

Università degli Studi di Padova – Dipartimento di Ingegneria dell'Informazione
Corso di Laurea in Ingegneria Informatica

Relazione per la prova finale
***«Un Framework Comportamentale per il Rilevamento
di Bot basato sulla funzione Log-Likelihood della
Dirichlet-Multinomiale»***

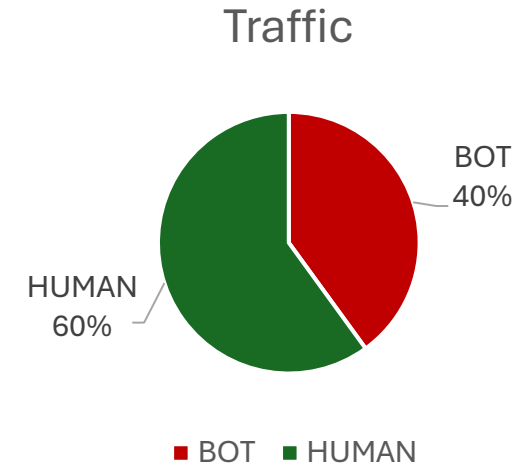
Relatore: Prof. Mauro Migliardi

Laureando: Silvio Genesin
Matricola n.: 2101832

Padova, 21/07/2026

Il Problema

- Nel 2025 il 60% del traffico internet è generato da esseri umani.
- Il 40 % rimanente è attribuito a bot.
- Si stimano perdite di decine di miliardi di dollari.



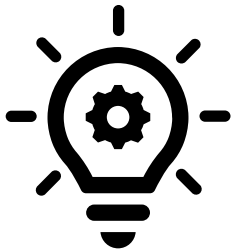
Gli agenti basati su LLM rendono sempre più difficile distinguere tra bot ed esseri umani.

- Rilevamento a livello di rete → Aggirabile tramite spoofing
- Tecniche Deception-based → Non affidabile su larga scala
- Biometria comportamentale → Testata solo su bot semplici, non LLM

Falliscono tutte contro bot che simulano il comportamento umano

I metodi attuali osservano COME si comporta un utente

Questo lavoro osserva COSA l'utente cerca



Cosa rende umano l'essere umano?

La capacità di spaziare tra temi diversi in una sola sessione → Eterogeneità

Misuriamo l'eterogeneità con un parametro chiamato sovradisersione (overdispersion)

Perché la Dirichlet Multinomiale?

Sovradispersione = eterogeneità del comportamento osservato

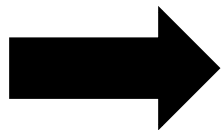
Multinomiale $\longrightarrow$ Assume probabilità fisse

+

Dirichlet $\longrightarrow$ Permette probabilità variabili

=

Multinomiale di Dirichlet



$p \sim \text{Dirichlet}(\alpha)$

α controlla quanto il comportamento osservato
è concentrato o disperso .

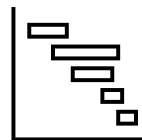
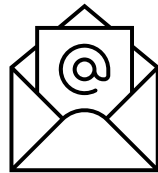
La Dirichlet-Multinomiale consente di modellare l'eterogeneità tra sessioni diverse.

La misura della sovradisersione

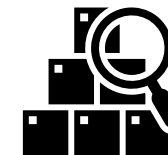
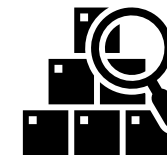
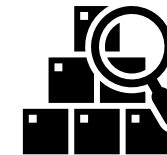
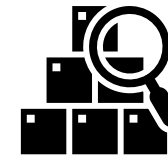
$$\Psi = \frac{1}{\sum_k \alpha_k}$$

Ψ è la sovradisersione, e misura l'eterogeneità.

Valori di Ψ grandi rappresentano una sessione con
temi diversi → UMANO



Valori di Ψ verso lo 0 rappresentano sessioni più
concentrate → BOT



Il Problema Computazionale

Quando $\Psi \rightarrow 0$ il calcolo della log-likelihood diventa instabile

Differenza tra due termini enormi è inaffidabile

$\Psi \rightarrow 0$ è il regime dei bot, quindi il più importante da rilevare

Problema identificato da Yu e Shaw, che però non lo risolvono totalmente

Serve una soluzione più stabile ed efficiente

La Procedura Languasco-Migliardi

Soluzione basata sulla formula di Eulero-MacLaurin

Consente approssimazione stabile della log-likelihood

Riduce errore numerico e complessità computazionale da quadratica a quasi lineare

Funziona anche dove prima il calcolo non partiva nemmeno

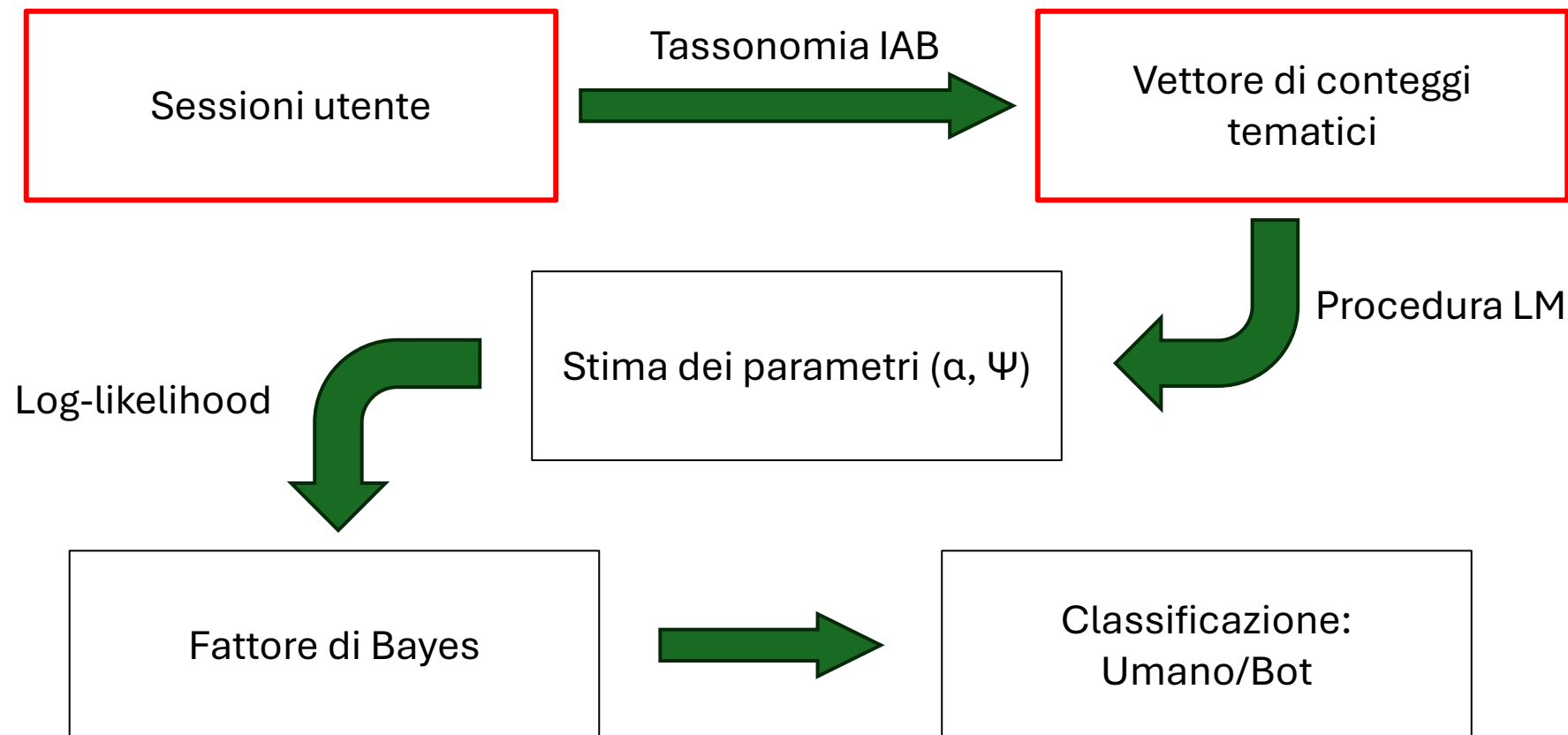
Rende possibile applicare la DM a grandi dataset reali

Lavori correlati

- Text mining (LDA, burstiness lessicale):
modellazione dell'eterogeneità tematica.
- Analisi del traffico IoT:
analisi di dati overdispersi
- Heard e Rubin-Delanchy:
rilevamento di anomalie in cybersecurity.
- Nessuna applicazione alla bot detection tematica.

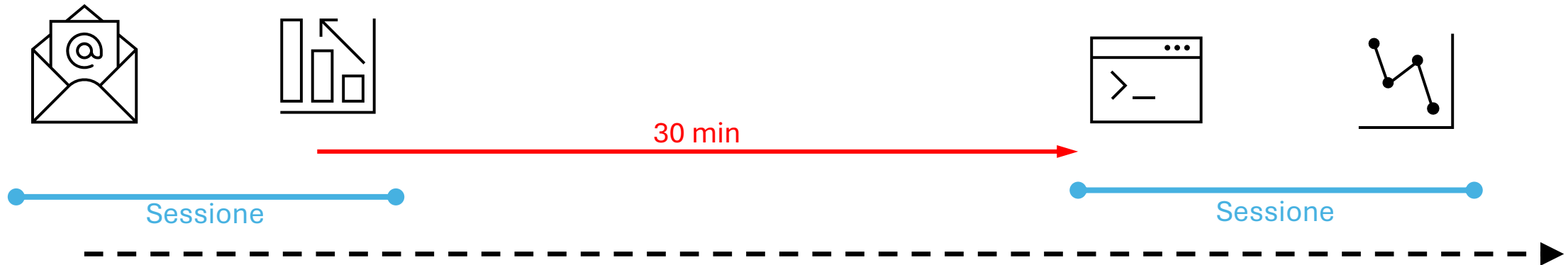
DM validata in ambiti affini, qui viene usata per la prima volta in bot detection basata sui temi delle sessioni

Panoramica del Framework



Dalla Sessione al Vettore Tematico

Sessione utente = Attività in una finestra temporale



- Scelta della tassonomia per catalogare le attività di una sessione
(Es. IAB ~700 categorie)

K categorie che corrispondono a n_k componenti del vettore $\mathbf{n}$

- Inserimento del numero di apparizioni delle categorie nel vettore $\mathbf{n} = (n_1, \dots, n_k)$

Stima dei Parametri e Bayes Factor

α è ricavato dall'algoritmo iterativo di Minka

Stima sotto due ipotesi: α_H e α_B rispettivamente umano e bot.

Calcolo il Bayes Factor: $\log \text{BF}(\mathbf{n}) = \log P(\mathbf{n} \mid \hat{\alpha}_H) - \log P(\mathbf{n} \mid \hat{\alpha}_B),$

$\text{BF} > 0 \rightarrow$ più probabile **umano**

$\text{BF} < 0 \rightarrow$ più probabile **bot**

Bayes Factor più robusto di una soglia su Ψ : si adatta a sessioni di lunghezza e strutture diverse

I limiti Principali

Agenti LLM con obiettivi eterogenei



Pattern tematici simili a quelli umani

Il problema della «ground truth»



Difficoltà nel calibrare il Bayes factor

Sessioni troppo corte



Scarsità di dati e BF meno affidabile

Privacy e Posizionamento

Bilanciare sicurezza e privacy è un tema regolato da norme europee (GDPR, AI act)



Questo framework è meno invasivo rispetto ad altri approcci:

- Usa solo metadati aggregati
- Nessuna biometria comportamentale
- I dati necessari sono già presenti nei log dei server



Coerente con Privacy by Design

Rischi non eliminati, ma ridotti.

Non è una soluzione standalone, si inserisce in un sistema a più livelli

Conclusioni e sviluppi futuri

- Affronta il gap della dispersione tematica nelle sessioni
- Ciò è possibile grazie alla procedura LM
- Limite principale: gli agenti LLM restano il caso più difficile
- Prossimo passo: Validazione empirica
- Integrabile in un sistema ensemble

**La dispersione tematica rappresenta una nuova
frontiera per la bot detection**