Capitolo 1

Results

Questo capitolo di questo lavoro di tesi dedicato alla presentazione dei risultati ottenuti dai vari esperimenti effettuati durante il lavoro. Nello specifico verranno presentati i risultati in fase di test dei modelli utilizzati oltre che i risultati in test ottenuti a partire dalle API del modello Gemini. Prima dei risultati, verranno introdotte le metriche scelte per valutare i modelli.

1.1 Metriche di Valutazione

Le metriche di valutazione sono utilizzate per misurare le prestazioni di un modello di classificazione. Esse sono calcolate confrontando le previsioni del modello con le etichette reali del dataset. Le metriche di valutazione pi comuni includono Accuracy, Precision, Recall e F1 Score. Queste metriche sono calcolate a partire dalle seguenti misurazioni: True Positives (TP), True Negatives (TN), False Positives (FP) e False Negatives (FN).

- **True Positives (TP)**: Il numero di istanze correttamente classificate come appartenenti a una certa classe.
- **True Negatives (TN)**: Il numero di istanze correttamente classificate come non appartenenti a una certa classe.
- **False Positives (FP)**: Il numero di istanze erroneamente classificate come appartenenti a una certa classe.
- **False Negatives (FN)**: Il numero di istanze erroneamente classificate come non appartenenti a una certa classe.

Le metriche, calcolate a partire da queste misurazioni, utilizzate in questo lavoro di tesi sono:

Accuracy

L'Accuracy misura la proporzione di previsioni corrette sul totale delle istanze e viene calcolata come:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Pro: Facile da interpretare e calcolare.
Contro: In un contesto multilabel con classi sbilanciate, pu essere fuorviante perch non considera la distribuzione delle classi.

Precision

La Precision (Micro, Macro, Weighted) misura la proporzione di istanze rilevanti tra quelle recuperate. Le formulazioni sono:

- **Micro Precision**: Aggrega i contributi di tutte le classi per calcolare la precision complessiva.

$$\text{Micro Precision} = \frac{\sum TP}{\sum TP + \sum FP}$$

- **Macro Precision**: Calcola la precision per ogni classe e poi ne fa la media.

$$\text{Macro Precision} = \frac{1}{N} \sum_{i=1}^{N} \text{Precision}_i$$

- **Weighted Precision**: Calcola la precision per ogni classe ponderata per il numero di veri positivi.

$$\text{Weighted Precision} = \frac{\sum_{i=1}^{N} \text{Precision}_i \times w_i}{\sum_{i=1}^{N} w_i}$$

Pro: Indica quanto rilevanti sono le previsioni fatte.
Contro: Pu essere ingannevole se il modello ha pochi falsi positivi ma molti falsi negativi.

Recall

La Recall (Micro, Macro, Weighted) misura la proporzione di istanze rilevanti che sono state recuperate. Le formulazioni sono:

- **Micro Recall**: Aggrega i contributi di tutte le classi per calcolare la recall complessiva.

$$\text{Micro Recall} = \frac{\sum TP}{\sum TP + \sum FN}$$

- **Macro Recall**: Calcola la recall per ogni classe e poi ne fa la media.

$$\text{Macro Recall} = \frac{1}{N} \sum_{i=1}^{N} \text{Recall}_i$$

- **Weighted Recall**: Calcola la recall per ogni classe ponderata per il numero di veri positivi.

$$\text{Weighted Recall} = \frac{\sum_{i=1}^{N} \text{Recall}_i \times w_i}{\sum_{i=1}^{N} w_i}$$

Pro: Indica quanto bene il modello in grado di identificare tutte le istanze rilevanti.
Contro: Pu essere fuorviante se il modello ha molti falsi positivi.

F1 Score

L'F1 Score (Micro, Macro, Weighted) la media armonica di precision e recall, offrendo un bilanciamento tra le due. Le formulazioni sono:

- **Micro F1**: Combina micro precision e micro recall.

$$\text{Micro F1} = 2 \times \frac{\text{Micro Precision} \times \text{Micro Recall}}{\text{Micro Precision} + \text{Micro Recall}}$$

- **Macro F1**: Calcola l'F1 score per ogni classe e poi ne fa la media.

$$\text{Macro F1} = \frac{1}{N} \sum_{i=1}^{N} \text{F1}_i$$

- **Weighted F1**: Calcola l'F1 score per ogni classe ponderata per il numero di veri positivi.

$$\text{Weighted F1} = \frac{\sum_{i=1}^{N} \text{F1}_i \times w_i}{\sum_{i=1}^{N} w_i}$$

Pro: Bilancia precision e recall in un'unica metrica.
Contro: Pu essere difficile da interpretare in presenza di classi molto sbilanciate.

1.2 Risultati

In questa sezione verranno presentati i risultati ottenuti dai vari esperimenti effettuati durante il lavoro di tesi. Verranno presentati i risultati di tutti gli esperimenti fatti, si quelli relativi ai modelli di classificazione utilizzati e presentati nel capitolo precedente che quelli relativi ai risultati ottenuti a partire delle API del modello Gemini.

1.2.1 Risultati Gemini

Metrica	Accuracy	Precision	Recall	F1 Score
Micro	-	0.3271	0.1872	0.2382
Macro	-	0.3137	0.1706	0.1800
Weighted	-	0.3490	0.1872	0.2056
Overall Accuracy	0.2754	-	-	-

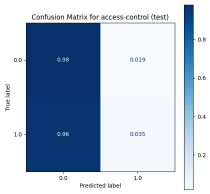
Tabella 1.1: Metriche di Valutazione del Modello Gemini

Il seguente classification report dettaglia le metriche di precision, recall e f1-score per ciascuna delle classi:

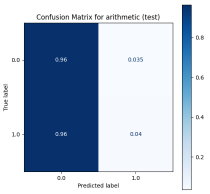
Classe	Precision	Recall	F1-Score	Support
access-control	0.24	0.04	0.06	171
arithmetic	0.19	0.04	0.07	198
other	0.28	0.49	0.36	282
reentrancy	0.41	0.08	0.14	364
unchecked-calls	0.45	0.20	0.28	475
micro avg	0.33	0.19	0.24	1490
macro avg	0.31	0.17	0.18	1490
weighted avg	0.35	0.19	0.21	1490
samples avg	0.18	0.12	0.13	1490

Tabella 1.2: Classification Report del Modello Gemini

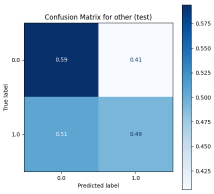
I risultati del modello Gemini mostrano una performance complessivamente modesta nelle metriche di valutazione. L'Accuracy del 27.54% suggerisce che meno di un terzo delle previsioni del modello sono corrette. La Precision micro di 32.71% indica che, tra le istanze classificate positivamente, solo un terzo sono effettivamente corrette, mentre la Precision macro e weighted, rispettivamente 31.37% e 34.90%, mostrano una variabilit nella performance sulle diverse classi. Il Recall micro di 18.72% evidenzia che il modello riesce a identificare meno di un quinto delle istanze rilevanti, con valori macro e weighted simili, indicando che il modello ha difficolt a catturare tutte le istanze corrette. L'F1 Score, che bilancia precision e recall, basso in tutte le versioni (micro: 23.82%, macro: 17.96%, weighted: 20.56%), suggerendo un compromesso non soddisfacente tra la capacit del modello di evitare falsi positivi e falsi negativi. Il report di classificazione per singole classi conferma queste osservazioni, mostrando prestazioni particolarmente deboli in classi come "access-control" e "arithmetic". Questi risultati indicano la necessit di miglioramenti significativi nel modello per una classificazione multilabel pi accurata e bilanciata. Mostriamo ora le confusion matrix per ogni classe per il modello Gemini.



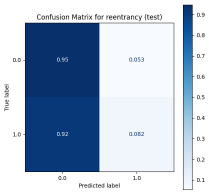
(a) Access-Control



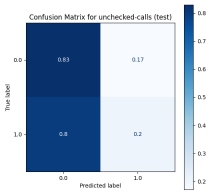
(b) Arithmetic



(c) Other



(d) Reentrancy



(e) Unchecked Calls

Figura 1.1: Confusion Matrices per le diverse classi