Università Magna Graecia di Catanzaro

Corso di Laurea Magistrale in Ingegneria Biomedica

Infrastrutture di calcolo e algoritmi efficienti per la biologia e medicina

Progetti didattici A.A. 2023/2024

Prof. Mario Cannataro,

Dott. Chiara Zucco

Text Analytics 2

Obiettivo del progetto è effettuare un tipico processo di Text Analysis, utilizzando tecniche statistiche congiunte ad algoritmi di NLP e Machine Learning per compiere un task di classificazione del dataset allegato relativo ad abstract di articoli scientifici scritti in lingua inglese e contenuti nel dataset denominato **“Coronavirus tweets NLP - Text Classification "**. Il dataset contiene tweet sul coronavirus annotati in base al sentiment espresso nel testo.

In particolare, ogni gruppo di candidati deve:

- Effettuare una descrizione introduttiva del dataset ed un’analisi esplorativa del dataset anche mediante grafici, dimostrando di:

· avere familiarità con il dataset stesso,

· conoscere la distribuzione delle varie classi

. provare ad identificare le 10 parole più comuni nel dataset.

- Effettuare, se opportuno, una fase di Text Cleaning e Feature extraction ed eventualmente ripetere l’analisi al punto precedente.

- Partendo da una dicotomizzazione del dataset (scartare i tweet neutri e considerare solo le classi positive e negative), definire una baseline di algoritmi di classificazione. Comparare i risultati ottenuti in modo consistente e commentare i risultati. Visualizzare matrice di confusione, curve ROC e curve di apprendimento/convalida; se si considera un approccio di tipo Deep Learning visualizzare la loss.

- Provare a migliorare i risultati ottenuti al punto precedente attraverso una opportuna strategia che comprenda ad esempio un diverso approccio di Text Representation/Embeddings, un tuning dei parametri dei classificatori o una classificazione attraverso tecniche di Deep Learning. Riportare tale strategia (preferibilmente in una unica pipeline) e salvare il modello ottenuto in formato pikle o joblib (o HDF5) per la fase di valutazione.

- Compito opzionale: classificazione multiclasse.

- Il/Gli script ottenuti vanno presentati in formato .py o ipynb insieme ad un breve Report (uno per ciascun gruppo) che contenga la descrizione argomentata delle attività eseguite (max 30 pagine A4) e le varie strategie vanno riassunte in una breve presentazione power point (una per ciascun gruppo).

- Si ricorda che la valutazione terrà conto dei seguenti tre aspetti:

o Performance del modello

o Originalità dell’approccio (se opportunamente giustificato)

o Presentazione del lavoro svolto (Qualità del materiale presentato)

**N.B. Oltre a quanto sopra richiesto per il progetto di gruppo, si ricorda che ogni studente deve scegliere uno tra gli argomenti trattati a lezione e comunicarlo al Docente. Gli approfondimenti effettuati dallo studente sull’argomento concordato con il Docente dovranno poi essere riportati in una ulteriore Tesina individuale da non confondere con il Report del progetto.**

**Allegato 2. Informazioni per la consegna**

Tutto il materiale utile ai fini della valutazione deve essere inviato con un'email al Professor Mario Cannataro (cannataro@unicz.it) (in CC chiara.zucco@ unicz.it) **ENTRO E NON OLTRE** 7 giorni prima la data in cui si desidera sostenere l’esame.

Alcune note:

1) Il materiale dovrà essere consegnato in un unico file .zip dal nome "Prog\_<NumeroGruppo>.zip" (dove <NumeroGruppo> è il numero del gruppo di progetto (come riportato nel file condiviso); ad esempio, se <NumeroGruppo>=3, allora il file da consegnare si chiamerà "Prog\_3.zip") e dovrà comprendere i seguenti file:

a) Cartella “Prog\_<NumeroGruppo>” comprensiva dei seguenti file:

i) “Codice\_ Prog\_<NumeroGruppo>”, contenente il codice relativo al progetto;

ii) “Relazione\_<NumeroGruppo>.pdf”, contenente la relazione tecnica descrivente le attività di progetto;

iii) Slides\_\_<NumeroGruppo>.pdf, contenente una presentazione sotto forma di slides descrivente le attività di progetto;

QUESTI DOCUMENTI SONO CONGIUNTAMENTE REDATTI DA TUTTI I COMPONENTI DEL GRUPPO.

b) Una cartella “Tesina\_<Cognome\_Nome>” relativa all’argomento a scelta che OGNI singolo candidato sceglierà e su cui verrà svolta una tesina di approfondimento, comprensiva di:

i) “Tesina\_Argomento\_<Cognome\_Nome>.pdf”, tesina su argomento a scelta;

ii) “Slides\_Argomento\_<Cognome\_Nome>.pdf/ppt”, presentazione sotto forma di slides dell’argomento a scelta;

iii) Eventuale codice nel caso in cui l’argomento personale presentato preveda un'applicazione pratica da fare valutare. **N.B.** In questo caso il codice deve essere corredato da breve file “Readme.txt” che spiega come eseguirlo. Se file Python deve essere opportunamente commentato.

QUESTI DOCUMENTI SONO REDATTI SINGOLARMENTE DA CIASCUN COMPONENTE DEL GRUPPO.

2) **IMPORTANTE**: sarà cura del/dei candidato/i accertarsi della presenza e della corretta collocazione dei file/directory necessari per l'esecuzione dei codici presentati. A questo fine, il file .zip deve contenere anche **tutti i file necessari** (quindi anche i file relativi ai dataset forniti), opportunamente richiamati nel codice. Se i candidati usano librerie non viste durante le esercitazioni, devono allegare un file requirements.txt contenente la lista delle suddette librerie. **L'impossibilità di eseguire il progetto avrà come conseguenza la valutazione negativa dell'elaborato.**

3) Si ricorda che la Presentazione Stilistica del Progetto di Gruppo è uno degli elementi di valutazione. Sarà cura del/dei candidato/i che gli eventuali grafici e tabelle risultino visibili, prevedere un indice, una sezione introduttiva e delle c

onclusioni, ecc ecc.