Filtro anti-hater per social network

Negli ultimi anni, la moderazione dei contenuti online è diventata una sfida cruciale per molte piattaforme, che si trovano ad affrontare un volume crescente di commenti potenzialmente dannosi. Questi commenti possono includere insulti, minacce, contenuti osceni o messaggi di odio. La moderazione manuale è inefficace su larga scala, e gli algoritmi tradizionali spesso non riescono a catturare la complessità e la varietà dei linguaggi offensivi.

DeepCortex AI Solutions

DeepCortex AI Solutions ha deciso di sviluppare un sistema avanzato basato su tecnologie di Deep Learning per automatizzare e migliorare il processo di moderazione. Il cuore del progetto è un modello di deep learning con **layer ricorrenti**, progettato per classificare commenti in più categorie di tossicità.

Il Problema da Risolvere

L'azienda **TechTalk**, un forum per appassionati di tecnologia, ha riscontrato che un numero significativo di commenti pubblicati nei thread della community contiene espressioni di odio e insulti che compromettono la qualità delle discussioni. Gli utenti hanno segnalato che la piattaforma, a causa della sua popolarità crescente, fatica a gestire il flusso di commenti dannosi con strumenti di moderazione tradizionali. **TechTalk** si è rivolta a **DeepCortex AI Solutions** per implementare una soluzione di moderazione automatica basata su Deep Learning, che sia in grado di filtrare in tempo reale i commenti tossici.

Caso d'Uso

Scenario reale: Mario Rossi, community manager di **TechTalk**, si occupa quotidianamente della moderazione manuale dei contenuti generati dagli utenti. Con l'aumento del traffico sulla piattaforma, Mario non riesce più a gestire manualmente la quantità di commenti dannosi, e deve trovare un modo per filtrare automaticamente i commenti offensivi, minacciosi o osceni senza rallentare l'esperienza utente.

Requisiti Tecnici del Modello

- **Task**: Classificazione multi-label dei commenti in 6 categorie:
    1. **Toxic** (Tossico)
    2. **Severely Toxic** (Super Tossico)
    3. **Obscene** (Osceno)
    4. **Threat** (Minaccia)
    5. **Insult** (Insulto)
    6. **Identity Hate** (Odio basato sull'identità)
- **Dataset**: Un dataset di 160.000 commenti sarà fornito, con ogni commento etichettato in una o più delle categorie sopra indicate. I commenti possono avere zero o più label attive.
- **Architettura**: Il modello deve includere **layer ricorrenti** (ad esempio, LSTM o GRU) per gestire la natura sequenziale dei commenti testuali.
- **Output**: A livello di inferenza, per ogni commento, il modello dovrà produrre un vettore di 6 elementi (uno per ogni label), con valori binari (0 o 1), dove 1 indica la presenza della label corrispondente e 0 la sua assenza.

Fasi del Progetto

1. **Preprocessing dei Dati**:
    - I commenti testuali devono essere convertiti in sequenze numeriche (tokenizzazione).
    - I dati devono essere normalizzati e bilanciati per garantire che tutte le categorie di tossicità siano rappresentate equamente.
2. **Sviluppo del Modello**:

- Il modello di deep learning sarà basato su un'architettura ricorrente, in grado di catturare le dipendenze a lungo termine tra le parole nei commenti.
- Verranno implementati strati ricorrenti (LSTM o GRU) per il task di classificazione multi-label.

3. **Training del Modello**:
    - Il dataset sarà suddiviso in training, validation e test set.
    - Utilizzo di tecniche di ottimizzazione avanzata per migliorare la convergenza del modello.
4. **Inferenza e Predizione**:
    - Durante il tempo di inferenza, per ogni commento, il modello restituirà un vettore di 6 elementi con 0 o 1, a seconda della presenza di tossicità in una o più delle categorie previste.
5. **Validazione**:
    - Il modello sarà valutato utilizzando metriche come **accuracy**, **F1-score** per ciascuna categoria, e **precisione** globale nella previsione delle label multiple.

Valore Aggiunto

- **Automazione**: Il modello ridurrà significativamente il carico di lavoro della moderazione manuale, permettendo a **TechTalk** di gestire un numero maggiore di commenti in tempo reale, mantenendo un ambiente sicuro per gli utenti.
- **Efficienza**: Grazie all'uso di layer ricorrenti, il modello sarà in grado di catturare meglio il contesto e le sfumature dei commenti testuali, migliorando l'accuratezza delle previsioni rispetto a metodi tradizionali.
- **Scalabilità**: Una volta implementato, il sistema sarà facilmente scalabile per gestire volumi crescenti di dati, adattandosi al crescente numero di utenti e commenti sulla piattaforma.
- **Integrazione**: La soluzione sarà integrata direttamente nel sistema di commenti di **TechTalk**, rendendo il filtraggio automatico immediatamente operativo e senza impattare negativamente l'esperienza degli utenti.

Dataset

Il dataset è scaricabile da questo link: https://proai-datasets.s3.eu-west-3.amazonaws.com/Filter_Toxic_Comments_dataset.csv

Conclusione

Il progetto di **DeepCortex AI Solutions** fornirà una soluzione avanzata e automatizzata per la moderazione dei contenuti tossici, migliorando significativamente la qualità delle discussioni online su **TechTalk**. Il sistema garantirà una gestione più efficiente e accurata dei commenti, offrendo una piattaforma sicura e inclusiva per tutti gli utenti.

Modalità di consegna:

Link pubblico a notebook di Google Colab