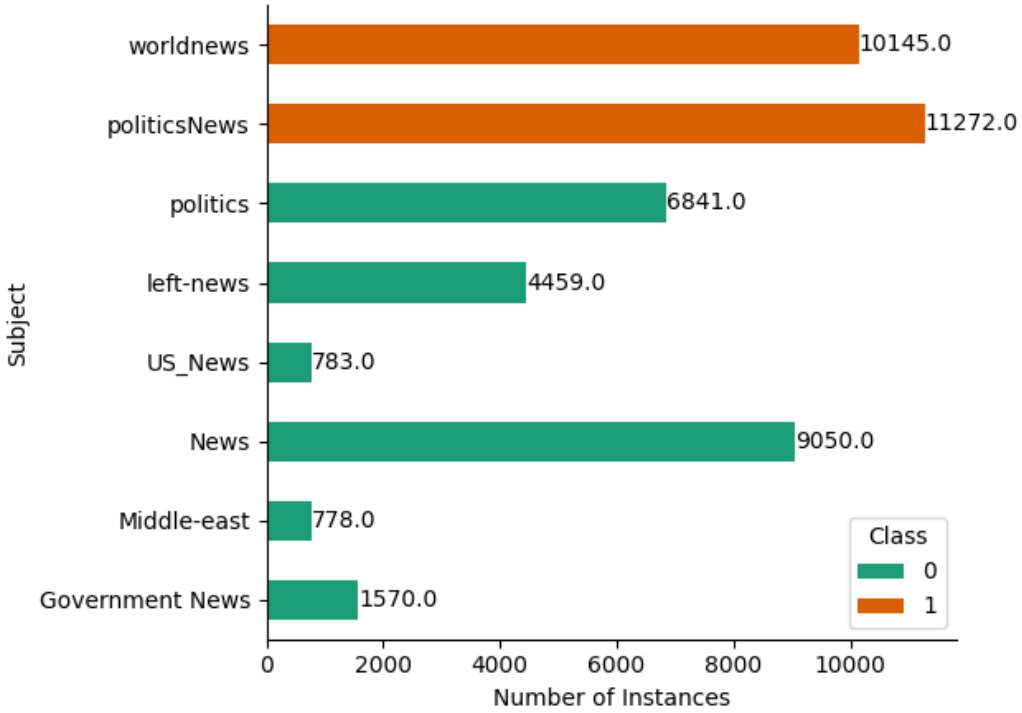
RIMOZIONE DEGLI ARTICOLI CON MENO DI 20 PAROLE

Sono stati **eliminati** 1131 articoli contenenti meno di 20 parole, riducendo leggermente il dataset ma migliorando la qualità complessiva dei dati e concentrando l'analisi sui **documenti più significativi e rappresentativi.**

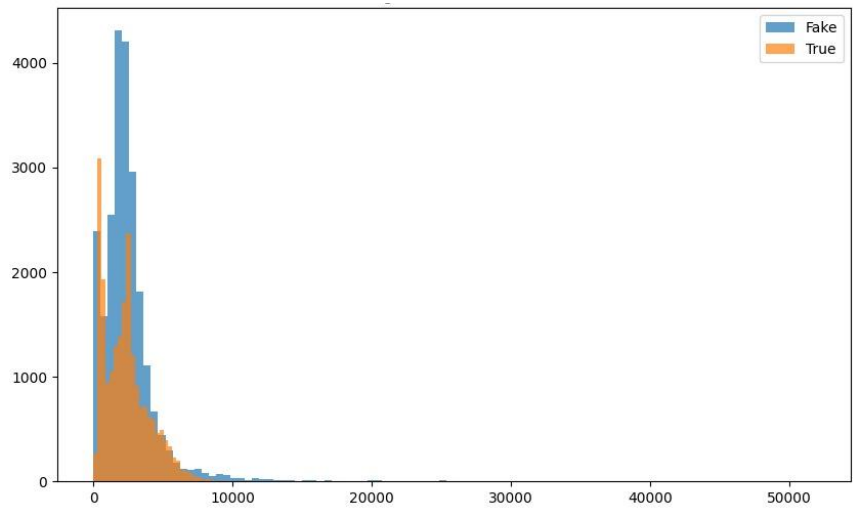
RIMOZIONE DI FEATURES

È stato **semplificato il dataset rimuovendo feature non cruciali**, ponendo così il focus sul testo dell'articolo e sulla sua classe, ottenendo quindi un dataset di **43,729 istanze**, di cui 22,313 fake news e 21,416 real news.
Il dataset generato è bilanciato.

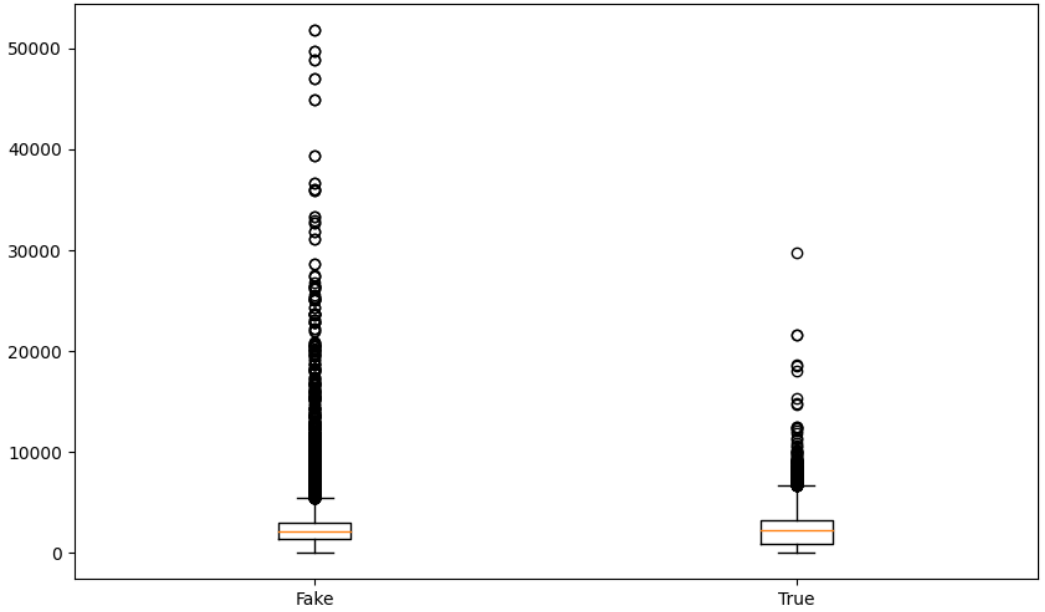
Distribuzione dei Topic data dall'unione dei due dataset



Statistiche sulla lunghezza dei testi:	
count	44898
mean	2469,109693
std	2.171.617.091
min	1.000.000
25%	1.234.000.000
50%	2.186.000.000
75%	3.105.000.000
max	51.794.000.000



Lunghezza delle istanze



BEFORE

index	title	text	subject	date	class
0	Donald Trump Sends Out Embarrassing New Year's Eve Message; This is Disturbing	Donald Trump just couldn t wish all Americans a Happy New Year and leave it at that. Instead, he had to give a shout out to his enemies, haters and the very dishonest fake news media. The former reality show star had just one job to do and he couldn t do it. As our Country rapidly grows stronger and smarter, I want to wish all of my friends, supporters, enemies, haters, and even the very dishonest Fake News Media, a Happy and Healthy New Year, President Angry Pants tweeted. 2018 will be a great year for America! As our Country rapidly grows stronger and smarter, I want to wish all of my friends, supporters, enemies, haters, and even the very dishonest Fake News Media, a Happy and Healthy New Year. 2018 will be a great year for America! Donald J. Trump (@realDonaldTrump)	News	December 31, 2017	0

index	title	text	subject	date	class
1	U.S., North Korea clash at U.N. forum over nuclear weapons	GENEVA (Reuters) - North Korea and the United States clashed at a U.N. forum on Tuesday over their military intentions towards one another, with Pyongyang s envoy declaring it would never put its nuclear deterrent on the negotiating table. Japan, well within reach of North Korea s missiles, said the world must maintain pressure on the reclusive country to rein in its nuclear and missile programs and now was not the time for a resumption of multi-party talks. North Korea has pursued its weapons programs in defiance of U.N. Security Council sanctions and ignored all calls, including from major ally China, to stop, prompting a bellicose exchange of rhetoric between the North and the United States. North Korea justifies its weapons programs, including its recent threat to fire missiles towards the U.S. Pacific territory of Guam, by pointing to perceived U.S. hostility, such as military exercises with South Korea this week. U.S. disarmament ambassador Robert	worldnews	August 22, 2017	1

AFTER

index	text	class
0	Donald Trump just couldn t wish all Americans ...	0
1	House Intelligence Committee Chairman Devin Nu...	0
2	BRUSSELS (Reuters) - NATO allies on Tuesday we...	1
3	LONDON (Reuters) - LexisNexis, a provider of l...	1

Preprocessing

Si converte il testo in minuscolo per garantire una rappresentazione uniforme del testo

Esegue una serie di operazioni di pre-processing, come la rimozione di link URL e tag HTML

La punteggiatura non aggiunge necessariamente significato al testo e può essere considerata "rumore"

Suddivisione del testo in unità più piccole, chiamate token

Rimozione di parole comuni come articoli e congiunzioni, prive di informazioni specifiche

Riduce ad una radice comune parole che hanno approssimativamente lo stesso significato

Funzione che processa il testo

- Sostituzione di '\n'
- Rimozione del testo tra parentesi quadre
- Rimozione dei link URL
- Rimozione dei tag HTML
- Rimozione dei numeri
- Rimozione del carattere '_'

Tokenizzazione

- Suddivisione in unità significative
- Standardizzazione del testo
- Creazione di vocabolari
- Preparazione per l'elaborazione automatica

Lemmatizzazione

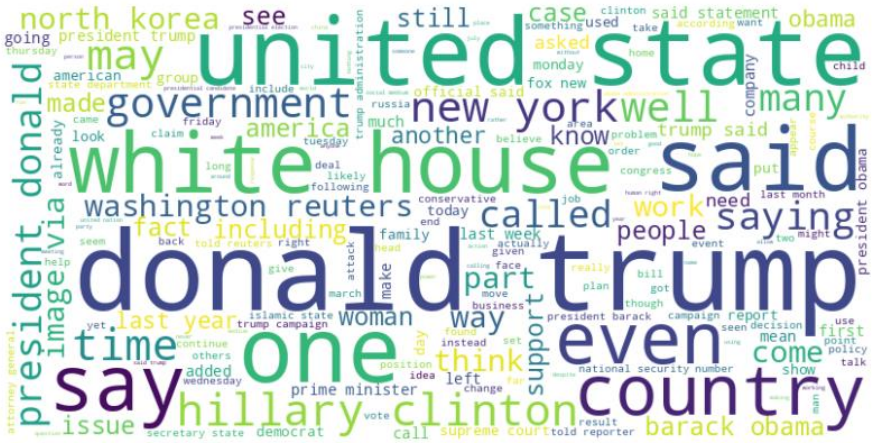
Questo procedimento non si basa semplicemente sul troncare la parola, ma su un meccanismo molto più complesso. Ad esempio, lemmatizzare la serie di parole correre, corro, corriamo e correremo, produrrebbe il lemma "correre".

BEFORE

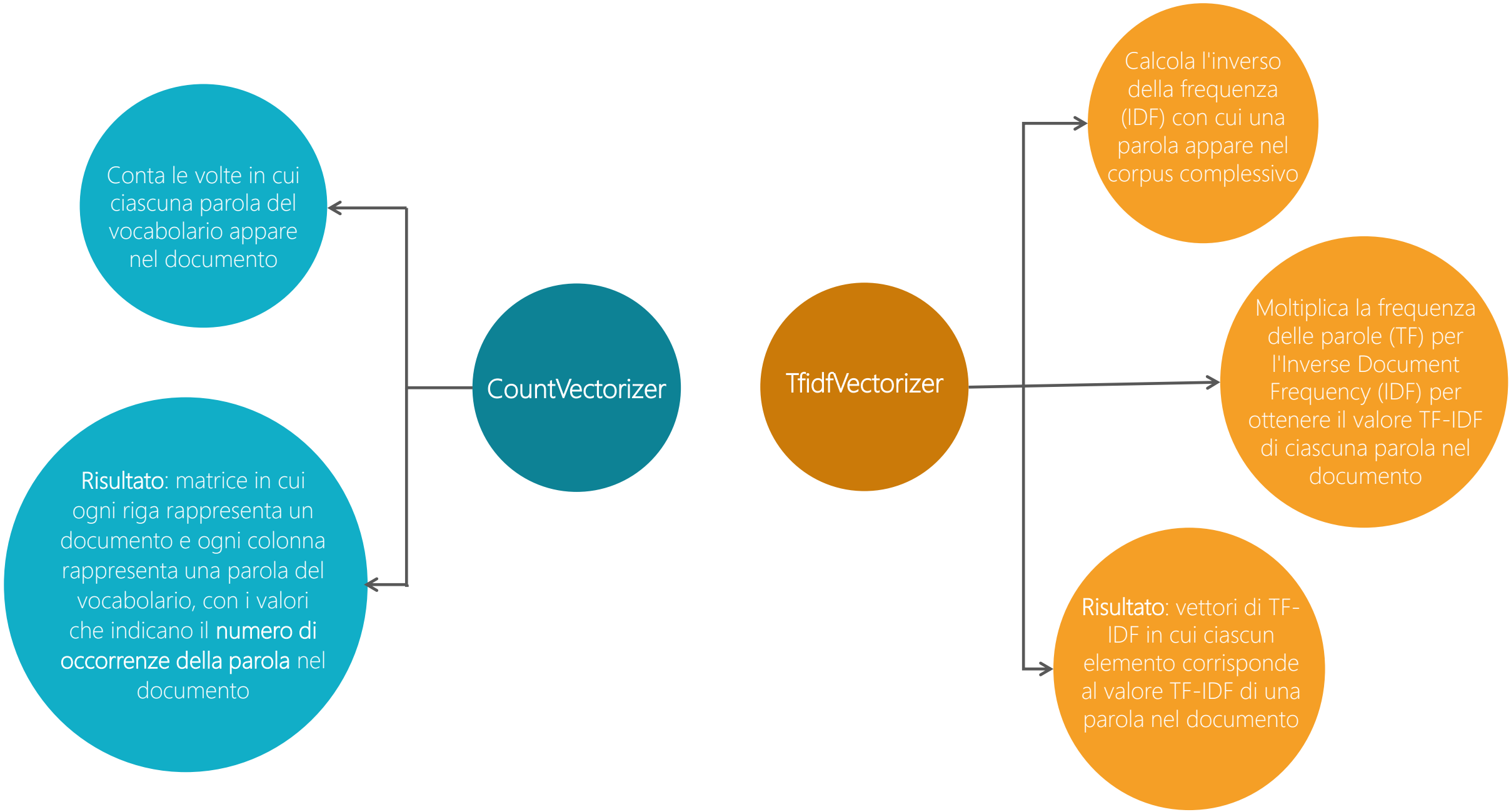
Index	Text
0	If there s one thing this election succeeded i...
1	MADRID (Reuters) - A Spanish audit office has ...
2	WASHINGTON (Reuters) - Some prominent Republic...
3	Who s laughing now funny guy?We asked everyo...
4	Wow! This young Asian student nails it! He spe...

AFTER

Index	Parola 1	Parola 2	Parola 3	Parola 4	Parola 5	...
0	one	thing	election	succeeded	bringing	...
1	madrid	reuters	spanish	audit	office	...
2	washington	reuters	prominent	republicans	...	...
3	laughing	funny	guywe	asked	everyone	...
4	wow	young	asian	student	nails	...



Vettorizzazione



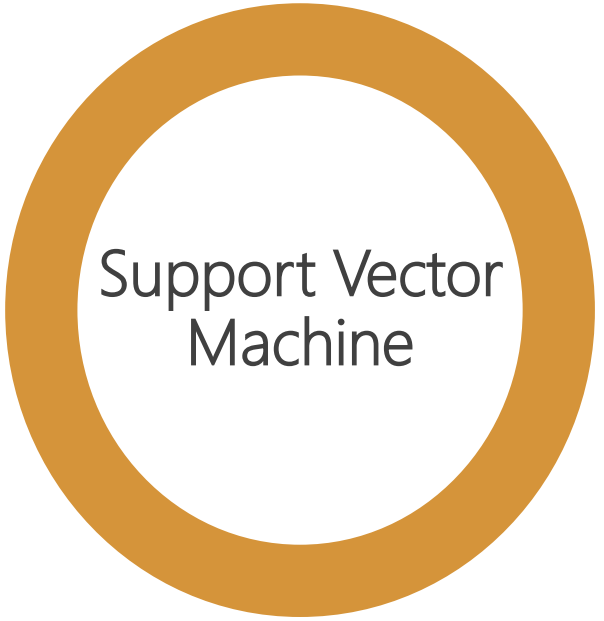
Addestramento

Modelli utilizzati:

Suddivide ricorsivamente l'insieme di dati in sottoinsiemi più omogenei rispetto alla variabile target, fino a raggiungere una decisione finale su quale classe attribuire a una specifica istanza.

Stima la probabilità che un'istanza appartenga ad una determinata classe utilizzando il teorema di Bayes, considerando le feature indipendenti e scegliendo la classe con la probabilità condizionata massima.

Definisce un iperpiano ottimale nel contesto di uno spazio ad alta dimensione, che massimizza il margine tra le diverse classi di dati. Questo iperpiano viene utilizzato per classificare nuove istanze di dati in base alla loro posizione rispetto alle diverse classi.



Classificatore Bayesiano

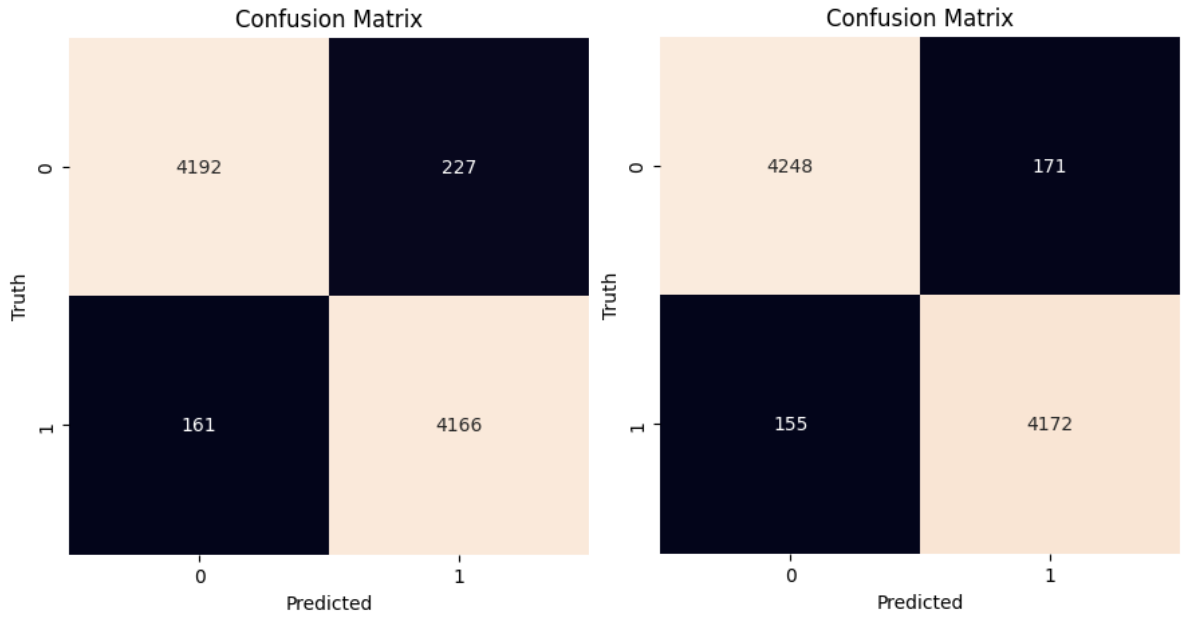
Performance evaluation with cv	
Accuracy	0.9556368625657443
Precision	0.9483268836785795
Recall	0.9627917725907095
F1	0.9555045871559633
ROC AUC	0.9557113422808718
Specificità	0.9486309119710342
Precision negativa:	0.9630140133241443

Performance evaluation with tfidf	
Accuracy	0.9392865309855934
Precision	0.9305807622504537
Recall	0.9480009244280102
F1	0.93921007441328
ROC AUC	0.9393772442913981
Specificità	0.9307535641547862
Precision negativa:	0.9481327800829875

Nella scelta del modello per la classificazione delle fake news, si è scelto il **Naive Bayes** principalmente per la sua **comprovata efficacia nell'elaborazione del testo**. Il Naive Bayes, sebbene di semplice implementazione, si distingue per la sua potenza nel trattare grandi volumi di dati testuali, rendendolo una scelta ideale per le applicazioni di elaborazione del linguaggio naturale.

Performance evaluation post tuning with tfidf	
Accuracy	0.957809284244226
Precision	0.9770973963355835
Recall	0.9366766813034435
F1	0.9564601769911505
ROC AUC	0.9575893024077752
Specificità	0.9785019235121069
Precision negativa:	0.9404088734232275

Dall'analisi della matrice di confusione emerge un miglioramento delle prestazioni del modello:



Confusion matrix con cv

Confusion matrix con cv dopo il tuning degli iperparametri

I tempi di calcolo sono stati molto brevi, con un'efficienza notevole riscontrata soprattutto nell'utilizzo di CountVectorizer:

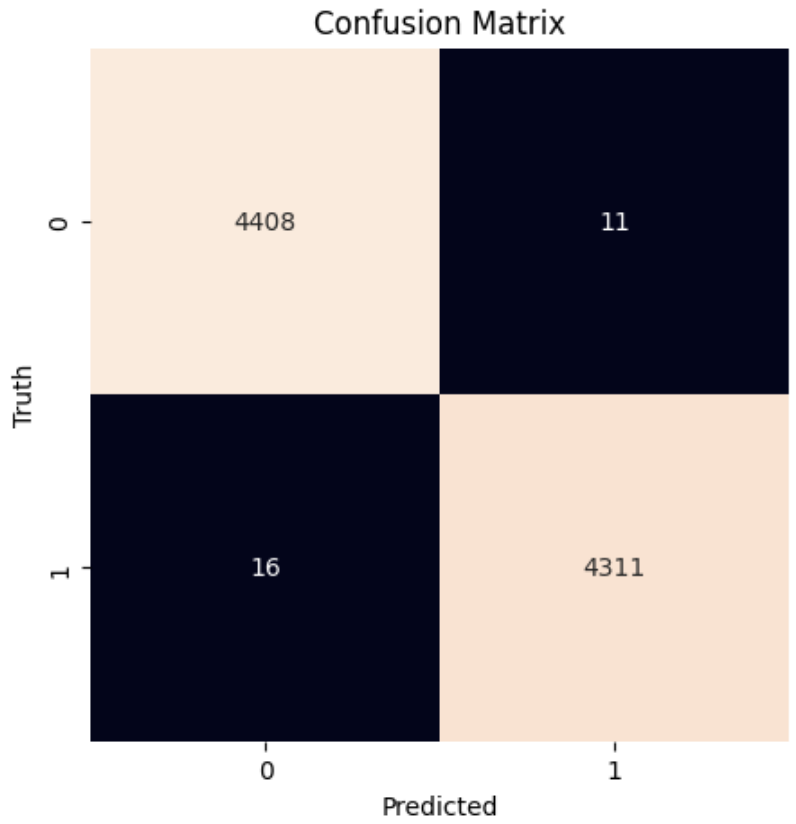
	tempo			
	Training	Predizione	Ricerca del Modello Migliore	Predizione del Modello Migliore
MultinomialNB con CV	0.036 secondi	0.01 secondi	2.323 secondi	0.011 secondi
MultinomialNB con TF-IDF	0.042 secondi	0.011 secondi	2.825 secondi	0.012 secondi

Per condurre la ricerca degli iperparametri in modo efficiente ed efficace, si è optato per l'utilizzo del metodo **grid** poiché, nonostante la notevole quantità di dati a disposizione, questo approccio ha dimostrato di comportarsi in modo robusto ed efficace, garantendo una copertura completa dello spazio degli iperparametri.

Decision Tree Classifier

La scelta del modello **Decision Tree** è stata guidata dalla sua natura esplicativa e dalla capacità di gestire dati non lineari, fornendo un approccio maggiormente interpretabile rispetto agli ali altri classificatori.

Performance evaluation with cv	
Accuracy	0.9969128744568946
Precision	0.9974548819990745
Recall	0.9963022879593252
F1	0.9968782518210197
ROC AUC	0.9969065184987846
Specificità	0.997510749038244
Precision negativa:	0.9963833634719711

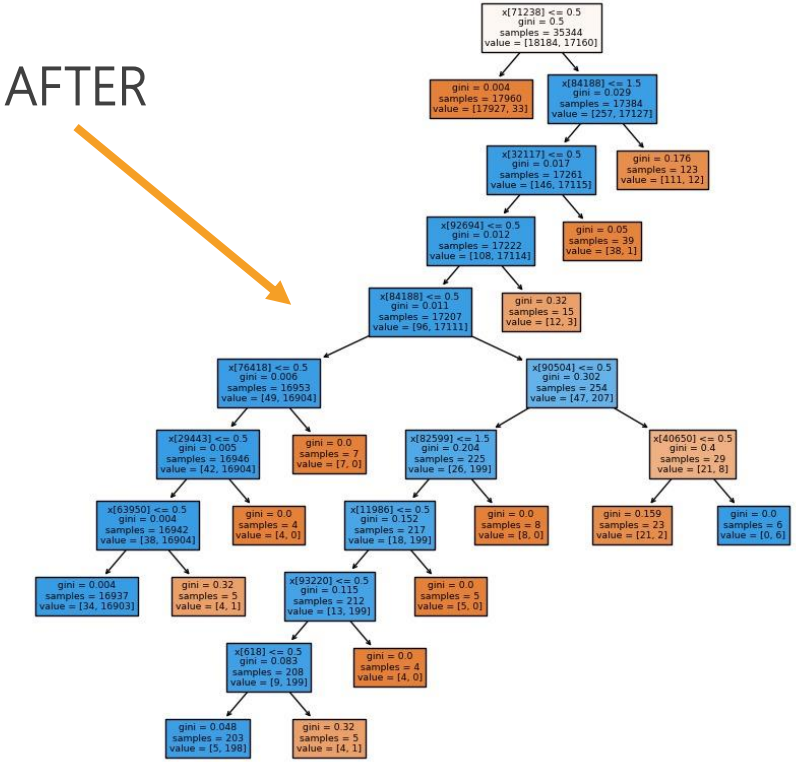


Confusion matrix con cv

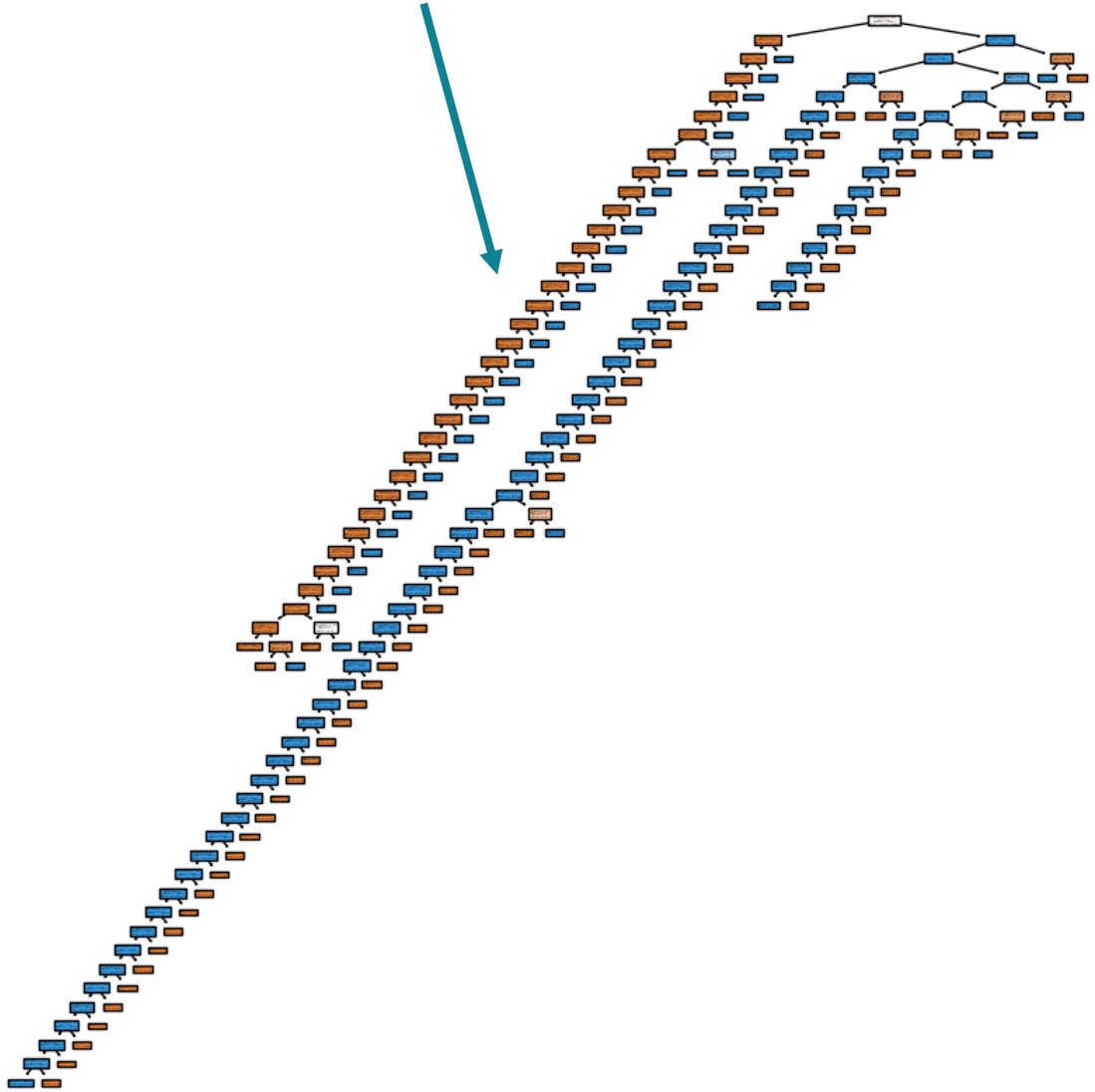
Le performance ottenute con tfidf sono le medesime (fino alla quarta cifra decimale).

Il modello mostra prestazioni eccezionali su diverse metriche di valutazione, indicando una classificazione accurata dei casi positivi e negativi. Tuttavia, la complessità dell'albero decisionale **solleva dubbi sull'overfitting** e la capacità di generalizzazione, suggerendo la necessità di valutare attentamente la sua complessità e considerare strategie di semplificazione come la potatura o l'ottimizzazione degli iperparametri.

AFTER



BEFORE



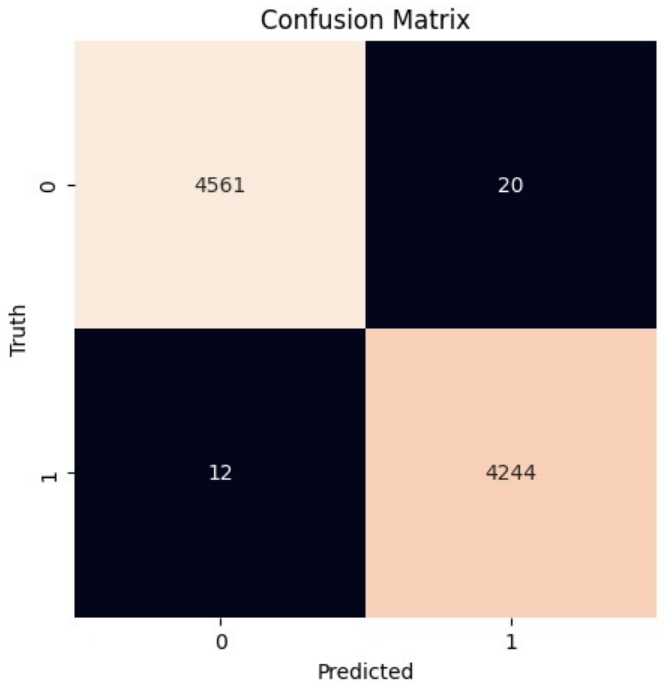
Per DecisionTree con **CV**, l'addestramento è efficiente e veloce, impiegando **8.759 secondi**, mentre il tempo di predizione è minimo, solo **0.1 secondi**, indicando un'istantaneità nelle predizioni una volta addestrato.

Per DecisionTree con **TF-IDF**, l'addestramento richiede molto più tempo, **23.029 minuti**, suggerendo una complessità maggiore rispetto al CountVectorizer. Tuttavia, il tempo di predizione rimane breve, solo **0.022 secondi**, indicando ancora una rapida capacità di predizione dopo l'addestramento.

	tempo			
	Training	Predizione	Ricerca del Modello Migliore	Predizione del Modello Migliore
DecisionTree con CountVectorizer	8.759 secondi	0.12 secondi	23.029 minuti	0.012 secondi
DecisionTree con TFIDF	24.504 secondi	0.022 secondi	60.05 minuti	0.03 secondi

SVM

Il modello di **Support Vector Machines** (SVM) è stato scelto per la sua capacità di gestire spazi ad alta dimensionalità e dati non lineari, offrendo flessibilità anche in scenari con un elevato numero di dimensioni rispetto ai campioni.



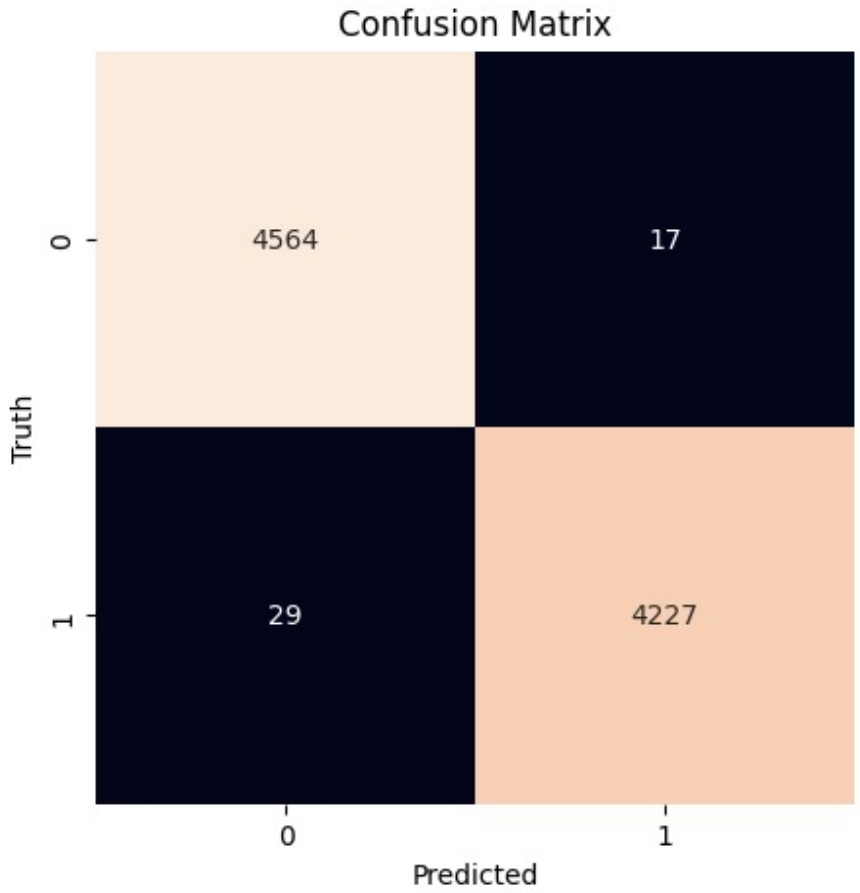
Confusion matrix con cv (kernel lineare)

Performance evaluation with cv (kernel lineare)	
Accuracy	0.996378861604617
Precision	0.9953095684803002
Recall	0.9971804511278195F1
F1	0.996244131455399ROC
ROC AUC	0.9964072960725323
Specificità	0.9956341410172451

	tempo			
	Training	Predizione	Ricerca del Modello Migliore	Predizione del Modello Migliore
SVM con CountVectorizer	131.336 secondi	15.531 secondi	1838.015 secondi	90.705 secondi
SVM con TFIDF	586.929 secondi	49.895 secondi	3015.806 secondi	115.89 secondi

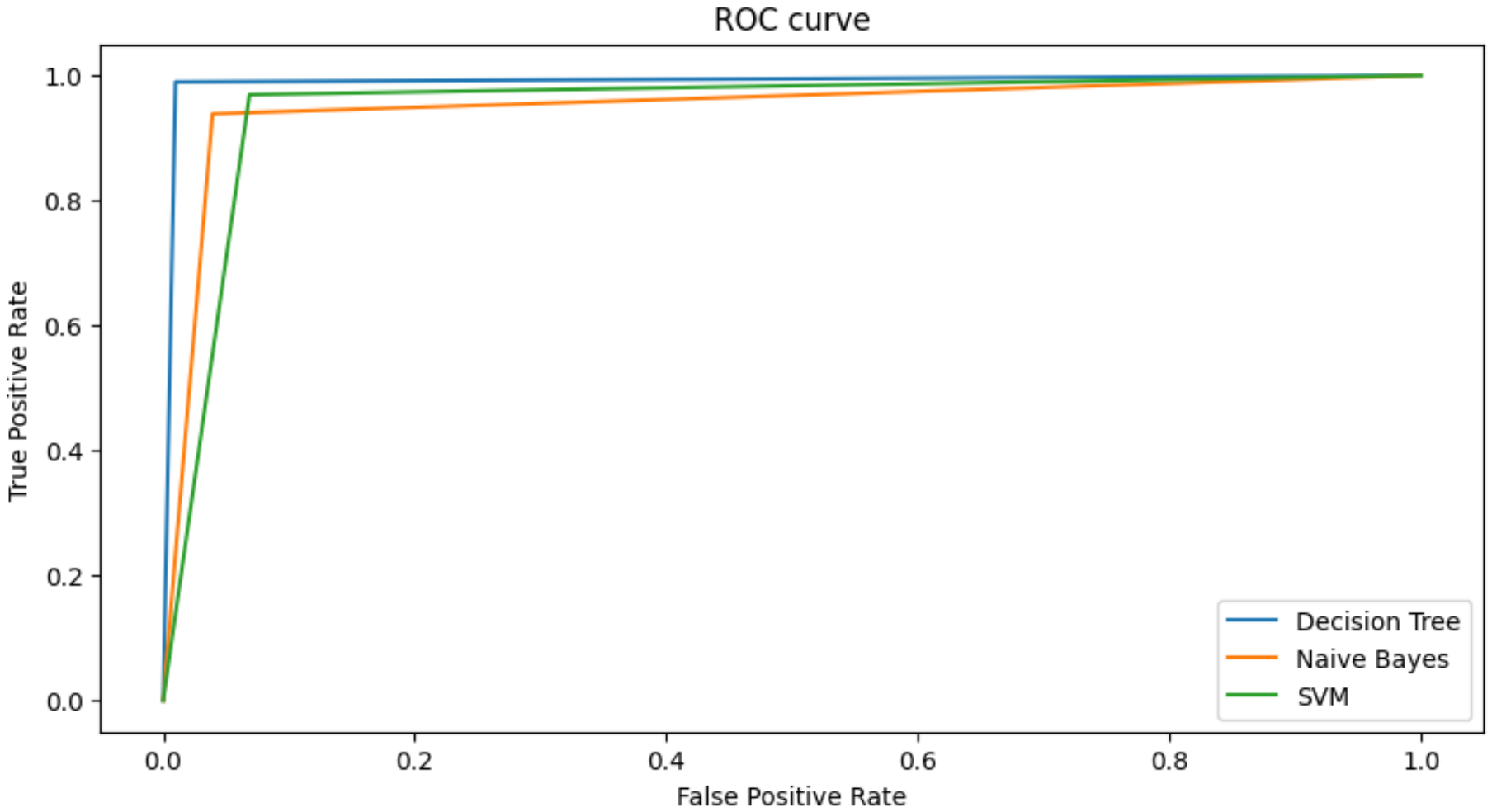
È stato scelto di ottimizzare i parametri utilizzando la **Grid Search** con una tecnica **k-fold**, suddividendo il set di allenamento in tre sottoinsiemi separati per trovare la migliore combinazione di iper-parametri.
Questo approccio ha permesso di massimizzare l'efficacia del classificatore nel distinguere tra notizie vere e false. La ricerca si è focalizzata solo sul kernel **rbf**, tuttavia, i risultati indicano che un kernel lineare riesce a separare le notizie con maggiore precisione e velocità.

Performance evaluation with cv	
Accuracy	0.9947946135566369
Precision	0.9959943449575872
Recall	0.9931860902255639F1
F1	0.9945882352941177
ROC AUC	0.9947375550451112
Specificità	0.9956341410172451



Confusion matrix con cv post tuning (kernel RBF)

Performance Evaluation



AUC DECISION TREE

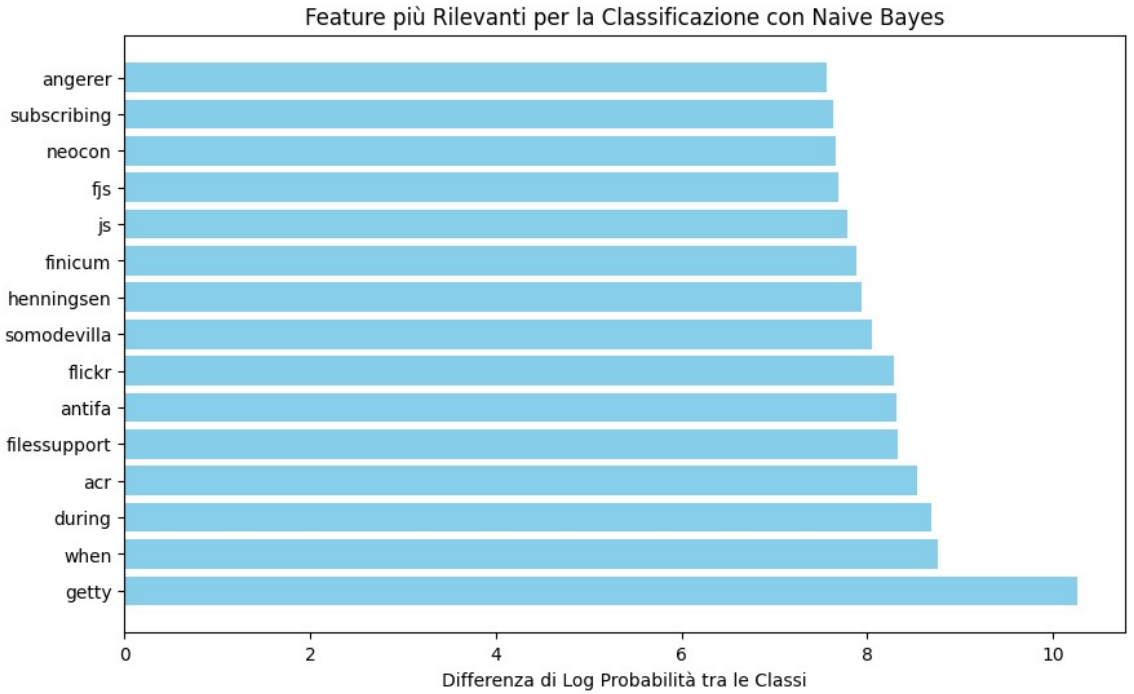
0.9899

AUC NAIVE BAYES

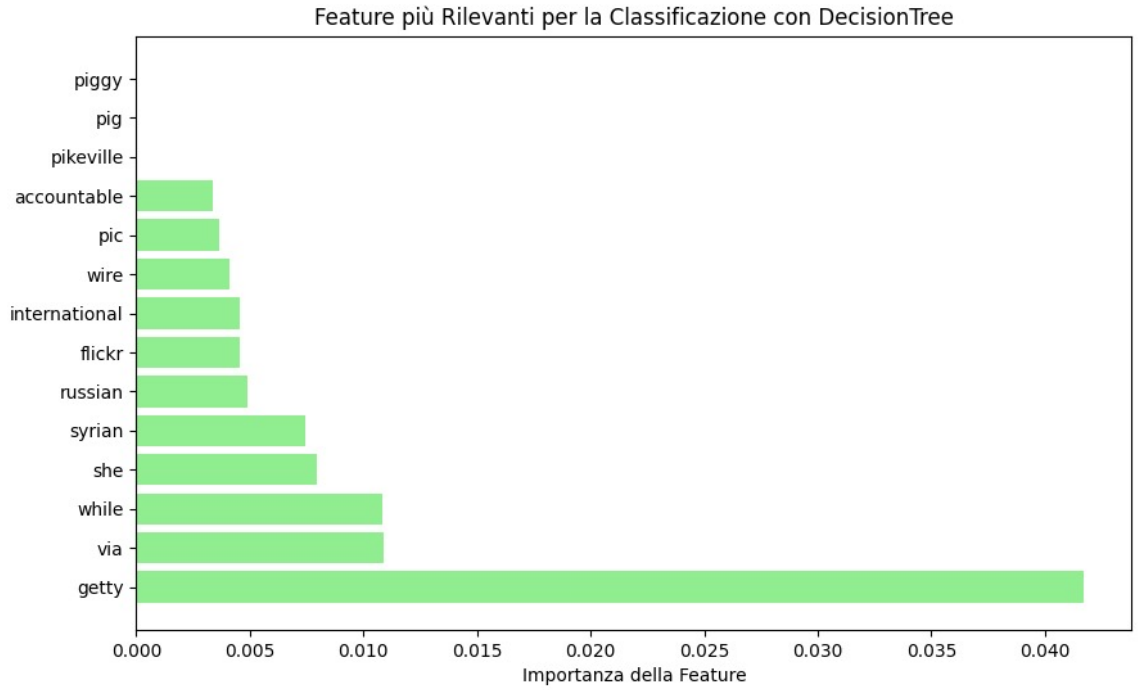
0.9497

AUC SVM

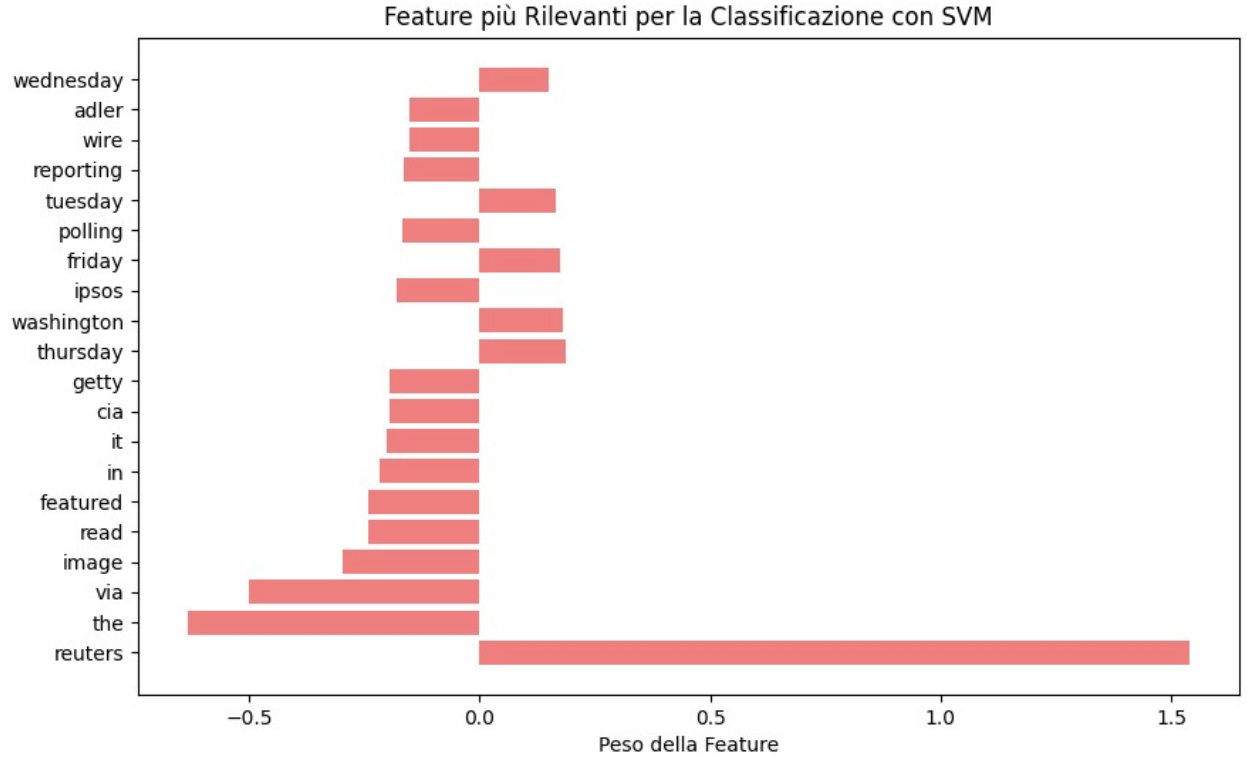
0.9503



Ogni barra rappresenta una feature specifica, mentre la lunghezza della barra riflette il peso della feature nella classificazione del testo.



Ogni barra rappresenta una feature specifica, mentre la lunghezza della barra riflette l'importanza relativa della feature nella presa di decisioni del modello



Le feature sono riportate sull'asse Y, mentre sull'asse X è rappresentato il peso associato a ciascuna feature. Le barre più lunghe indicano feature che hanno un maggiore impatto nella capacità del modello di discriminare tra le classi.

Ogni feature contribuisce in modo proporzionale al suo peso nel calcolo della funzione di decisione lineare. L'output di questa funzione viene poi utilizzato per determinare la classe predetta.

External Validation

I modelli migliori prodotti precedentemente ottengono buone prestazioni anche su dataset differenti.
Qui di seguito un esempio:

DECISION TREE	
Accuracy	0.8114737482262315
Precision	0.9827700463883366
Recall	0.6212819438625891
F1	0.7612936344969198

SVM	
Accuracy	0.8467464017839044
Precision	0.9771796372147454
Recall	0.6996229576874738F1
F1	0.8154296875

NAIVE BAYES	
Accuracy	0.8295155078045814
Precision	0.8838133068520357
Recall	0.7457059069962296
F1	0.8089070665757782

Sviluppi futuri

1

Ottimizzazione del pre-processing: È proposta l'analisi approfondita delle parole per individuare e rimuovere quelle non rilevanti, migliorando l'efficienza computazionale.

Stratificazione dei vettorizzatori: Si suggerisce di bilanciare l'importanza dei vettorizzatori come word2vec, GloVe e BERT per migliorare la completezza della rappresentazione del testo.

2

3

Clustering semantico con word2vec: L'implementazione di clustering non supervisionato basato sul significato semantico delle frasi potrebbe rivelare pattern nascosti nel dataset, arricchendo la comprensione delle relazioni tra le notizie.

Grazie