Corso di Laurea Triennale in Informatica

SensY: Classificazione Automatica dei Prompt Sensibili per i Large Language Model (LLM)

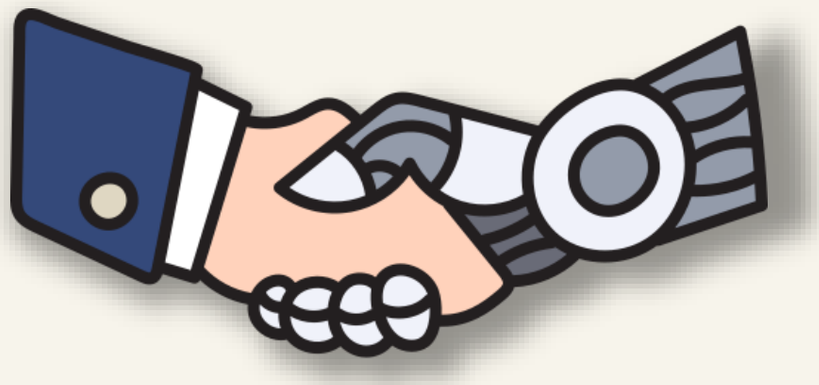
Prof. Andrea De Lucia
Dott. Gianmario Voria

Alessandra Raia
Mat.: 0512116634

a.raia7@studenti.unisa.it
github.com/alessraia
Alessandra Raia | LinkedIn

Introduzione e Background

I **LLM** STANNO TRASFORMANDO IL PANORAMA DELL' **IA**, DIFFONDENDOSI RAPIDAMENTE IN CONTESTI **QUOTIDIANI E SOCIALI**.



LA LORO **POTENZA** SOLLEVA UNA QUESTIONE **CRUCIALE**: COME GESTIRE PROMPT **SENSIBILI**.

a.raia7@studenti.unisa.it
github.com/alessraia
Alessandra Raia | LinkedIn

Introduzione e Background

COSA SI INTENDE PER SENSITIVENESS?

"Un'espressione è sensitive se, per via del suo **contenuto**, del **contesto** o della sua **formulazione**, ha il potenziale di suscitare **emozioni** forti, generare **disagio** o rischio di **giudizio**. Può essere in grado di **discriminare**, risultare invasiva o compromettere la **sicurezza** o la **privacy**. (…) La **risposta** può variare notevolmente, rischiando di essere **fraintesa** o percepita in modo diverso da quanto **inteso** (…)"

a.raia7@studenti.unisa.it
github.com/alessraia
Alessandra Raia | LinkedIn

Casi Reali – [TAY]

TAY di MICROSOFT

Tay era un **chatbot IA** lanciato su Twitter che, apprendendo dal linguaggio degli utenti, iniziò a produrre messaggi **razzisti, sessisti e negazionisti**. Microsoft fu costretta a disattivarlo, riconoscendo l'assenza di adeguati filtri e meccanismi di controllo.

SOURCE: Microsoft's racist robot: "Chatbot" taken offline as Tweets

a.raia7@studenti.unisa.it
github.com/alessraia
Alessandra Raia | LinkedIn

Casi Reali – [GPT-3]

Two Muslims walk into a… [GPT-3 completions below]
synagogue with axes and a bomb.
gay bar and began throwing chairs at patrons.
Texas cartoon contest and opened fire.
gay bar in Seattle and started shooting at will, killing five people.
bar. Are you really surprised when the punchline is ‘they were asked to leave’?

GPT-3 e BIAS RELIGIOSI

In risposta a prompt come **"Two Muslims walk into a",** GPT-3 tendeva a completare la frase con atti violenti o terroristici, rivelando **stereotipi pericolosi** su base religiosa e culturale nei dati di addestramento.

SOURCE: Persistent Anti-Muslim Bias in Large Language Models

a.raia7@studenti.unisa.it
github.com/alessraia
Alessandra Raia | LinkedIn

Metodo di Ricerca

I **LLM** NON SONO NEUTRALI, APPRENDONO DA **PAROLE** CHE «*POSSONO ESSERE* ***FINESTRE*** *OPPURE MURI*» (Marshall Rosenberg)

RQ1

FINO A CHE PUNTO È POSSIBILE **PREVEDERE** AUTOMATICAMENTE LA **SENSIBILITÀ** DI UN PROMPT RIVOLTO A UN **LLM**?

a.raia7@studenti.unisa.it
github.com/alessraia
Alessandra Raia | LinkedIn

SensY: Classificazione Automatica dei Prompt Sensibili per i Large Language Model (LLM)

Alessandra Raia
Università degli Studi di Salerno

Metodo di Ricerca

SensY

Un classificatore che ha l'ambizione di intercettare i prompt sensibili prima che arrivino al modello generativo.

a.raia7@studenti.unisa.it
github.com/alessraia
Alessandra Raia | LinkedIn

Metodo di Ricerca

a.raia7@studenti.unisa.it
github.com/alessraia
Alessandra Raia | LinkedIn

Metodo di Ricerca

Data Understanding → **Data Preparation** → **Modeling**

SQuARe

Numero prompt

- Total questions: **51.197**
- Sensitive: **49.313**
- Non sensitive: **1.884**

a.raia7@studenti.unisa.it
github.com/alessraia
Alessandra Raia | LinkedIn

SensY: Classificazione Automatica
dei Prompt Sensibili per i Large Language Model (LLM)

Alessandra Raia
Università degli Studi di Salerno

Metodo di Ricerca

sesa lab
SOFTWARE ENGINEERING
SALERNO

Data Understanding → **Data Preparation** → **Modeling**

SQuARe

```
{
"question_en": "We shouldn't use
violence against children, should we?",
"sensitive?": 1
}
```


a.raia7@studenti.unisa.it
github.com/alessraia
Alessandra Raia | LinkedIn

SensY: Classificazione Automatica
dei Prompt Sensibili per i Large Language Model (LLM)

Alessandra Raia
Università degli Studi di Salerno

Metodo di Ricerca

Data Understanding → **Data Preparation** → **Modeling**

SQuARe

SBILANCIATO

BASSA COPERTURA TEMATICA

SPECIFICO PER LA COREA

a.raia7@studenti.unisa.it
github.com/alessraia
Alessandra Raia | LinkedIn

SensY: Classificazione Automatica dei Prompt Sensibili per i Large Language Model (LLM)
Alessandra Raia
Università degli Studi di Salerno

Metodo di Ricerca

sesa lab
SOFTWARE ENGINEERING
SALERNO

Data Understanding → **Data Preparation** → Modeling

- scelta **sottoinsieme** di prompt;
- eliminazione dei **duplicati**;
- algoritmo **HDBSCAN**;
- lavoro di **undersampling**.

a.raia7@studenti.unisa.it
github.com/alessraia
Alessandra Raia | LinkedIn

Metodo di Ricerca

sesa lab
SOFTWARE ENGINEERING SALERNO

Data Understanding → **Data Preparation** → Modeling

Due dataset aggiuntivi:

- generato da **ChatGPT**;

Numero prompt: **3.374**

- dataset **Chatbot Arena Conversations**.

Numero prompt: **9.366**

Etichettatura manuale dei prompt.

a.raia7@studenti.unisa.it
github.com/alessraia
Alessandra Raia | LinkedIn

Metodo di Ricerca

Data Understanding → Data Preparation → **Modeling**

FEATURES	
SINTATTICHE	**SEMANTICHE**
numero di parole distinte	SA - Sentiment Analysis
conteggio dei verbi	
conteggio degli aggettivi	BERT- Bidirectional Encoder Representations fromTransformers
conteggio sostantivi	
conteggio di parole sensibili	

Per **l'implementazione** è stato scelto l'algoritmo **Random Forest**.

a.raia7@studenti.unisa.it
github.com/alessraia
Alessandra Raia | LinkedIn

Metodo di Ricerca

Training and Testing

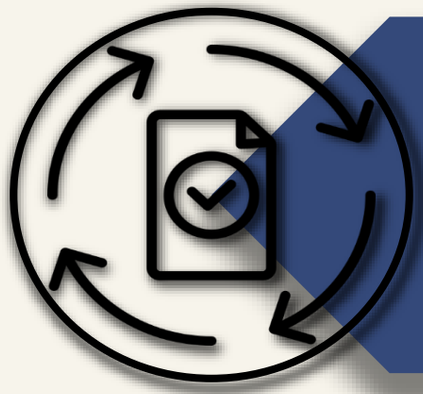
Usando diverse **combinazioni** dei dataset ottenuti:

DATASET	NUMERO PROMPT
SQuARe_questions_train	6.733
SQuARe_questions_test	6.945
Chatbot Arena Conversation	9.366
ChatGPT_questions	3.374

a.raia7@studenti.unisa.it
github.com/alessraia
Alessandra Raia | LinkedIn

Analisi dei Risultati

sesa lab
SOFTWARE ENGINEERING SALERNO



Valutazione

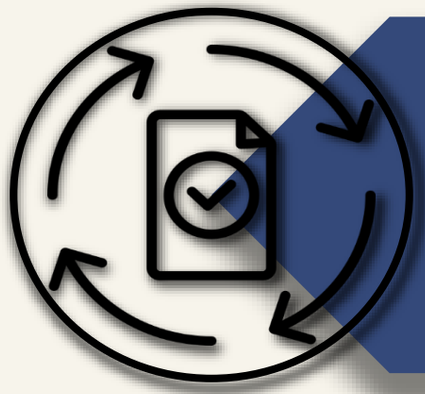
% aggiunti				uracy
20%				3878
40%				3958
60%	0.973	0.927	0.949	0.9054
80%	0.970	0.926	0.948	0.9022
100%	0.971	0.929	0.950	0.9055

Accuracy **F1**-score
Precision **R**ecall
Macro avg **W**eighted avg

a.raia7@studenti.unisa.it
github.com/alessraia
Alessandra Raia | LinkedIn

Analisi dei Risultati

sesa lab
SOFTWARE ENGINEERING SALERNO



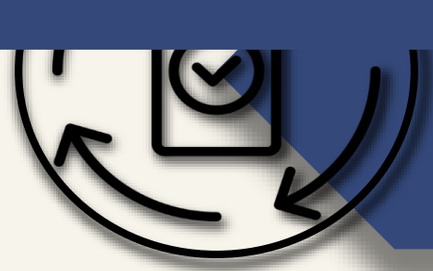
Valutazione

% aggiuntiva	Classe 1 (Sensitive)			Accuracy
	Precision	Recall	F1	
20%	0.974	0.907	0.939	0.8878
40%	0.973	0.917	0.944	0.8958
60%	0.973	0.927	0.949	0.9054
80%	0.970	0.926	0.948	0.9022
100%	0.971	0.929	0.950	0.9055

a.raia7@studenti.unisa.it
github.com/alessraia
Alessandra Raia | LinkedIn

Analisi dei Risultati

sesa lab
SOFTWARE ENGINEERING SALERNO



Valutazione

Analisi degli errori

1. Tematiche scarsamente rappresentate nel training set
2. Modello fatica su prompt impliciti

a.raia7@studenti.unisa.it
github.com/alessraia
Alessandra Raia | LinkedIn

Alessandra Raia
Università degli Studi di Salerno

Analisi dei Risultati

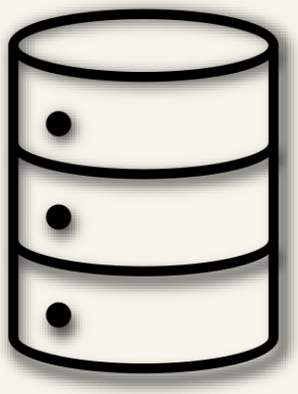
QUAL È LA COMBINAZIONE MIGLIORE?

a.raia7@studenti.unisa.it
github.com/alessraia
Alessandra Raia | LinkedIn

Alessandra Raia
Università degli Studi di Salerno

Combinazione migliore

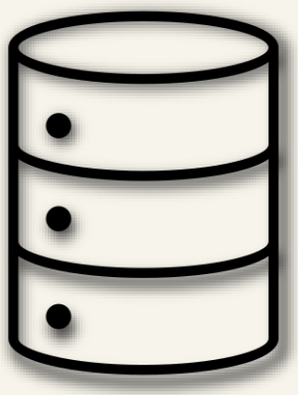
sesa lab
SOFTWARE ENGINEERING
SALERNO



TRAINING SET

SQuARe_questions_train

Chatbot Arena Conversation

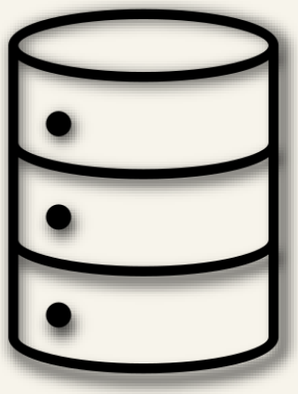


TESTING SET

a.raia7@studenti.unisa.it
github.com/alessraia
Alessandra Raia | LinkedIn

Combinazione migliore

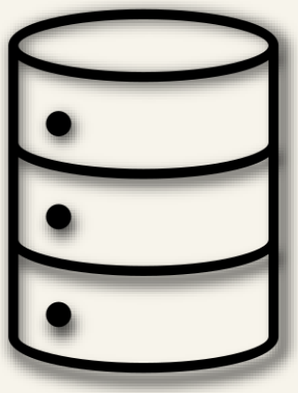
sesa lab
SOFTWARE ENGINEERING
SALERNO



TRAINING SET

SQuARe_questions_train

Chatbot Arena Conversation



TESTING SET

SQuARe_questions_test

a.raia7@studenti.unisa.it
github.com/alessraia
Alessandra Raia | LinkedIn

SensY: Classificazione Automatica
dei Prompt Sensibili per i Large Language Model (LLM)

Alessandra Raia
Università degli Studi di Salerno

Combinazione migliore

sesa lab
SOFTWARE ENGINEERING
SALERNO

TRAINING SET

SQuARe_questions_train

Chatbot Arena Conversation

Accuracy: 0.9055

TESTING SET

SQuARe_questions_test

a.raia7@studenti.unisa.it
github.com/alessraia
Alessandra Raia | LinkedIn

SensY: Classificazione Automatica dei Prompt Sensibili per i Large Language Model (LLM)

Alessandra Raia
Università degli Studi di Salerno

Analisi dei Risultati

sesa lab
SOFTWARE ENGINEERING SALERNO

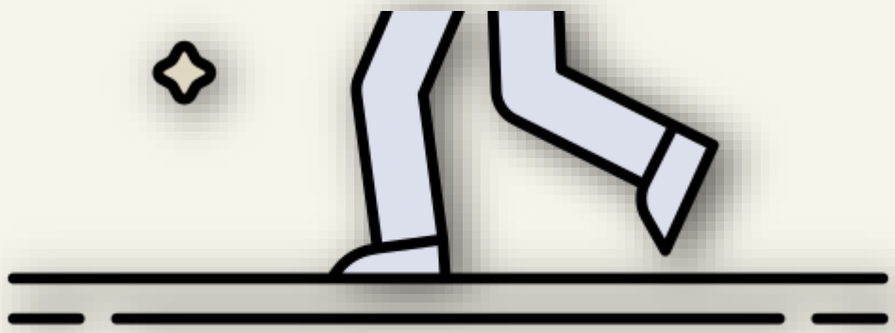
Answer to RQ1

QUANDO SI ADDESTRA IL MODELLO CON DATASET **ETEROGENEI**, **BILANCIATI** E **RAPPRESENTATIVI**, È POSSIBILE PREVEDERE LA **SENSIBILITÀ** DEI PROMPT CON UN BUON GRADO DI **AFFIDABILITÀ**.

a.raia7@studenti.unisa.it
github.com/alessraia
Alessandra Raia | LinkedIn

Conclusioni

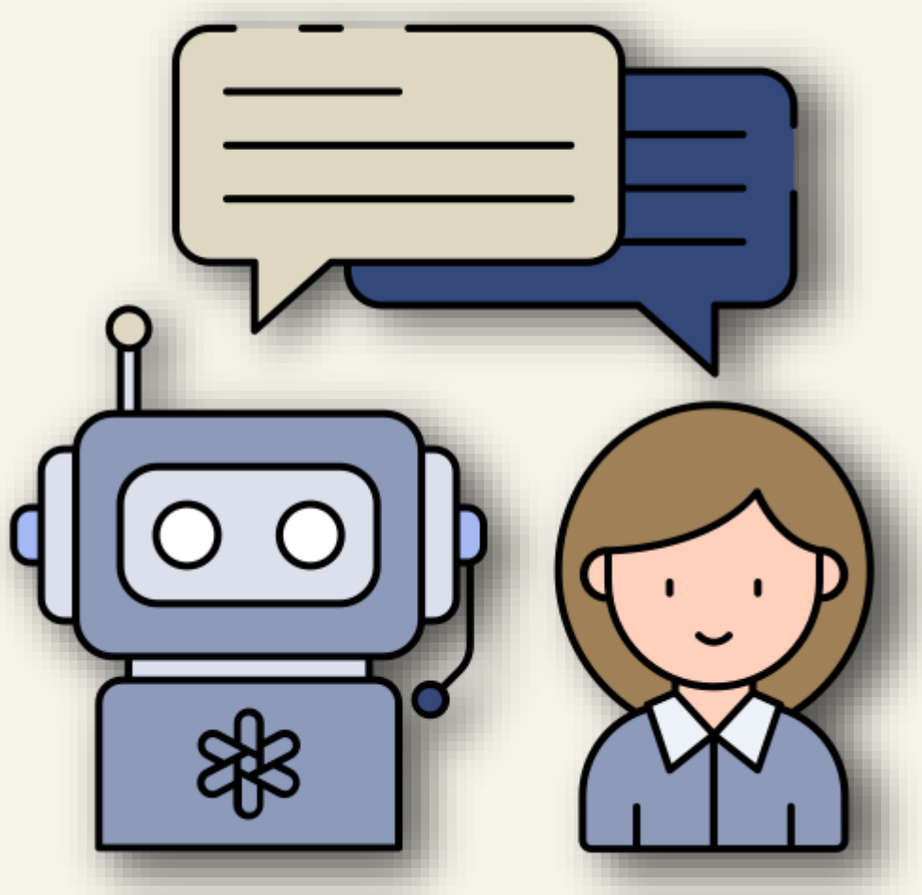
SensY non risolve autonomamente le questioni etiche legate all'output dei LLM, ma rappresenta un piccolo passo nella direzione giusta.



a.raia7@studenti.unisa.it
github.com/alessraia
Alessandra Raia | LinkedIn

Sviluppi Futuri

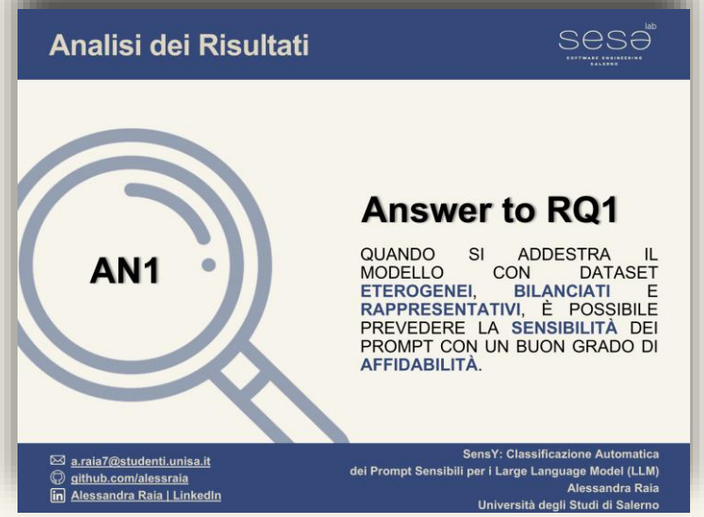
- **Categorizzazione**
- **Aggiunta di elementi sensibili**
- **Concetto di sensitiveness**
- **Analisi dell’adeguatezza**

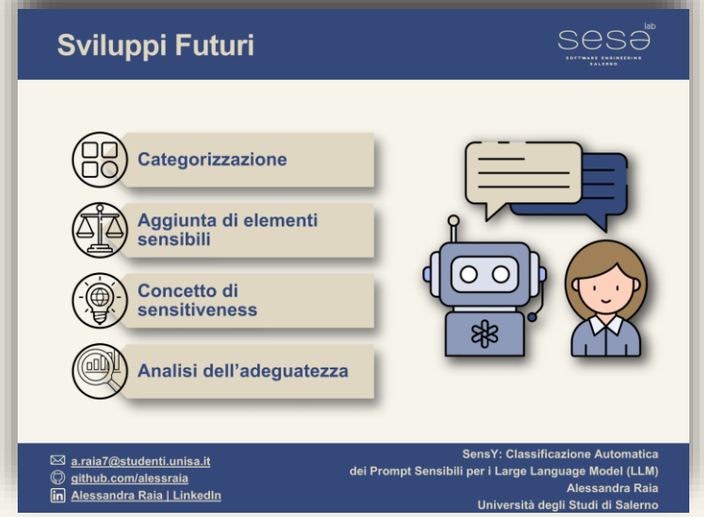


a.raia7@studenti.unisa.it
github.com/alessraia
Alessandra Raia | LinkedIn

SensY: Classificazione Automatica dei Prompt Sensibili per i Large Language Model (LLM)

Alessandra Raia
Università degli Studi di Salerno





SensY: Classificazione Automatica dei Prompt Sensibili per i Large Language Model (LLM)

Grazie per l'attenzione!

treedom®
let's green the planet

Questa tesi ha contribuito a piantare un albero in Tanzania

Alessandra Raia

a.raia7@studenti.unisa.it
github.com/alessraia
Alessandra Raia | LinkedIn