# MASTER DI I LIVELLO IN
# ANALISI DATI PER LA BUSINESS INTELLIGENCE E DATA SCIENCE

Candidato: Dott. Gianpiero Drigo

Relatore: Prof. Daniele Paolo Radicioni

Tutor aziendale: dott. Enrico Mensa

## IL CONTRIBUTO DELLE TECNICHE DI NLP NELLA CATEGORIZZAZIONE DI DOCUMENTI TESTUALI: ANALISI DI DUE CASI D'USO, OPEN DOMAIN E DOMAIN-SPECIFIC

## ABSTRACT

Il lavoro rappresenta una sintesi degli argomenti trattati nel corso del tirocinio svolto sotto la supervisione del Dipartimento di Informatica dell'Università di Torino.

Quale argomento principale, si è cercato di evidenziare se, e nel caso quanto, l'applicazione delle principali tecniche di NLP - Data Preprocessing incidesse sul rendimento di classificazione di documenti testuali relativi a contesti differenti: un open domain "classico" per le attività NLP (dataset 20_NEWSGROUP) ed un dominio specifico rappresentato da una raccolta di dati infortunistici pubblicati dall'ente americano di sicurezza nei luoghi di lavoro OSHA (dataset OSHA_INJURY). È stato inoltre testato un terzo contesto ibrido, formato dall'unione parziale dei due dataset per record assimilabili in quanto a tematica trattata.

Le valutazioni sono state effettuate testando il rendimento di un classificatore Naive Bayes con approccio Tf-idf in tre differenti situazioni:

- condizione iniziale senza applicazione di tecniche di preprocess
- condizione intermedia con applicazione di tecniche di preprocess
- condizione ottimale con applicazione di tecniche di preprocess e parametri ottimizzati

Sono state effettuate classificazioni multi target e single target, oltre a classificazioni binarie per campi specifici del dataset OSHA, operando su configurazioni dei dataset differenti per contenuto testuale.

I risultati ottenuti sono stati valutati sia in termini di metriche (accuratezza con cross fold validation; precision; recall; F-1 score) che di matrici di confusione.

Secondariamente, sono state ricercate eventuali differenze, in termini di performance computazionali o di rendimento, nelle diverse modalità di preparazione dei dati, testando l'applicazione delle tecniche di pre-

process direttamente tramite il modulo di tokenizzazione ed esternamente tramite utilizzo di strumenti di gestione dataframe Pandas e libreria NLTK.

Infine si è cercato di valutare se l'utilizzo associato di varie tecniche di pre-process in modo sistematico comporti incrementi maggiori nel risultato di classificazione rispetto all'applicazione singola e sequenziale.

La base tecnica del lavoro è stata il linguaggio Python con le librerie specifiche NLTK (Natural Language Tool Kit); Pandas e SciKit (SKLEARN). Per quanto inerente al classificatore, si è optato per il Naive Bayes Classifier (MultinomialNB).

I risultati ottenuti hanno messo in evidenza un impatto positivo delle tecniche di data preprocessing sul rendimento di classificazione di documenti testuali, condizionato però in modo importante dal dataset in uso: si sono riscontrate infatti forti differenze tra 20_Newsgroup e Osha_injury. Le problematiche relative ai dataset emerse durante lo svolgimento dei test sono infine state oggetto di approfondimento per cercare di evidenziarne le cause.

Al fine di rendere più semplice la lettura del lavoro, i risultati sono presentati e commentati in modo descrittivo, con richiami a grafici riassuntivi, metriche e matrici di confusione relativi ai singoli test eseguiti.