MODULO 09 · S4.1 · AGOSTO 2026

AI Applicata II: NLP e Analisi Investigativa

Interrogare centinaia di migliaia di messaggi:
estrarre entità, relazioni e anomalie dal linguaggio

NER • topic modeling • social network analysis • RAG

Corso di Laurea Magistrale in Sicurezza Informatica

Al termine del modulo lo studente sarà in grado di

01 · ESTRARRE

Entità (persone, luoghi, organizzazioni, importi) da chat, email e documenti con tecniche di NER, anche multilingue

02 · RICOSTRUIRE

La rete relazionale di un corpus di comunicazioni e interpretarne le metriche (centralità, comunità, broker)

03 · INTEGRARE

LLM nel workflow DFIR per sintesi e ricerca semantica, in ambienti controllati e con tracciabilità completa

Vincolo probatorio. L'NLP fornisce indicazioni di lettura. In relazione entra sempre il messaggio originale, identificato e verificato; le parafrasi del modello non si citano.

SCHEMA DEL MODULO

Dal testo grezzo alla rete di relazioni

01 · ESTRAZIONE

Strutturare il testo

- NER
- Topic modeling
- Normalizzazione

→

02 · SEMANTICA

Cercare il significato

- Embedding
- Ricerca semantica
- Corpora multilingue

→

03 · RELAZIONI

Chi parla con chi

- Grafi relazionali
- Centralità
- Comunità

→

04 · AUTORIA E LLM

Le frontiere

- Stilometria
- LLM nel DFIR
- RAG e verificabilità

Le quattro aree si susseguono nelle slide del modulo nello stesso ordine dello schema: ciascuna presuppone la precedente.

Dal testo grezzo alla struttura interrogabile

Tecnica	Che cosa produce	Cautela
Named Entity Recognition	Persone, luoghi, organizzazioni, date, importi	Soprannomi e gergo spesso sfuggono
Topic modeling	Temi ricorrenti del corpus e loro evoluzione temporale	I topic sono statistici: l'etichetta la dà l'analista
Ricerca semantica	Recupero per affinità di significato	Dichiarare il modello di embedding
Analisi multilingue	Trattamento uniforme di più lingue	Qualità da validare per lingua

Pipeline documentata

Preprocessing, modello e versione, parametri, data di esecuzione: ogni passo riportato in relazione, perché l'esito dipende da ciascuno.

Dal risultato al reperto

Ogni entità o tema rilevante rimanda, con identificativo univoco, ai messaggi che lo fondano: in relazione si citano questi messaggi.

La NER in pratica: un messaggio, annotato

MESSAGGIO (fittizio, dal caso d'esercitazione)

«Di' al **Lungo** [PER?] che venerdì porto i **15k** [MONEY] al solito posto vicino al **porto di Bari** [LOC], stesso orario della volta con **Sandro** [PER].»

Il modello estrae persone, luoghi, importi e date e li **normalizza** (venerdì → data assoluta dal contesto del messaggio), popolando un indice interrogabile, per esempio sulle menzioni di «Sandro», sugli importi sopra soglia o sui luoghi citati.

Dove la NER inciampa

- **Soprannomi**: «il Lungo» è una persona, ma il modello generico spesso non lo vede
- **Gergo**: «15k», «il posto»: entità implicite che richiedono contesto
- **Coreferenze**: collegare «lui», «il Lungo» e un nome anagrafico è un compito ulteriore, da riscontrare a mano

Rimedio: modelli adattati al dominio + revisione umana delle entità a bassa confidenza.

Regola di tracciabilità: ogni entità nell'indice conserva il riferimento al messaggio d'origine, che è la fonte da citare in relazione.

Interpretare i risultati del topic modeling

Il topic modeling raggruppa i documenti per co-occorrenza di termini, restituendo per ogni «tema» le parole più caratteristiche e i documenti più rappresentativi. Su un corpus di decine di migliaia di email è la mappa tematica che nessun umano potrebbe tracciare a mano.

```
Topic 7 (esempio su corpus Enron):
«reserve, mark-to-market, SPE, hedge,
quarter, restate» (412 email, picco Q3)
```

Il modello consegna **parole e raggruppamenti**; l'etichetta («contabilità delle partite fuori bilancio») la assegna l'analista, leggendo i documenti rappresentativi. Per questo in relazione l'etichetta va motivata con i documenti letti.

Uso investigativo corretto

- **Orientamento**: quali temi esistono, quanto pesano, come evolvono nel tempo
- **Selezione**: da quale sottoinsieme iniziare la lettura umana
- **Anomalia**: un tema che appare o scompare in coincidenza con i fatti d'interesse

In relazione si scrive «il topic ha indirizzato verso i documenti X, Y, Z, dai quali risulta…»; va evitata la formula «il topic prova che…».

Il significato come geometria: gli embedding

Un embedding traduce un testo in un **vettore in uno spazio a centinaia di dimensioni**, costruito in modo che testi di significato simile finiscano vicini. Su questo si basa la **ricerca semantica**: cercando «il pagamento di cui parlavamo» si trova anche «quella cosa da sistemare», perché i due testi, pur senza parole in comune, occupano la stessa regione dello spazio.

Per l'investigatore la ricerca passa dalle **keyword** (che i soggetti evitano con gerghi e codici) alla **somiglianza di significato**. I limiti vanno dichiarati: la soglia di «vicinanza» è una scelta dell'analista; i modelli risentono della lingua e del dominio di addestramento; la vicinanza semantica serve soltanto a **selezionare i testi** da leggere.

In relazione: modello e versione, soglia usata, e la precisazione che la ricerca semantica ha orientato la revisione umana senza sostituirla.

ESEMPIO · IL LINGUAGGIO CIFRATO

In 80.000 messaggi, nessuna occorrenza di termini legati ai pagamenti illeciti: i soggetti parlano di «documenti», «pratiche», «caffè».

La ricerca semantica a partire da tre messaggi noti recupera conversazioni dallo stesso profilo semantico: contesti di scambio, quantità, urgenza. Leggendo le conversazioni recuperate, l'analista riconosce il codice.

APPROFONDIMENTO · CORPORA MULTILINGUE

La traduzione nelle indagini multilingue

Il problema operativo

I corpora reali mescolano lingue, dialetti, traslitterazioni e **code-switching** nella stessa frase. Due passi sono automatizzabili: l'identificazione della lingua per segmento e una traduzione di cortesia per il triage.

Embedding cross-lingua

I modelli multilingue proiettano lingue diverse **nello stesso spazio semantico**: una ricerca in italiano trova messaggi pertinenti in altre lingue senza tradurre prima il corpus. La qualità varia tra lingue ad alte e basse risorse e va dichiarata.

Il valore probatorio

La traduzione automatica serve all'analisi; ciò che entra nel fascicolo richiede un **traduttore o interprete qualificato**. La macchina sbaglia soprattutto su gergo, ironia e riferimenti culturali, spesso decisivi per il significato.

ESEMPIO · Frodi transnazionali, 40.000 messaggi in quattro lingue: dopo identificazione delle lingue, traduzione per il triage e indicizzazione, gli analisti selezionano 240 conversazioni. Si deposita solo la versione del traduttore giurato.

Chi parla con chi: la struttura come indizio

Dai metadati delle comunicazioni (mittente, destinatario, momento, frequenza) si costruisce un grafo: nodi le persone, archi le relazioni. La struttura emerge anche dove il contenuto è cifrato o reticente.

Metriche e loro lettura investigativa

- **Centralità di grado**: chi comunica di più (non necessariamente chi decide)
- **Betweenness**: i broker, snodi tra gruppi altrimenti separati
- **Comunità**: sottogruppi coesi e loro confini
- **Evoluzione temporale**: ingressi, uscite e silenzi dopo un evento

Limiti interpretativi

Il grafo descrive soltanto il corpus acquisito: i canali mancanti ne deformano la struttura.

Dalla centralità si può ipotizzare un ruolo gerarchico, da corroborare con il contenuto dei messaggi.

In relazione ogni metrica va accompagnata dalla domanda investigativa a cui risponde.

APPROFONDIMENTO · METRICHE DI RETE

Le centralità con un esempio numerico

Rete fittizia di 5 soggetti; archi = comunicazioni dirette: A–B, A–C, B–C, C–D, D–E.

Nodo	Grado	Betweenness (cammini mediati)
A	2	0 (mai di passaggio)
B	2	0 (mai di passaggio)
C	3	Alta: ogni cammino tra {A,B} e {D,E} passa da C
D	2	Media: unico accesso a E
E	1	0 (terminale)

Lettura investigativa

Il grado di C supera di poco quello degli altri, ma C è **lo snodo**: rimuoverlo spezza la rete in due. È il profilo tipico del broker.

E, periferico e silenzioso, raggiungibile solo via D, è il pattern tipico della figura schermata: pochi contatti, tutti mediati.

La posizione nella rete non basta a stabilire la gerarchia: l'ipotesi di ruolo si corrobora con i contenuti.

Il limite del perimetro: la rete descrive i canali acquisiti. Se C ed E comunicano su un canale non nel corpus, la «mediazione» di D è un artefatto del perimetro, e va sempre dichiarato.

APPROFONDIMENTO · STILOMETRIA

Authorship attribution: metodo e limiti

La stilometria misura le regolarità inconsce della scrittura: frequenze delle **parole funzionali** (congiunzioni, preposizioni), lunghezze di frase, punteggiatura, n-grammi di caratteri, errori ricorrenti. Su questi profili, i classificatori confrontano un **testo contestato** con campioni noti dei candidati autori.

I limiti vanno dichiarati: servono **campioni ampi e dello stesso registro** (le chat non si confrontano con le relazioni di lavoro); il risultato è probabilistico e misura la **compatibilità fra ipotesi alternative**; lo stile cambia con contesto, tempo ed età; l'imitazione deliberata è un'ipotesi da escludere espressamente.

In relazione: metodo, corpus di riferimento, tasso d'errore su casi noti e conclusione probabilistica rispetto a ipotesi esplicite.

ESEMPIO · LA LETTERA ANONIMA AZIENDALE

Una lettera diffamatoria circola in azienda. Tre dipendenti hanno movente; di ciascuno esistono email di pari registro.

L'analisi stilometrica indica forte compatibilità con uno dei tre e incompatibilità con gli altri due. La relazione lo formula come rapporto di verosimiglianza fra le tre ipotesi. Il dato orienta l'indagine interna, che trova poi il riscontro decisivo nei log di stampa (modulo 3): l'attribuzione definitiva si fonda su questo artefatto.

LLM NEL WORKFLOW DFIR

L'LLM come assistente di lettura

Usi appropriati

- Sintesi orientativa di thread lunghi, con rimando ai messaggi originali
- Ricerca semantica e riformulazione di query
- Supporto alla redazione, su contenuti già verificati dal perito
- RAG su corpus del caso: risposte ancorate a documenti citabili

Usi inappropriati

- Citare in relazione testo generato come se fosse del corpus
- Delegare valutazioni di rilevanza o attribuzioni di responsabilità
- Caricare reperti reali su servizi esterni non controllati
- Usare output non riproducibili senza fissare modello e parametri

Requisiti di impiego

- Ambiente locale o istanza dedicata e documentata
- Log integrale di prompt e risposte
- Versione del modello e parametri fissati
- Verifica umana di ogni affermazione riusata

APPROFONDIMENTO · RETRIEVAL-AUGMENTED GENERATION

RAG: ancorare l'LLM ai documenti del caso

Nel RAG un motore di ricerca semantica recupera prima i passaggi pertinenti dal corpus del caso; l'LLM formula poi la risposta **sulla base di quei passaggi**, citandoli.

1. La domanda viene trasformata in embedding e confrontata con l'indice del corpus
2. I passaggi più vicini (con i loro identificativi) vengono recuperati
3. Il modello genera la risposta vincolata a quei passaggi, con i riferimenti

Vantaggio forense: ogni affermazione è **riconducibile a un documento citato** e si verifica aprendo quel documento.

ESEMPIO · DOMANDA ANCORATA

«Chi menziona il conto estero prima di marzo?» → il sistema restituisce: «Due messaggi: [MSG-0412] del 12/02 (mittente X) e [MSG-0788] del 27/02 (mittente Y)», con i passaggi testuali.

L'analista apre MSG-0412 e MSG-0788, li verifica e cita in relazione **i due messaggi originali**.

Il RAG riduce le allucinazioni senza eliminarle: il modello può fraintendere un passaggio recuperato o combinarlo male. Il recupero ha inoltre falsi negativi: un documento che l'indice non trova resta escluso dalla risposta, anche se esiste.

LABORATORIO · LETTURE · VERIFICA

Il corpus Enron come palestra di metodo

Laboratorio

- NER e topic modeling sul **corpus Enron**
- Grafo relazionale e centralità
- Anomalie e **validazione manuale** su campione

`DELIVERABLE: dossier analitico con rimandi ai messaggi`

Letture obbligatorie

- Jurafsky & Martin, *Speech and Language Processing*, capitoli su NER e topic model
- Articoli FSI Digital Investigation su NLP forense
- Documentazione spaCy e BERTopic

Verifica

Si valuta la tracciabilità: ogni affermazione del dossier deve risalire a messaggi identificati; il protocollo di validazione (modulo 8) è parte del voto.

Prossimo modulo: S4.2 LLM, RAG e visione artificiale nelle indagini.

> In giudizio il perito depone sui messaggi letti e verificati, che l'NLP ha aiutato a individuare.

Grazie per l'attenzione

Vincenzo Calabrò

Domande, materiali del corso e contatti sul sito del docente

vincenzocalabro.it