

**TAG-it**

**Topic, Age and Gender prediction for Italian**

*Evalita task 2020*

Andreozzi Gianni

Toti Riccardo

*Progetto per l’esame di Text Analytics*

*a.a. 2020/21*

# Introduzione

EVALITA è una campagna di valutazione periodica di strumenti di Natural Language Processing (NLP) e vocali per la lingua italiana.

L'obiettivo generale di EVALITA è quello di promuovere lo sviluppo di tecnologie linguistiche e vocali per la lingua italiana, fornendo un quadro condiviso in cui diversi sistemi e approcci possono essere valutati in modo coerente.

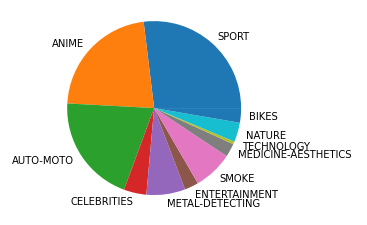
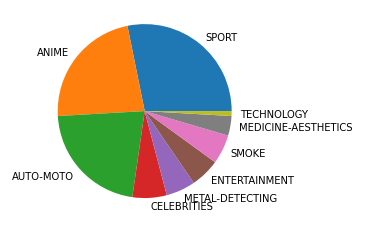
Nel 2020 uno dei task proposti era TAG-IT, il quale obiettivo era uno di profiling, più precisamente si mirava a identificare il genere dell’utente utilizzando i suoi post provenienti da un blog.

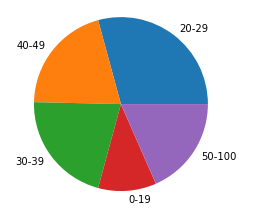
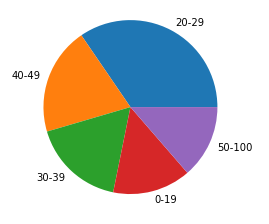
Data Understanding

Il dataset è formato da 1109 record e 5 colonne: *id, age, topic, gender, post*. Ogni record contiene tutti i post di quel determinato utente. Il training set è formato da 999 record e il test set da 110. I record con label “M” (maschio) sono l’82.1% nel training set e l’87.3% nel test set, quelli con label “F” (femmina) sono quindi 17.9% nel training set e 12.7% nel test set. In tabella 1 sono riportate le variabili categoriche con i rispettivi valori che queste possono assumere, la figura 1 mostra la distribuzione di età e topic divisa per gender.

|  |  |
| --- | --- |
| **Variabile** | **Valori** |
| *Age* | ['0-19' '20-29' '30-39' '40-49' '50-100'] |
| *Topic* | ['ANIME' 'BIKES' 'SPORTS' 'AUTO-MOTO' 'NATURE' 'METAL-DETECTING'  'MEDICINE-AESTHETICS' 'CELEBRITIES' 'SMOKE' 'ENTERTAINMENT' 'TECHNOLOGY'] |
| *Gender* | [M F] |

*Tabella 1 – Variabili categoriche contenute nel dataset*



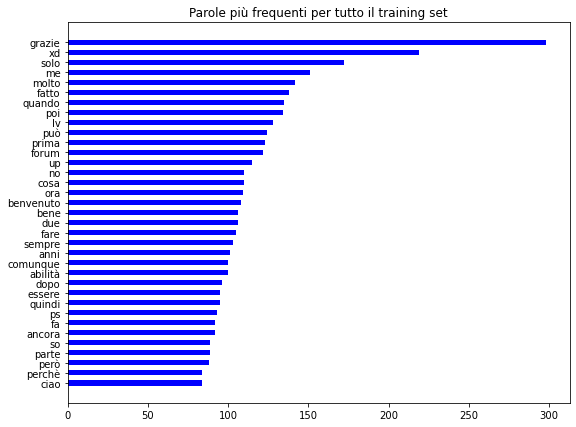


*Figura 1 – Distribuzione age e topic divisa per gender*

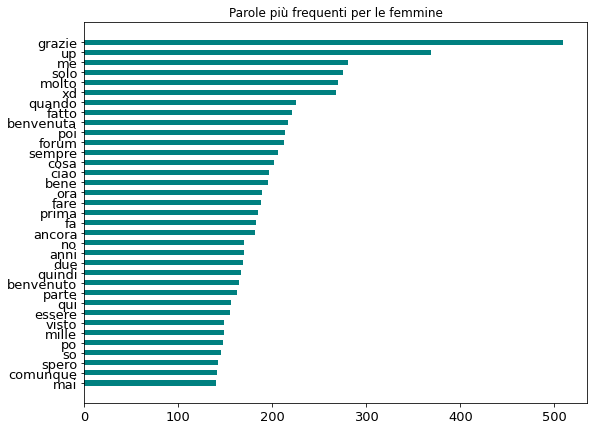
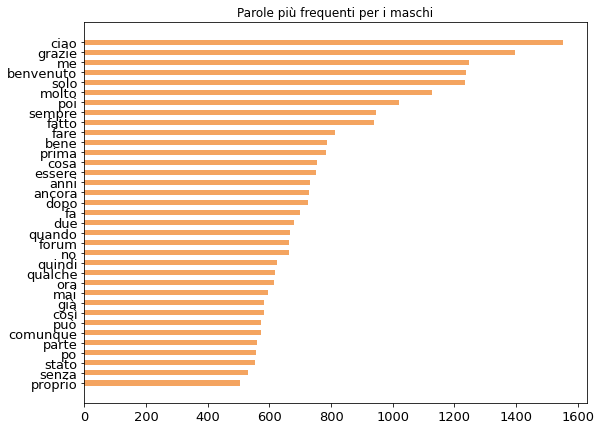
È stato inoltre esplorata la distribuzione delle variabili ‘age’ e ‘topic’ sia nel training che nel test set. Per l’attributo age la distribuzione rimane simile per entrambi i set con la maggioranza degli utenti nel range di età 20-29. Mentre per i topic le categorie maggiormente presenti sono ‘Sport’, ‘Anime’ e ‘Auto-Moto’ per entrambi i set. Nel test set non sono presenti utenti che hanno scritto post inerenti ai topic ‘Nature’ e ‘Bikes’.

È stata effettuata della pulizia del testo prima di procedere con l’analisi. In particolare, sono stati rimossi punteggiatura, stopwords, numeri e alcuni spazi superflui. Questa scelta è stata effettuata dopo aver osservato un miglioramento dei risultati della classificazione rispetto a quelli ottenuti mantenendo queste caratteristiche del testo.

Nelle figure 2, 3 e 4 sono riportate le parole più frequenti, rispettivamente, nell’intero training set, nelle sole femmine e nei soli maschi.

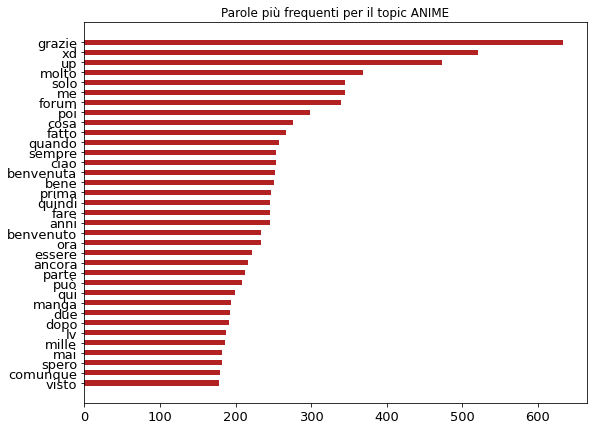
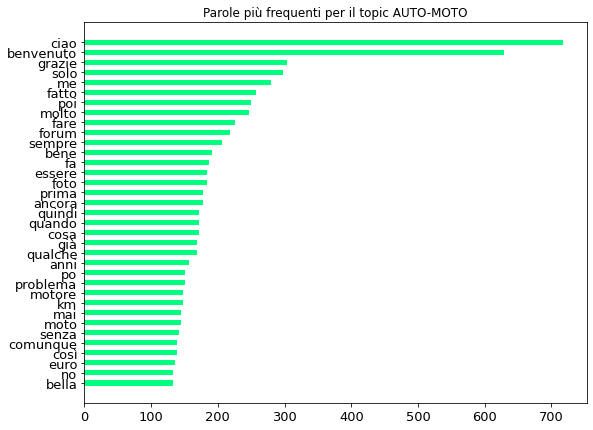


*Figura 2 – Parole più frequenti nell’intero training set*

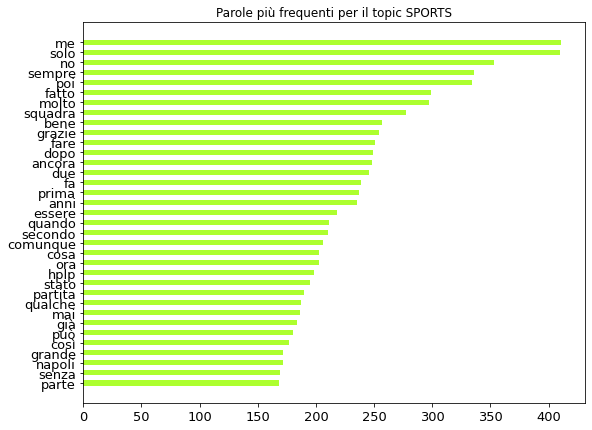


*Figura 3 – Parole più frequenti nella categoria ‘Male’ Figura 4 – Parole più frequenti nella categoria “Female”*

Sono stati analizzati i vocabolari per topic. Sono riportate in figura 5, figura 6 e figura 7 le 40 parole più frequenti per i topic più presenti sia nel training che nel test set: ‘ANIME’,’AUTO-MOTO’ e ‘SPORTS’.



*Figura 5 – parole più frequenti per il topic AUTO-MOTO Figura 6 – parole più frequenti per il topic ANIME*



*Figura 7 – parole più frequenti per il topic SPORTS*

È stato possibile osservare che le parole con maggiore frequenza combaciano per tutti i topic ma alcune sono caratterizzanti del topic in questione. Infatti, si può notare come nel topic ‘ANIME’ la parola *‘manga’* sia presente più di 200 volte, nel topic ‘AUTO-MOTO’ siano presenti le parole *‘motore’*, *‘km’* e *‘consumi’* e infine nel topic ‘SPORTS’ siano presenti *‘squadra’*, *‘partita’* e *‘napoli’*. Il testo estratto da social network riporta molte espressioni tipiche di una piattaforma di networking come, per esempio, l’utilizzo dell’abbreviazione ‘x’ in sostituzione della preposizione semplice ‘per’ e l’utilizzo dell’espressione ‘up’ che viene impiegata per aumentare la visibilità di un post all’interno dei feed degli altri utenti.

Gender Prediction

Precedentemente alle operazioni di classificazione, è stata effettuata la lemmatization del testo e la divisione di questo in bigrammi e trigrammi.

Il dataset è sbilanciato per la classe Gender: l’82% dei record nel training set sono maschi mente nel test set la percentuale è 87%.

Utilizzando l’intero dataset è stato osservato come i classificatori riuscissero a predirre con dei risultati soddisfacenti il gender ‘male’ mentre il gender ‘female’ riportava dei risultati peggiori.

Di seguito il risultato dei classificatori applicati all’intero dataset.

|  |  |  |  |
| --- | --- | --- | --- |
| LINEAR SVM | | | |
|  | **Precision** | **Recall** | **F1** |
| F | 0.60 | 0.64 | 0.62 |
| M | 0.95 | 0.94 | 0.94 |
| Accuracy |  |  | 0.90 |
| Macro average | 0.77 | 0.79 | 0.78 |
| Weighted average | 0.90 | 0.90 | 0.90 |

|  |  |
| --- | --- |
| 9 | 5 |
| 6 | 90 |

*Tabella 2 – Risultati per Linear SVM*

|  |  |  |  |
| --- | --- | --- | --- |
| DECISION TREE | | | |
|  | **Precision** | **Recall** | **F1** |
| F | 0.38 | 0.36 | 0.37 |
| M | 0.91 | 0.91 | 0.92 |
| Accuracy |  |  | 0.90 |
| Macro average | 0.65 | 0.64 | 0.64 |
| Weighted average | 0.84 | 0.85 | 0.84 |

|  |  |
| --- | --- |
| 5 | 9 |
| 8 | 88 |

*Tabella 3 – Risultati per Decision Tree*

|  |  |  |  |
| --- | --- | --- | --- |
| NAIVE BAYES | | | |
|  | **Precision** | **Recall** | **F1** |
| F | 0.44 | 0.86 | 0.59 |
| M | 0.98 | 0.84 | 0.91 |
| Accuracy |  |  | 0.85 |
| Macro average | 0.71 | 0.85 | 0.75 |
| Weighted average | 0.91 | 0.85 | 0.86 |

|  |  |
| --- | --- |
| 12 | 2 |
| 15 | 81 |

*Tabella 4 – Risultati per Naive Bayes*

|  |  |  |  |
| --- | --- | --- | --- |
| LOGISTIC REGRESSION | | | |
|  | **Precision** | **Recall** | **F1** |
| F | 0.35 | 0.79 | 0.49 |
| M | 0.96 | 0.79 | 0.87 |
| Accuracy |  |  | 0.79 |
| Macro average | 0.66 | 0.79 | 0.68 |
| Weighted average | 0.88 | 0.79 | 0.82 |

|  |  |
| --- | --- |
| 11 | 3 |
| 20 | 76 |

*Tabella 5 – Risultati per la Logistic Regression*



*Figura 8 – Learning Curves per Naïve Bayes classifier e Linear SVC sull’intero training set*

Per il Multinomial Naïve Bayes, sia il validation score che il training score convergono verso un valore che è abbastanza basso con l'aumentare della dimensione del set di addestramento. Quindi, probabilmente non trarremmo molto beneficio aumentando il numero di record nel training set.

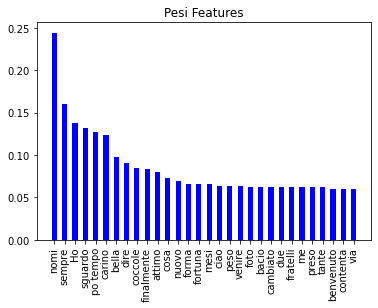
Mentre il LinearSVC presenta un gap tra le due curve del training score e validation score che corrisponde ad un'alta varianza, presentando un principio di overfitting del modello.

Campionamento del dataset

Nel tentativo di migliorare le performance dei classifier sono state applicate delle tecniche di campionamento stratificato al fine di riequilibrare il training set per la classe ‘Gender’. Il training set risultante dal campionamento contiene 250 record distribuiti equamente tra maschi e femmine, mentre nessuna operazione è stata effettuata sul test set che contiene 96 maschi e 14 femmine.

Nel rispetto del subtask2a, in cui la predizione della classe Gender assume che il *topic* sia casuale, è stato necessario produrre un campione nel quale non prevalesse nessun *topic* specifico.

I pesi per le features da utilizzare nei vari classifier sono stati assegnati utilizzando la misura tf-idf. Nella figura 9 sono riportate le prime 30 parole in ordine decrescente di peso.



*Figura 9 – features ordinate per peso*

Di seguito sono riportati i risultati della classificazione dopo il campionamento.

|  |  |  |  |
| --- | --- | --- | --- |
| LINEAR SVM | | | |
|  | **Precision** | **Recall** | **F1** |
| F | 0.28 | 0.93 | 0.43 |
| M | 0.98 | 0.66 | 0.79 |
| Accuracy |  |  | 0.69 |
| Macro average | 0.63 | 0.79 | 0.61 |
| Weighted average | 0.90 | 0.69 | 0.74 |

|  |  |
| --- | --- |
| 63 | 33 |
| 1 | 13 |

*Tabella 6 – Risultati per Linear SVM*

|  |  |  |  |
| --- | --- | --- | --- |
| DECISION TREE | | | |
|  | **Precision** | **Recall** | **F1** |
| F | 0.21 | 0.79 | 0.33 |
| M | 0.95 | 0.57 | 0.71 |
| Accuracy |  |  | 0.60 |
| Macro average | 0.58 | 0.68 | 0.52 |
| Weighted average | 0.85 | 0.60 | 0.67 |

|  |  |
| --- | --- |
| 55 | 41 |
| 3 | 11 |

*Tabella 7 – Risultati per Decision Tree*

|  |  |  |  |
| --- | --- | --- | --- |
| NAIVE BAYES | | | |
|  | **Precision** | **Recall** | **F1** |
| F | 0.29 | 0.71 | 0.42 |
| M | 0.95 | 0.75 | 0.84 |
| Accuracy |  |  | 0.75 |
| Macro average | 0.62 | 0.73 | 0.63 |
| Weighted average | 0.86 | 0.75 | 0.78 |

|  |  |
| --- | --- |
| 72 | 24 |
| 4 | 10 |

*Tabella 8 – Risultati per Naive Bayes*

|  |  |  |  |
| --- | --- | --- | --- |
| LOGISTIC REGRESSION | | | |
|  | **Precision** | **Recall** | **F1** |
| F | 0.29 | 0.93 | 0.43 |
| M | 0.98 | 0.66 | 0.79 |
| Accuracy |  |  | 0.69 |
| Macro average | 0.63 | 0.79 | 0.61 |
| Weighted average | 0.90 | 0.69 | 0.74 |

|  |  |
| --- | --- |
| 12 | 2 |
| 29 | 63 |

*Tabella 9 – Risultati per la Logistic Regression*



*Figura 10 – Learning Curves per Naïve Bayes classifier e Linear SVC per il training set post-campionamento*

Per il campionamento del dataset i due miglior modelli presentano un peggioramento in performance notevole, soprattutto il MultinomialNB che beneficiava del numero del maggior numero di record diminuisce in accuracy presentando un principio di underfitting non riuscendo a generalizzare e discretizzare con meno records.

Il linearSVM presenta come in figura 8 un’alta varianza che suggerisce un fenomeno di overfitting ancor più accentuato rispetto al modello applicato al dataset intero.

Gender Prediction – Topic control

Le analisi sono state ripetute controllando per topic. In particolare, sono stati analizzati i topic ‘Anime’ e ‘Celebrities’. La scelta di questi è dovuta al fatto che essi contenevano un numero sufficiente di osservazioni per entrambe le classi da prevedere, mentre all’interno degli altri topic il dataset era altamente sbilanciato, con alcuni che presentavano unicamente utenti maschi.

Per il topic ‘Anime’ sono stati individuati 222 utenti (117 maschi e 105 femmine) per il training e 25 user per il test set, mentre per il topic ‘Celebrities’ ne sono stati individuati 42 nel training set (19 maschie 23 femmine) e 7 nel test set.

Nelle tabelle 10, 11, 12 e 13 sono riportati i risultati dei classifier sul subset del topic ‘Anime’. I risultati migliori in termini di F1 score a precisione sono stati ottenuti utilizzando una Support Vector Machine lineare (0.80 e 0,79 rispettivamente). Gli indici sono stati notevolmente peggiori utilizzando il Decision Tree e solo poco peggiori utilizzando un Naive Bayes classifier.

|  |  |  |  |
| --- | --- | --- | --- |
| LINEAR SVM | | | |
|  | **Precision** | **Recall** | **F1** |
| F | 0.75 | 0.67 | 0.71 |
| M | 0.82 | 0.88 | 0.85 |
| Accuracy |  |  | 0.80 |
| Macro average | 0.79 | 0.77 | 0.78 |
| Weighted average | 0.80 | 0.80 | 0.80 |

|  |  |
| --- | --- |
| 6 | 3 |
| 2 | 14 |

*Tabella 10 – Risultati per Linear SVM*

|  |  |  |  |
| --- | --- | --- | --- |
| DECISION TREE | | | |
|  | **Precision** | **Recall** | **F1** |
| F | 0.50 | 0.67 | 0.57 |
| M | 0.77 | 0.62 | 0.69 |
| Accuracy |  |  | 0.64 |
| Macro average | 0.63 | 0.65 | 0.63 |
| Weighted average | 0.67 | 0.64 | 0.65 |

|  |  |
| --- | --- |
| 6 | 3 |
| 6 | 10 |

*Tabella 11 – Risultati per Decision Tree*

|  |  |  |  |
| --- | --- | --- | --- |
| NAIVE BAYES | | | |
|  | **Precision** | **Recall** | **F1** |
| F | 0.64 | 0.78 | 0.70 |
| M | 0.86 | 0.75 | 0.80 |
| Accuracy |  |  | 0.76 |
| Macro average | 0.75 | 0.76 | 0.75 |
| Weighted average | 0.78 | 0.76 | 0.76 |

|  |  |
| --- | --- |
| 6 | 3 |
| 4 | 12 |

*Tabella 12 – Risultati per Naive Bayes*

|  |  |  |  |
| --- | --- | --- | --- |
| LOGISTIC REGRESSION | | | |
|  | **Precision** | **Recall** | **F1** |
| F | 0.64 | 0.78 | 0.70 |
| M | 0.86 | 0.75 | 0.80 |
| Accuracy |  |  | 0.76 |
| Macro average | 0.75 | 0.76 | 0.75 |
| Weighted average | 0.78 | 0.76 | 0.76 |

|  |  |
| --- | --- |
| 6 | 3 |
| 4 | 12 |

*Tabella 13 – Risultati per Logistic Regression*

Nelle tabelle sono riportati i risultati dei classifier sul subset del topic ‘Celebrities’. La numerosità del test set, solo 7 osservazioni, fa sì che gli indici dei vari classifier abbiano valori simili e che la differenza ottenuta per il Decision Tree sia interamente dovuta ad una singola osservazione che da vero negativo per gli altri classifier, diventa falso negativo.

|  |  |  |  |
| --- | --- | --- | --- |
| LINEAR SVM | | | |
|  | **Precision** | **Recall** | **F1** |
| F | 0.67 | 0.67 | 0.67 |
| M | 0.75 | 0.75 | 0.75 |
| Accuracy |  |  | 0.71 |
| Macro average | 0.71 | 0.71 | 0.71 |
| Weighted average | 0.71 | 0.71 | 0.71 |

|  |  |
| --- | --- |
| 2 | 1 |
| 1 | 3 |

*Tabella 14 – Risultati per Linear SVM*

|  |  |  |  |
| --- | --- | --- | --- |
| DECISION TREE | | | |
|  | **Precision** | **Recall** | **F1** |
| F | 0.50 | 0.33 | 0.40 |
| M | 0.60 | 0.75 | 0.67 |
| Accuracy |  |  | 0.57 |
| Macro average | 0.55 | 0.54 | 0.53 |
| Weighted average | 0.56 | 0.57 | 0.55 |

|  |  |
| --- | --- |
| 1 | 2 |
| 1 | 3 |

*Tabella 15 – Risultati per Decision Tree*

|  |  |  |  |
| --- | --- | --- | --- |
| NAIVE BAYES | | | |
|  | **Precision** | **Recall** | **F1** |
| F | 0.67 | 0.67 | 0.67 |
| M | 0.75 | 0.75 | 0.75 |
| Accuracy |  |  | 0.71 |
| Macro average | 0.71 | 0.71 | 0.71 |
| Weighted average | 0.71 | 0.71 | 0.71 |

|  |  |
| --- | --- |
| 2 | 1 |
| 1 | 3 |

*Tabella 16 – Risultati per Naive Bayes*

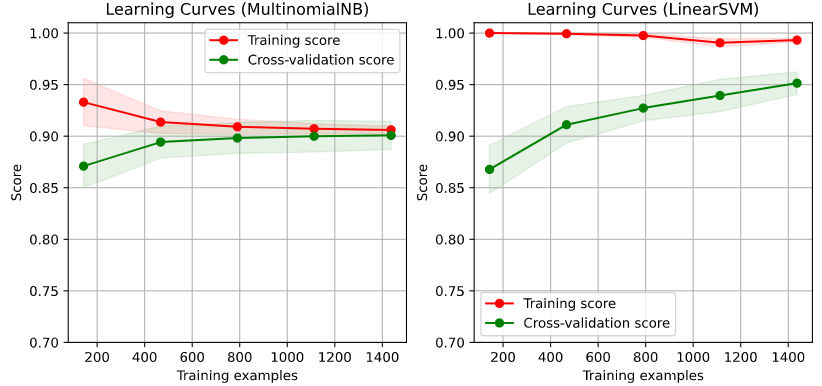
|  |  |  |  |
| --- | --- | --- | --- |
| LOGISTIC REGRESSION | | | |
|  | **Precision** | **Recall** | **F1** |
| F | 0.67 | 0.67 | 0.67 |
| M | 0.75 | 0.75 | 0.75 |
| Accuracy |  |  | 0.71 |
| Macro average | 0.71 | 0.71 | 0.71 |
| Weighted average | 0.71 | 0.71 | 0.71 |

|  |  |
| --- | --- |
| 2 | 1 |
| 1 | 3 |

*Tabella 17 – Risultati per la Logistic Regression*



*Figura 11 – Learning Curves per Naïve Bayes e Linear SVM controllando per il topic ANIME*



*Figura 12 – Learning Curves per Naïve Bayes e Linear SVM controllando per il topic ANIME*

Osservando le learning curves sulla predizione del genere per topic è possibile osservare che il ridimensionamento e la focalizzazione permettono ai due modelli di classificazione di operare in modo migliore. La varianza è notevolmente diminuita nel LinearSVC, potendo quindi affermare che i risultati ottenuti migliorano sensibilmente quando il la predizione è svolta su un particolare Topic.

Age prediction

Il task successivo riguardava la predizione della variabile età. La variabile è divisa in cinque classi: 0-19, 20-29, 30-39, 40-49 e 50+. Analogamente alla predizione del genere, la classificazione è stata effettuata sia per tutto il dataset sia controllando per i topic che lo permettevano, dove il principale fattore di *ostruzione… no, qualcosa* era la numerosità campionaria. Nelle tabelle 18 e 19 sono riportati i risultati per la classificazione sull’intero dataset.

|  |  |  |  |
| --- | --- | --- | --- |
| LINEAR SVM | | | |
|  | **Precision** | **Recall** | **F1** |
| 0-19 | 0.50 | 0.25 | 0.33 |
| 20-29 | 0.46 | 0.58 | 0.51 |
| 30-39 | 0.29 | 0.21 | 0.24 |
| 40-49 | 0.21 | 0.23 | 0.22 |
| 50+ | 0.31 | 0.33 | 0.32 |
| Accuracy |  |  | 0.36 |
| Macro average | 0.35 | 0.32 | 0.33 |
| Weighted average | 0.36 | 0.36 | 0.35 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 4 | 9 | 0 | 2 | 1 |
| 3 | 22 | 4 | 9 | 0 |
| 0 | 6 | 4 | 4 | 5 |
| 1 | 6 | 5 | 5 | 5 |
| 0 | 5 | 1 | 4 | 5 |

*Tabella 18 – Risultati per Linear SVM*

|  |  |  |  |
| --- | --- | --- | --- |
| LINEAR SVM (ONE VS ONE) | | | |
|  | **Precision** | **Recall** | **F1** |
| 0-19 | 0.60 | 0.19 | 0.29 |
| 20-29 | 0.46 | 0.66 | 0.54 |
| 30-39 | 0.31 | 0.21 | 0.25 |
| 40-49 | 0.28 | 0.32 | 0.30 |
| 50+ | 0.38 | 0.33 | 0.36 |
| Accuracy |  |  | 0.40 |
| Macro average | 0.41 | 0.34 | 0.35 |
| Weighted average | 0.41 | 0.40 | 0.38 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 3 | 10 | 1 | 2 | 0 |
| 2 | 25 | 5 | 6 | 0 |
| 0 | 8 | 4 | 5 | 2 |
| 0 | 6 | 3 | 7 | 6 |
| 0 | 5 | 0 | 5 | 5 |

*Tabella 19 – Risultati per One Vs One (Linear SVM)*

|  |  |  |  |
| --- | --- | --- | --- |
| LOGISTIC REGRESSION | | | |
|  | **Precision** | **Recall** | **F1** |
| 0-19 | 0.42 | 0.31 | 0.36 |
| 20-29 | 0.38 | 0.26 | 0.31 |
| 30-39 | 0.22 | 0.32 | 0.26 |
| 40-49 | 0.32 | 0.32 | 0.32 |
| 50+ | 0.30 | 0.47 | 0.37 |
| Accuracy |  |  | 0.32 |
| Macro average | 0.33 | 0.34 | 0.32 |
| Weighted average | 0.34 | 0.32 | 0.32 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 5 | 7 | 0 | 1 | 3 |
| 5 | 10 | 10 | 9 | 4 |
| 1 | 5 | 6 | 3 | 4 |
| 1 | 3 | 6 | 7 | 5 |
| 0 | 1 | 5 | 2 | 7 |

*Tabella 18 – Risultati per la Logistic Regression*

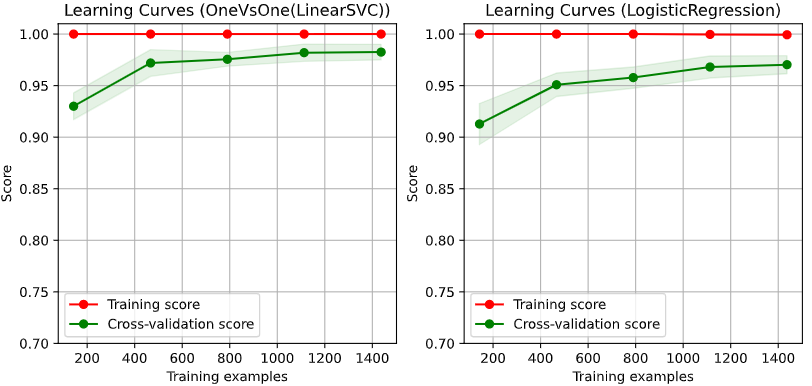


Figura 13 – Learning Curves per OneVsOne e Logistic Regression

Per la multi-class classification i risultati che possono essere estratti dalle learning curves suggeriscono un precision maggiore da parte del classificatore OneVsOne con LinearSVC che raggiunge una Weighted Accuracy di 0.40 e una varianza che tende a diminuire al crescere dei records, segno che una quantità di dati maggiori avrebbe potuto migliorare la precisione del classifier. La Logistic Regression raggiunge dei risultati peggiori pur mantenendo una varianza tra il training e la cross-validation accettabile.

Age prediction controllando per topic

Le analisi svolte per la previsione della classe di età sono state poi ripetute controllando per topic, in particolare sono stati analizzati i topic ANIME, SPORTS e AUTO-MOTO. Per il primo di questi, le distribuzioni dell’età nei 222 record del training set e nei 25 record del test set, sono riportate in figura 13.

*Figura 13 – Distribuzioni età nel training e nel test set per il topic ANIME*

Nelle tabelle 20 e 21 sono riportati i risultati della classificazione.

|  |  |  |  |
| --- | --- | --- | --- |
| LINEAR SVM | | | |
|  | **Precision** | **Recall** | **F1** |
| 0-19 | 0.83 | 0.50 | 0.62 |
| 20-29 | 0.50 | 0.89 | 0.64 |
| 30-39 | 0.31 | 0.21 | 0.25 |
| 40-49 | 0.28 | 0.32 | 0.30 |
| 50+ | 0.38 | 0.33 | 0.36 |
| Accuracy |  |  | 0.52 |
| Macro average | 0.27 | 0.28 | 0.25 |
| Weighted average | 0.51 | 0.52 | 0.48 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 5 | 3 | 1 | 1 | 0 |
| 1 | 8 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 |
| 0 | 2 | 1 | 0 | 0 |
| 0 | 2 | 0 | 0 | 0 |

*Tabella 20 – Risultati per LinearSVM*

|  |  |  |  |
| --- | --- | --- | --- |
| ONE VS ONE (LINEAR SVM) | | | |
|  | **Precision** | **Recall** | **F1** |
| 0-19 | 0.75 | 0.30 | 0.43 |
| 20-29 | 0.44 | 0.89 | 0.59 |
| 30-39 | 0.00 | 0.00 | 0.00 |
| 40-49 | 0.00 | 0.00 | 0.00 |
| 50+ | 0.00 | 0.00 | 0.00 |
| Accuracy |  |  | 0.44 |
| Macro average | 0.24 | 0.24 | 0.20 |
| Weighted average | 0.46 | 0.44 | 0.38 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 3 | 5 | 1 | 1 | 0 |
| 1 | 8 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 |
| 0 | 2 | 1 | 0 | 0 |
| 0 | 2 | 0 | 0 | 0 |

*Tabella 21 – Risultati per One Vs One (Linear SVM)*

SPORTS

Nel topic **SPORTS** il numero di record era di 269. Rispetto al topic ANIME, gli utenti hanno mediamente un’età maggiore e la classe di età 0-19 è la meno rappresentata, con meno del 10% di record sia per il training che per il test set, mentre la percentuale era oltre il 30% nel precedente topic. La figura 14 riporta le distribuzioni di età per le 269 osservazioni del training set e le 31 del test set.

*Figura 14 – Distribuzioni età nel training e nel test set per il topic SPORTS*

Nelle tabelle 22 e 23 sono riportati i risultati per i classifier.

|  |  |  |  |
| --- | --- | --- | --- |
| LINEAR SVM | | | |
|  | **Precision** | **Recall** | **F1** |
| 0-19 | 0.00 | 0.00 | 0.00 |
| 20-29 | 0.70 | 0.47 | 0.56 |
| 30-39 | 0.33 | 0.80 | 0.47 |
| 40-49 | 0.22 | 0.40 | 0.29 |
| 50+ | 0.00 | 0.00 | 0.00 |
| Accuracy |  |  | 0.42 |
| Macro average | 0.25 | 0.33 | 0.26 |
| Weighted average | 0.43 | 0.42 | 0.39 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 0 | 1 | 1 | 0 | 0 |
| 0 | 7 | 3 | 5 | 0 |
| 0 | 0 | 4 | 1 | 0 |
| 0 | 1 | 2 | 2 | 0 |
| 0 | 1 | 2 | 1 | 0 |

*Tabella 22 – Risultati per Linear SVM*

|  |  |  |  |
| --- | --- | --- | --- |
| ONE VS ONE (LINEAR SVM) | | | |
|  | **Precision** | **Recall** | **F1** |
| 0-19 | 0.00 | 0.00 | 0.00 |
| 20-29 | 0.75 | 0.40 | 0.52 |
| 30-39 | 0.33 | 0.80 | 0.47 |
| 40-49 | 0.22 | 0.40 | 0.29 |
| 50+ | 0.00 | 0.00 | 0.00 |
| Accuracy |  |  | 0.39 |
| Macro average | 0.25 | 0.33 | 0.26 |
| Weighted average | 0.43 | 0.42 | 0.37 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 0 | 1 | 1 | 0 | 0 |
| 0 | 6 | 4 | 5 | 0 |
| 0 | 0 | 4 | 0 | 1 |
| 0 | 1 | 1 | 2 | 1 |
| 0 | 0 | 2 | 2 | 0 |

*Tabella 23 – Risultati per One Vs One (Linear SVM)*

AUTO-MOTO

Per il topic AUTO-MOTO i record nel training set erano 202 e quelli nel test 24. La classe 0-19 era quella meno rappresentata con le altre classi sufficientemente bilanciate. Nella figura 15 sono riportate le distribuzioni per la classe età per training set e test set.

*Figura 15 – Distribuzioni età nel training e nel test set per il topic AUTO-MOTO*

Nelle tabelle 24 e 25 sono riportati i risultati dei classifier.

|  |  |  |  |
| --- | --- | --- | --- |
| LINEAR SVM | | | |
|  | **Precision** | **Recall** | **F1** |
| 0-19 | 0.00 | 0.00 | 0.00 |
| 20-29 | 0.33 | 0.25 | 0.29 |
| 30-39 | 0.00 | 0.00 | 0.00 |
| 40-49 | 0.23 | 0.43 | 0.30 |
| 50+ | 0.50 | 0.33 | 0.40 |
| Accuracy |  |  | 0.25 |
| Macro average | 0.21 | 0.20 | 0.20 |
| Weighted average | 0.25 | 0.25 | 0.24 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 0 | 0 | 1 | 1 | 0 |
| 0 | 1 | 2 | 1 | 0 |
| 0 | 1 | 0 | 4 | 0 |
| 0 | 1 | 1 | 3 | 2 |
| 0 | 0 | 0 | 4 | 2 |

*Tabella 24 – Risultati per Linear SVM*

|  |  |  |  |
| --- | --- | --- | --- |
| ONE VS ONE (LINEAR SVM) | | | |
|  | **Precision** | **Recall** | **F1** |
| 0-19 | 0.00 | 0.00 | 0.00 |
| 20-29 | 0.33 | 0.25 | 0.29 |
| 30-39 | 0.00 | 0.00 | 0.00 |
| 40-49 | 0.18 | 0.29 | 0.22 |
| 50+ | 0.29 | 0.33 | 0.31 |
| Accuracy |  |  | 0.21 |
| Macro average | 0.16 | 0.17 | 0.16 |
| Weighted average | 0.18 | 0.21 | 0.19 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 0 | 0 | 0 | 2 | 0 |
| 0 | 1 | 2 | 0 | 1 |
| 0 | 1 | 0 | 4 | 0 |
| 0 | 0 | 1 | 2 | 4 |
| 0 | 1 | 0 | 3 | 2 |

*Tabella 25 – Risultati per One Vs One (Linear SVM)*

Le previsioni si sono rivelate migliori per i modelli considerati controllando per topic, rispetto a quelli testati sull’intero dataset. I risultati migliori in termini di F1 e Accuracy, sono stati ottenuti nel topic ANIME (0.52 e 0.27 rispettivamente) sebbene le classi fossero più sbilanciate rispetto agli altri topic considerati. Le performance dei modelli sono generalmente molto scarse ed è sconsigliato utilizzarli per fare delle vere previsioni.

GENDER PREDICTION CON SIMPLE TRANSFORMER

Nel tentativo di esplorare più approfonditamente la predizione del*gender* è stato implementato un approccio non supervisionato utilizzando un modello *pre-trained* di derivazione BERT.

È stato scelto un modello che si avvicinasse il più possibile al tipo di testo che il dataset riporta, testi che comprendono termini utilizzati prettamente nel contesto dei social network, specificatamente su Twitter e un linguaggio per la maggior parte di basso livello. In particolare, questo modello utilizza quindi caratteristiche del testo proprie dei social media come emoticons, mentions e altre feature che si possono trovare come URL, indirizzi email eccetera.

È stato perciò scelto il modello **ALBERTo-it**, *pre-trained* su tweet italiani tra il 2019/20, creato e messo a disposizione sulle piattaforme Github e Hugginface da Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro e Valerio Basile.

È stato notato un incremento di performance all’aumentare delle epoch. Qui di seguito vengono riportati i risultati con un numero di epoch fissato a 10.

La classification report di Simple Transformer riporta il risultato dell’accuracy del modello con la **Matthews correlation coefficient** (Mcc), una metrica estremamente diffusa per la valutazione della classificazione binaria di un modello. La formula per calcolare questa metrica è:

In tabella 26 sono riportati i risultati della classificazione.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Auprc | auroc | eval\_loss | MCC | ACCURACY |
| 0.976 | 0.860 | 0.877 | **0.384** | **0.854** |

|  |  |
| --- | --- |
| 87 | 7 |
| 9 | 7 |

*Tabella 26 – Risultati per SimpleTransformers*

**CONCLUSIONI**

La challenge proposta da EVALITA richiedeva di classificare il genere e l’età di alcuni utenti, i quali commenti sono stati raccolti da piattaforme social. Il dataset fornito presentava delle sfide per quello che riguardava il numero di dati sia nel training set che nel test set. In particolare, per il *gender* il dataset si presentava sbilanciato in favore dei maschi, quindi delle tecniche di bilanciamento sono state applicate con moderato successo per ottenere risultati migliori nella classificazione.

In generale il modello migliore si è rivelato essere la Linear SVM, sia nell’analisi svolta sul dataset intero che controllando per *topic*, per entrambi i task di classificazione, *gender* e *età*. Controllando per *topic*, nella classificazione dell’età, i risultati si sono rivelati migliori rispetto a quelli ottenuti sull’intero dataset. Questo potrebbe essere dovuto alla maggiore omogeneità dei dati e al ridotto rumore contenuto negli stessi.

Esplorando le potenzialità di BERT (Biderectional Encoder Representations from Transformers) e utilizzando un suo derivato ALBERTo-it, si sono ottenuti i risultati migliori con un’accuracy dell’85.4%.

Alla base dei limiti di quest’analisi, c’è la qualità dei dati utilizzati. Un maggior numero di training data sarebbe consigliato per migliorare la predizione della classificazione. Questo può aiutare anche ad avere una distribuzione più omogenea dei topic stessi e delle classi prese in considerazione al loro interno.