

FAKE NEWS



DEDECTION

BIG DATA ENGINEERING & SOLUTIONS ARCHITECT

INTRODUZIONE

 **NOME E COGNOME:** DAVIDE PEDRETTI

 **INTERESSI:** DATA SCIENCE, MACHINE LEARNING, LINGUISTICA COMPUTAZIONALE, SOFTWARE ENGINEERING

 **COMPETENZE APPLICATE NEL PROGETTO:** PYTHON, MONGO DB, ALGORITMI DI ML,NLP, TEAM WORKING, PENSIERO CRITICO

 **MOTIVAZIONE:** LE FAKE NEWS OGGI SONO UNA SFIDA APERTA E UN TEMA CHE MI APPASSIONA.

METTERE LE MIE COMPETENZE TECNICHE AL SERVIZIO DI QUESTO PROBLEMA MI È SEMBRATA UNA SCELTA NATURALE.

LA MOTIVAZIONE È NATA ANCHE VEDENDO I LAVORI DI THOMAS MIKOLOV SUGLI ALGORITMI PER TRASFORMARE LE PAROLE IN VETTORI.

INDICE DEI CONTENUTI

CONTESTO & OBIETTIVI

FONTE DATI

ARCHITETTURA

PRE-PROCESSING DEL TESTO

TECNICHE DI WORD EMBEDDING

ALGORITMI DI MACHINE LEARNING

METRICHE DI VALUTAZIONE

RISULTATI

CONCLUSIONI

I PROSSIMI PASSI

RINGRAZIAMENTI

CONTESTO E OBIETTIVO



CONTESTO ATTUALE

La diffusione capillare delle fake news sui social media ha reso necessaria l'elaborazione di metodi di rilevamento sempre più sofisticati per preservare l'integrità dell'informazione

OBIETTIVO

L'obiettivo di questo lavoro è valutare l'efficacia di diverse tecniche di word embedding applicate a modelli di machine learning e deep learning per la classificazione delle fake news. Attraverso l'analisi comparativa dei risultati, il progetto mira a identificare le metodologie più performanti e a costruire una pipeline modulare e replicabile per il rilevamento automatico delle notizie false

FONTE DATI



COSA E' KAGGLE?

Kaggle è una piattaforma online che serve come comunità globale per appassionati e professionisti di data science e machine learning. Offre strumenti per partecipare a competizioni, condividere e utilizzare dataset, creare modelli e collaborare tramite notebook

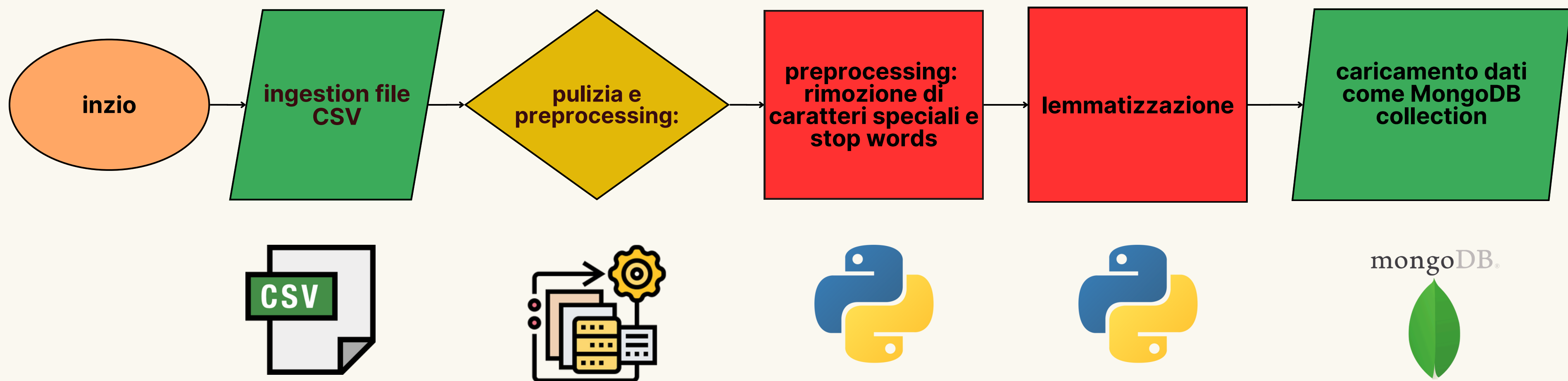
FAKE AND REAL NEWS DATASET

Il dataset "Fake and Real News Dataset" raccoglie articoli di notizie etichettati come "fake" o "real", ed è strutturato tipicamente con colonne per titolo, testo e label che indica se la notizia è vera o falsa

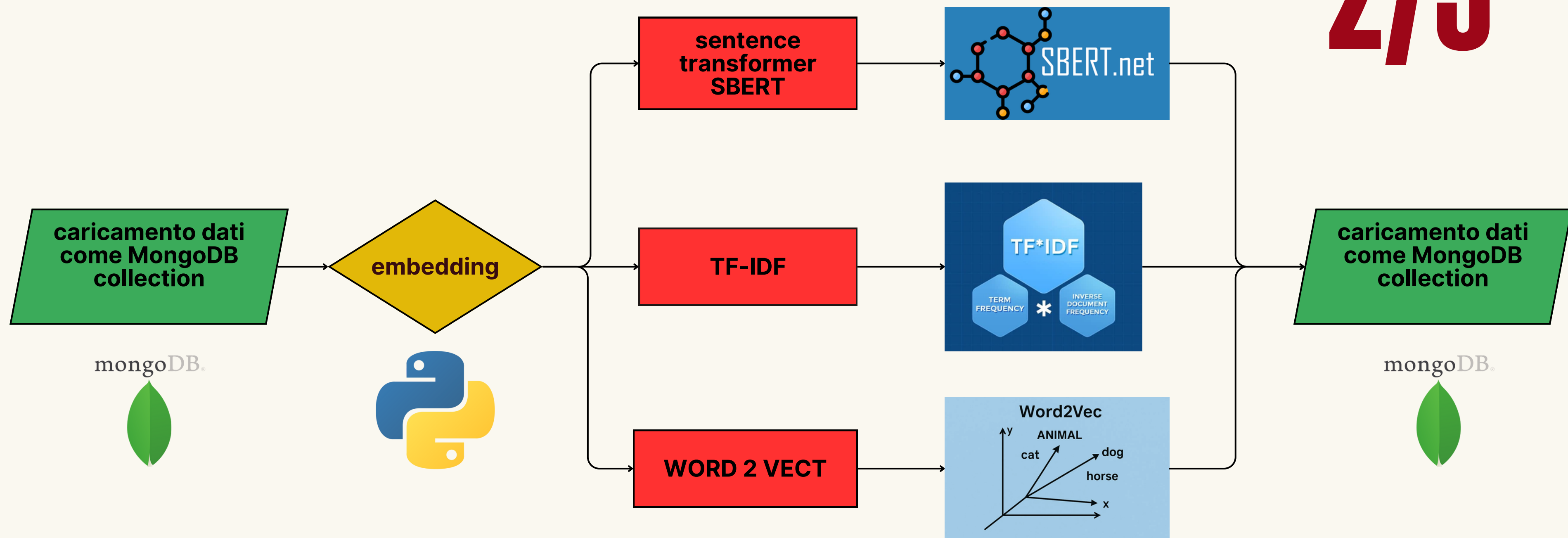
kaggleTM

ARCHITETTURA

1/3



2/3



3/3

caricamento dati
come MongoDB
collection

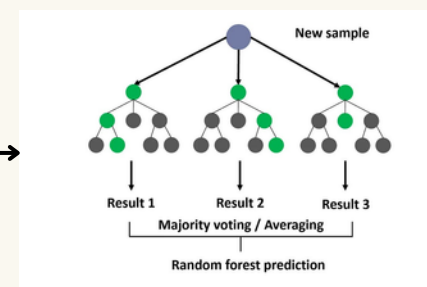
mongoDB.



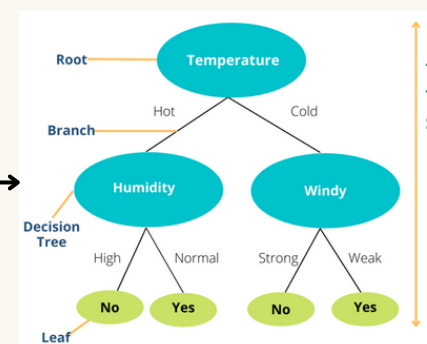
algoritmi di
classificazione



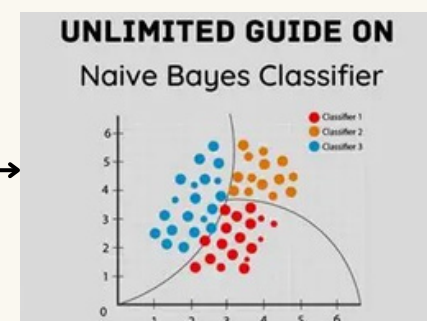
Random forest



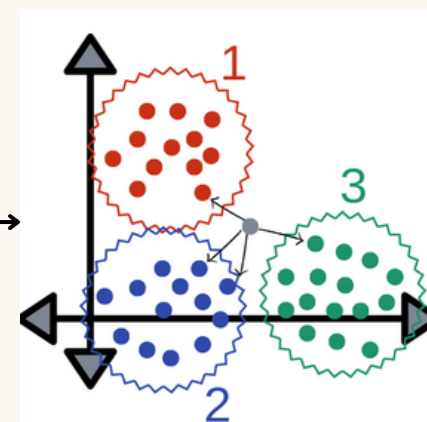
Decision Tree



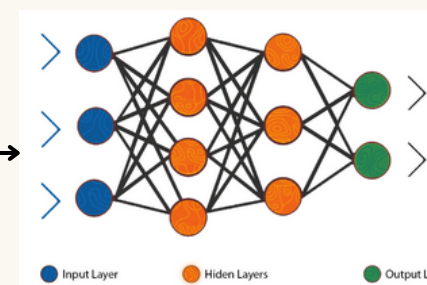
Naive Bayes



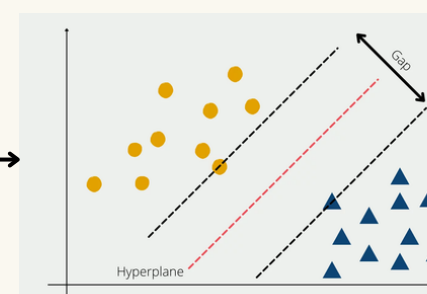
KNN



MLP



SVM



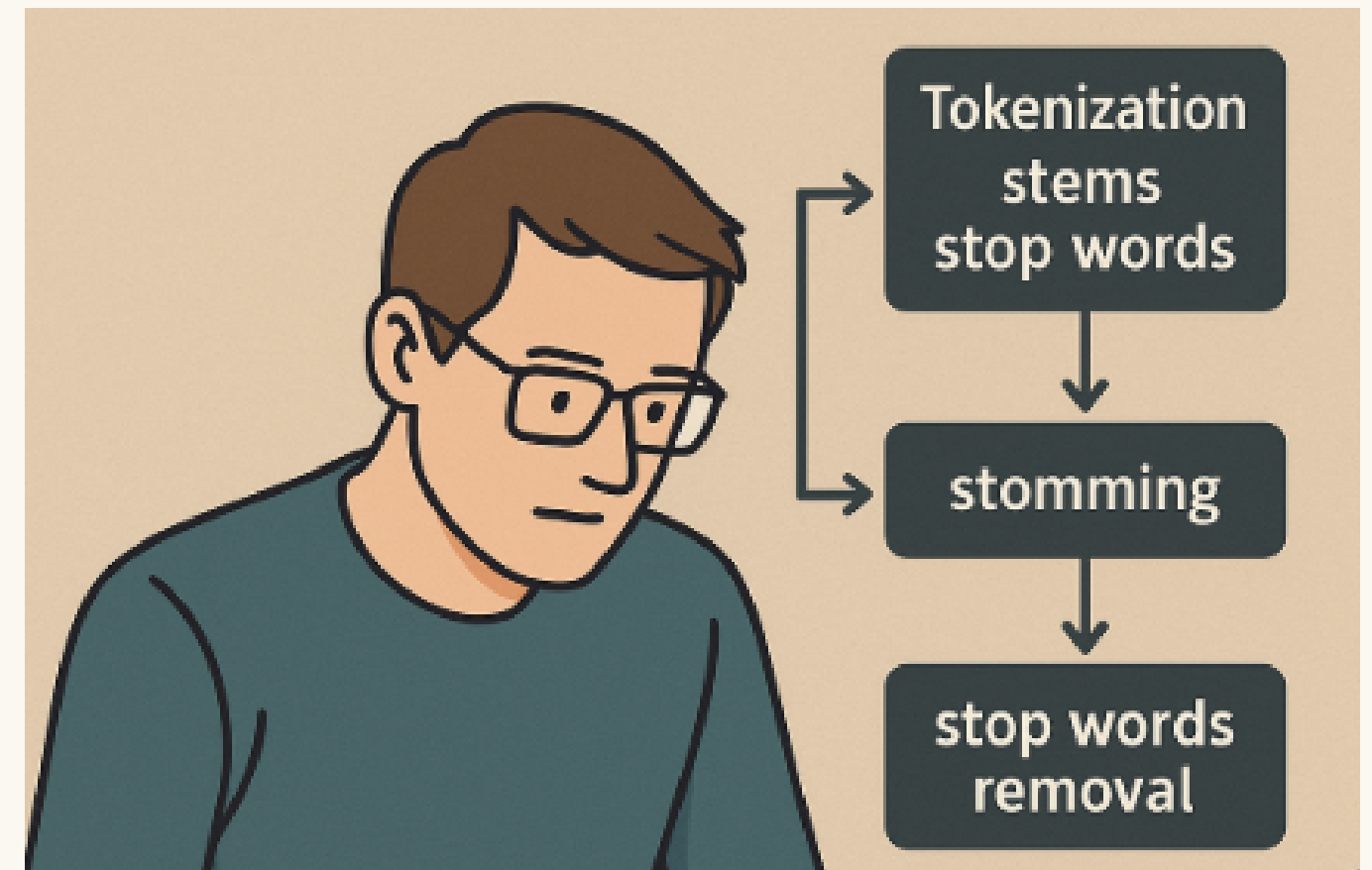
Generazione file
CSV con
metriche



Creazione
report Streamlit

Streamlit

TEXT PRE PROCESSING



IMPORT DATA

IMPORTDATA.PY GESTISCE LA CONNESSIONE A MONGODB E L'IMPORTAZIONE DEL DATASET IN FORMATO CSV, OCCUPANDOSI ANCHE DELLA PULIZIA PRELIMINARE DEI DATI. IN QUESTA FASE VENGONO ELIMINATI CAMPI NON VALIDI, GESTITI I VALORI MANCANTI E CREATI INDICI PER OTTIMIZZARE LE QUERY

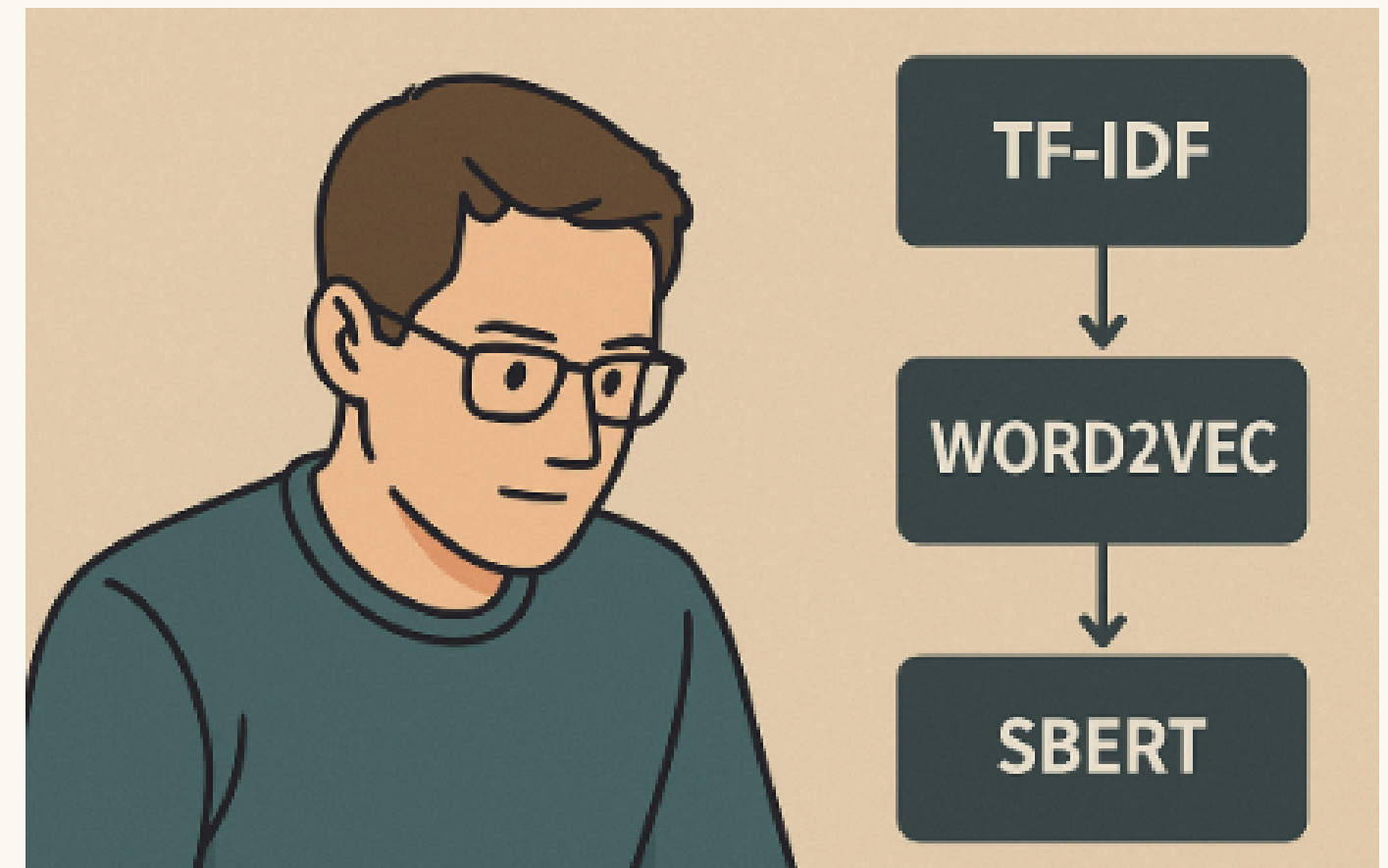
PRE PROCESSING

PREPROCESSING.PY IMPLEMENTA LE OPERAZIONI DI PULIZIA E NORMALIZZAZIONE DEL TESTO, COME LA RIMOZIONE DI URL, HASHTAG, CARATTERI SPECIALI E LA STANDARDIZZAZIONE IN MINUSCOLO. SUCCESSIVAMENTE APPLICA TOKENIZZAZIONE E FILTRAGGIO DELLE STOPWORDS, PREPARANDO I DATI PER L'EMBEDDING

LEMMATIZZAZIONE

LA LEMMATIZZAZIONE RIDUCE OGNI PAROLA ALLA SUA FORMA BASE TRAMITE POS TAGGING, ELIMINANDO LE VARIANTI MORFOLOGICHE. QUESTO PROCESSO PERMETTE DI DIMINUIRE L'AMBIGUITÀ LINGUISTICA E DI RENDERE IL TESTO PIÙ UNIFORME PER L'ANALISI

WORD EMBEDDING



TF - IDF

LA TF-IDF MISURA L'IMPORTANZA DI UNA PAROLA COMBINANDO FREQUENZA NEL DOCUMENTO E RARITÀ NEL CORPUS, EVIDENZIANDO I TERMINI PIÙ RAPPRESENTATIVI. È UNA TECNICA SEMPLICE MA MOLTO EFFICACE PER LA CLASSIFICAZIONE TESTUALE

WORD 2 VECT

WORD2VEC È UNA TECNICA DI EMBEDDING CHE UTILIZZA RETI NEURALI PER APPRENDERE RAPPRESENTAZIONI VETTORIALI DENSE DELLE PAROLE. SI BASA SU DUE ARCHITETTURE PRINCIPALI, CBOW E SKIP-GRAM, CHE PERMETTONO DI COGLIERE RELAZIONI SEMANTICHE E SINTATTICHE TRA TERMINI

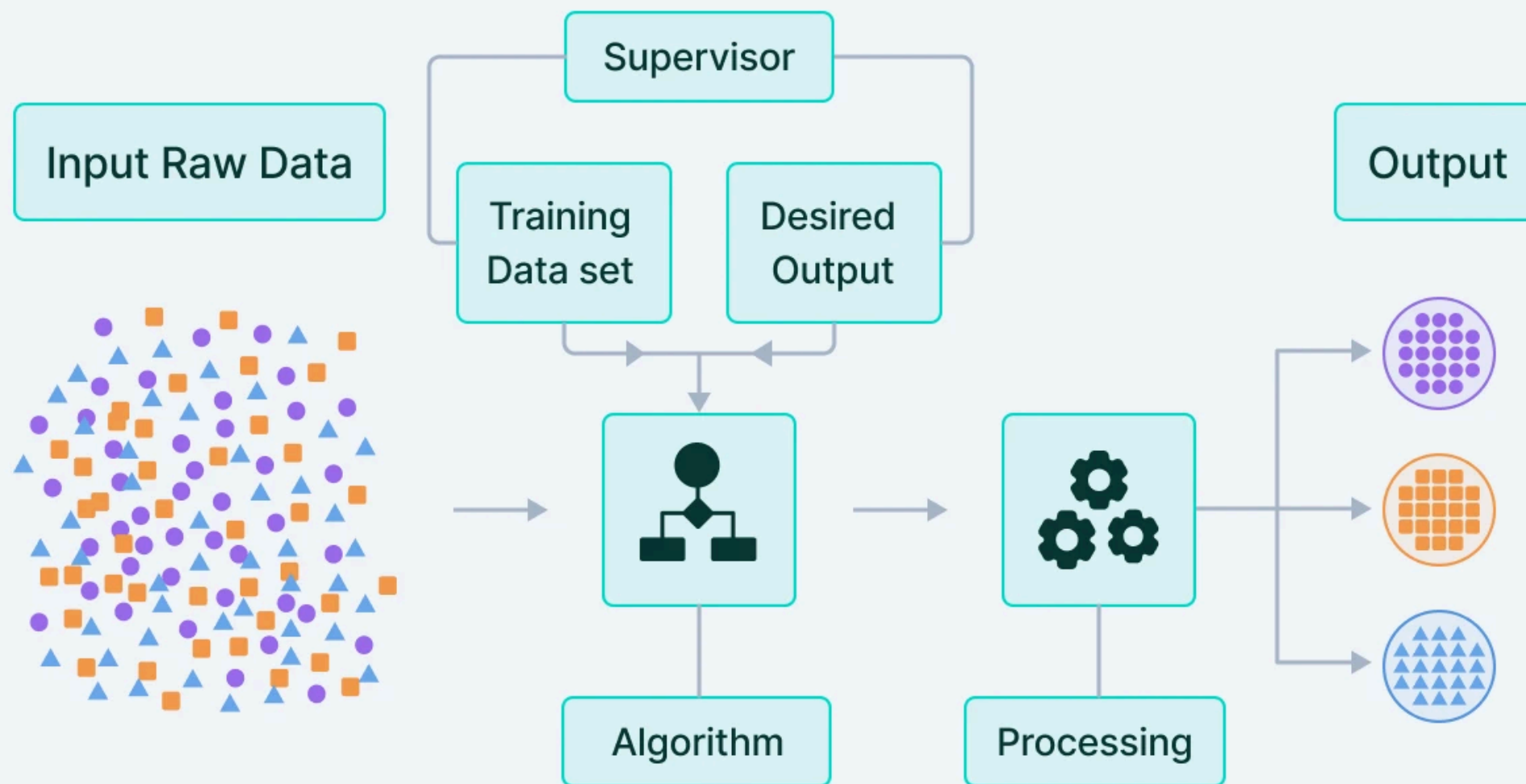
TRANSFORMERS

I TRANSFORMERS INTRODUCONO EMBEDDING DINAMICI CHE CONSIDERANO IL CONTESTO IN CUI LE PAROLE COMPAIONO. QUESTO APPROCCIO CONSENTE DI CATTURARE MEGLIO LE DIPENDENZE TRA TERMINI E LE SFUMATURE SEMANTICHE NEI TESTI.

MACHINE LEARNING & NEURAL NETWORK



Supervised Learning



ALGORITMI DI ML TRADIZIONALI

Gli algoritmi tradizionali di machine learning, come Naive Bayes, Logistic Regression, KNN, SVM, Decision Tree e Random Forest, sono stati utilizzati per affrontare la classificazione delle fake news.

RETE NEURALE (MLP)

Il Multilayer Perceptron (MLP) rappresenta invece un modello di rete neurale in grado di catturare relazioni complesse tra le feature grazie a strati multipli e funzioni di attivazione non lineari. Pur richiedendo maggiori risorse computazionali, si rivela adatto a compiti che necessitano di una modellazione più sofisticata.

VALUTAZIONE DEL MODELLO



ACCURACY

Accuracy misura la proporzione complessiva di previsioni corrette sul totale delle istanze analizzate. È un indicatore generale delle prestazioni, ma può risultare poco rappresentativo in presenza di dataset sbilanciati

PRECISION

Precision valuta la capacità del modello di ridurre i falsi positivi, indicando quante delle notizie classificate come false lo siano realmente. È cruciale in contesti in cui i falsi allarmi hanno un costo elevato.

RECALL

Recall rappresenta la capacità del modello di individuare tutte le notizie false presenti nel dataset. Un valore elevato di recall riduce il rischio che fake news sfuggano al sistema di rilevamento

F1 - SCORE

F1-score è la media armonica tra precision e recall, fornendo una misura bilanciata delle performance. Questo parametro è particolarmente utile quando occorre mantenere un equilibrio tra sensibilità e accuratezza del modello

RISULTATI

Il confronto sperimentale mostra che TF-IDF ottiene le performance più elevate, in particolare con Linear SVM e Logistic Regression, mentre Word2Vec risulta meno efficace senza fine-tuning. Gli embedding basati su Transformers (SBERT) garantiscono buoni risultati e maggiore capacità di generalizzazione, pur senza superare TF-IDF

Embedding	Accuracy media	F1 media
TF-IDF	0.929	0.936
Word2Vec	0.944	0.944
SBERT (all-MiniLM-L6-v2)	0.920	0.923

CONCLUSIONI

Il progetto ha evidenziato come diverse tecniche di word embedding influenzino in modo significativo le prestazioni dei modelli di machine learning nella rilevazione delle fake news. In particolare, TF-IDF si è dimostrato molto efficace in scenari supervisionati con dataset ampi e bilanciati

Allo stesso tempo, Word2Vec e SBERT hanno mostrato una maggiore capacità di cogliere relazioni semantiche e contestuali, rendendosi utili in scenari che richiedono generalizzazione. La pipeline sviluppata, supportata da una dashboard interattiva, ha confermato l'efficacia di un approccio modulare e replicabile



I PROSSIMI PASSI . . .



❖ RETI CONVOLUZIONALI (CNN)

L'integrazione di CNN potrebbe migliorare la capacità di rilevare pattern complessi e relazioni non lineari nei testi. Questo approccio permetterebbe di catturare caratteristiche latenti utili per distinguere in modo più accurato tra notizie reali e fake news.

❖ FONTI MULTIMODALI

L'estensione del sistema ad altre tipologie di contenuti, come immagini manipolate o video, aprirebbe la strada a un rilevamento più completo. L'obiettivo è sviluppare un modello multimodale capace di analizzare simultaneamente testi, contenuti visivi e dati social.



THANK YOU



So Much