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 modelli sul Bytecode

In questa sezione verranno mostrati i risultati ottenuti dai modelli di classificazione utilizzati per la classificazione del bytecode. I modelli utilizzati sono stati allenati sul dataset di training, validati sul dataset di validazione e testati sul dataset di test. I risultati ottenuti sono stati calcolati utilizzando le metriche di valutazione precedentemente descritte.
Per quanto riguarda il bytecode stato dapprima utilizzato un modello BERT ed un modello CodeBert per la classificazione del bytecode utilizzando 512 token in input. Successivamente, poich i risultati migliori sono stati ottenuti dal modello CodeBert, questo stato il modello scelto per gli esperimenti successivi.

BERT

L'accuratezza del modello del 70,48%. La precision, recall e f1-score per ciascuna delle classi sono mostrate nel classification report sottostante.

Class	**Precision**	**Recall**	**F1-Score**	**Support**
access-control	0.87	0.67	0.76	2331
arithmetic	0.88	0.59	0.71	2708
other	0.81	0.73	0.76	4193
reentrancy	0.88	0.78	0.83	4838
unchecked-calls	0.90	0.87	0.88	7276
Micro avg	0.8726	0.7654	0.8155	21346
Macro avg	0.8694	0.7287	0.7895	21346
Weighted avg	0.8724	0.7654	0.8126	21346
Samples avg	0.5200	0.5000	0.5000	21346

Tabella 1.1: Classification Report

CodeBert

L'accuratezza del modello del 72,54%. La precision, recall e f1-score per ciascuna delle classi sono mostrate nel classification report sottostante.

Class	Precision	Recall	F1-Score	Support
access-control	0.87	0.72	0.79	2331
arithmetic	0.81	0.69	0.75	2708
other	0.85	0.73	0.78	4193
reentrancy	0.88	0.81	0.84	4838
unchecked-calls	0.93	0.86	0.89	7276
Micro avg	0.8800	0.7869	0.8309	21346
Macro avg	0.8661	0.7622	0.8104	21346
Weighted avg	0.8787	0.7869	0.8298	21346
Samples avg	0.5400	0.5100	0.5100	21346

Tabella 1.2: Classification Report del modello CodeBert sul bytecode

1.2.2 Risultati sul Codice Sorgente Solidity

Questa sezione presenta i risultati ottenuti nei modelli di classificazione utilizzati che prendevano in input in codice sorgente Solidity. Anche in questo caso, i modelli utilizzati sono stati allenati sul dataset di training, validati sul dataset di validazione e testati sul dataset di test ed i risultati ottenuti sono stati calcolati utilizzando le metriche di valutazione precedentemente descritte.
Allo stesso modo del bytecode, stato dapprima utilizzato un modello BERT ed un modello CodeBert per la classificazione del codice sorgente Solidity utilizzando 512 token in input. Successivamente, poich i risultati migliori sono stati ottenuti dal modello CodeBert, stato utilizzato questo modello per gli esperimenti successivi in cui si utilizzavano porzioni piu ampie dei dati a disposizione.

BERT

CodeBert

DistilBert

CodeBert Concatenazione di due chunk

L'accuratezza del modello del 76,36%. La precision, recall e f1-score per ciascuna delle classi sono mostrate nel classification report sottostante.

Class	Precision	Recall	F1-Score	Support
access-control	0.86	0.78	0.82	2331
arithmetic	0.81	0.81	0.81	2708
other	0.82	0.78	0.80	4193
reentrancy	0.90	0.80	0.85	4838
unchecked-calls	0.94	0.91	0.93	7276
Micro avg	0.8819	0.8342	0.8574	21346
Macro avg	0.8662	0.8172	0.8406	21346
Weighted avg	0.8821	0.8342	0.8571	21346
Samples avg	0.5700	0.5600	0.5500	21346

Tabella 1.3: Classification Report del modello codeBERT con concatenazione di due chunk

CodeBert Concatenazione di tre chunk

L'accuratezza del modello del 79,39%. La precision, recall e f1-score per ciascuna delle classi sono mostrate nel classification report sottostante.

Class	Precision	Recall	F1-Score	Support
access-control	0.87	0.82	0.84	2331
arithmetic	0.90	0.81	0.85	2708
other	0.87	0.79	0.83	4193
reentrancy	0.91	0.84	0.87	4838
unchecked-calls	0.95	0.93	0.94	7276
Micro avg	0.9103	0.8561	0.8824	21346
Macro avg	0.8994	0.8387	0.8677	21346
Weighted avg	0.9093	0.8561	0.8816	21346
Samples avg	0.5900	0.5700	0.5700	21346

Tabella 1.4: Classification Report del modello codeBERT con concatenazione di tre chunk

CodeBert Aggregazione di due chunk

Aggregazione con funzione Max

L'accuratezza del modello del 76.50%. La precisione, il recall e l'F1-score per ciascuna delle classi sono mostrati nel classification report sottostante.

Class	Precision	Recall	F1-Score	Support
access-control	0.87	0.77	0.82	2331
arithmetic	0.83	0.80	0.82	2708
other	0.80	0.80	0.80	4193
reentrancy	0.90	0.81	0.85	4838
unchecked-calls	0.94	0.91	0.93	7276
Micro avg	0.8816	0.8369	0.8586	21346
Macro avg	0.8687	0.8181	0.8422	21346
Weighted avg	0.8821	0.8369	0.8585	21346
Samples avg	0.5700	0.5600	0.5600	21346

Tabella 1.5: Classification Report del modello codeBERT con aggregazione di due chunk

Aggregazione con funzione Mean

L'accuratezza del modello del 75.84%. La precisione, il recall e l'F1-score per ciascuna delle classi sono mostrati nel classification report sottostante.

Class	Precision	Recall	F1-Score	Support
access-control	0.84	0.77	0.81	2331
arithmetic	0.82	0.80	0.81	2708
other	0.81	0.77	0.79	4193
reentrancy	0.90	0.80	0.85	4838
unchecked-calls	0.94	0.90	0.92	7276
Micro avg	0.8830	0.8233	0.8521	21346
Macro avg	0.8657	0.8068	0.8350	21346
Weighted avg	0.8833	0.8233	0.8520	21346
Samples avg	0.5700	0.5500	0.5500	21346

Tabella 1.6: Classification Report del modello con aggregazione di due chunk

CodeBert Aggregazione di tre chunk

Aggregazione con funzione Max

L'accuratezza del modello del 79.08%. La precisione, il recall e l'F1-score per ciascuna delle classi sono mostrati nel classification report sottostante.

Class	Precision	Recall	F1-Score	Support
access-control	0.87	0.83	0.85	2331
arithmetic	0.86	0.84	0.85	2708
other	0.84	0.82	0.83	4193
reentrancy	0.91	0.84	0.87	4838
unchecked-calls	0.95	0.94	0.94	7276
Micro avg	0.8983	0.8671	0.8824	21346
Macro avg	0.8855	0.8524	0.8685	21346
Weighted avg	0.8983	0.8671	0.8822	21346
Samples avg	0.5900	0.5800	0.5700	21346

Tabella 1.7: Classification Report del modello con aggregazione di tre chunk e funzione Max

Aggregazione con funzione Mean
L'accuratezza del modello del 78.97%. La precisione, il recall e l'F1-score per ciascuna delle classi sono mostrati nel classification report sottostante.

Class	Precision	Recall	F1-Score	Support
access-control	0.89	0.82	0.85	2331
arithmetic	0.89	0.81	0.85	2708
other	0.83	0.83	0.83	4193
reentrancy	0.92	0.82	0.86	4838
unchecked-calls	0.95	0.93	0.94	7276
Micro avg	0.9047	0.8566	0.8800	21346
Macro avg	0.8959	0.8411	0.8672	21346
Weighted avg	0.9051	0.8566	0.8797	21346
Samples avg	0.5900	0.5700	0.5700	21346

Tabella 1.8: Classification Report del modello codeBERT con aggregazione mean di tre chunk

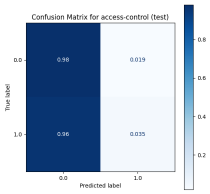
1.2.3 Risultati Gemini

In questa sezione verranno presentati i risultati ottenuti a partire dalle API del modello Gemini. I risultati ottenuti sono stati calcolati utilizzando le metriche di valutazione precedentemente descritte.
L'accuratezza del modello del 27.54%. La precisione, il recall e l'F1-score per ciascuna delle classi sono mostrati nel classification report sottostante.

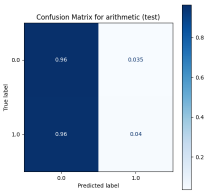
Class	Precision	Recall	F1-Score	Support
access-control	0.24	0.04	0.06	171
arithmetic	0.19	0.04	0.07	198
other	0.28	0.49	0.36	282
reentrancy	0.41	0.08	0.14	364
unchecked-calls	0.45	0.20	0.28	475
Micro avg	0.3271	0.1872	0.2382	1490
Macro avg	0.3137	0.1706	0.1800	1490
Weighted avg	0.3490	0.1872	0.2056	1490
Samples avg	0.1800	0.1200	0.1300	1490

Tabella 1.9: Classification Report del modello Gemini

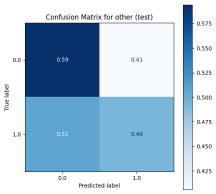
I risultati del modello Gemini mostrano una performance complessivamente modesta nelle metriche di valutazione. L'Accuracy del 27.54% suggerisce che meno di un terzo delle previsioni del modello sono corrette. La Precision micro di 32.71% indica che, tra le istanze classificate positivamente, solo un 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.



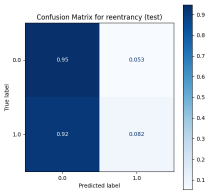
(a) Access-Control



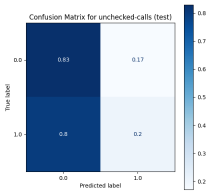
(b) Arithmetic



(c) Other



(d) Reentrancy



(e) Unchecked Calls

Figura 1.1: Confusion Matrices per le diverse classi