



UNIVERSITA' DEGLI STUDI DI
NAPOLI FEDERICO II

Scuola Politecnica e delle Scienze di Base
Corso di Laurea Magistrale in Ingegneria Informatica

Tesi di Laurea Magistrale in Software Security

*Un Framework di Cyber Threat
Intelligence per l'analisi semantica
di campagne di disinformazione*

Anno Accademico 2024/2025

Relatore

Ch.mo Prof. Roberto Natella

Correlatore

Ing. Vittorio Orbinato

Candidato

Francesca Di Martino

matr. M63001478

Alla mia famiglia...

Abstract

La crescente diffusione della disinformazione impone approcci innovativi che vadano oltre la semplice classificazione delle notizie in vere o false. Questa tesi introduce un framework di Cyber Threat Intelligence semantica che adatta OpenCTI dall'analisi di minacce informatiche tradizionali all'intelligence narrativa per l'attribuzione automatica di campagne di disinformazione.

L'innovazione principale consiste nella reinterpretazione degli oggetti STIX per rappresentare narrative: gli oggetti pensati per descrivere attacchi informatici vengono utilizzati creativamente per modellare relazioni semantiche nel formato soggetto-verbo-oggetto. Questa trasformazione concettuale è implementata attraverso estensioni custom STIX 2.1 che arricchiscono ogni entità con metadati semantici specifici, preservando la compatibilità con l'ecosistema OpenCTI esistente. Il sistema implementa un'architettura modulare Producer-Consumer-Generatore: il Producer estrae automaticamente triple semantiche da contenuti testuali tramite LLM e crea un knowledge graph strutturato secondo lo standard STIX caricato in OpenCTI. Il Consumer sincronizza queste

informazioni semantiche e implementa un classificatore multi-componente per l'attribuzione automatica. Il modulo Generatore estende la pipeline creando contenuti sintetici di disinformazione per testare la robustezza del sistema e simulare evoluzioni narrative future. L'architettura relazionale sviluppata in OpenCTI abilita navigazione esplorativa attraverso query semantiche, permettendo agli analisti di mappare istantaneamente l'intero panorama narrativo di una campagna ed esplorare collegamenti strategici attraverso un'interfaccia interattiva. La validazione sperimentale su FakeCTI (12.155 articoli, 43 campagne) presenta tre configurazioni: classificazione su dati reali (82,76% accuratezza), classificazione su contenuti generati artificialmente (74,14% accuratezza), e confronto con approcci tradizionali (FakeBERT), dimostrando la superiorità dell'attribuzione semantica rispetto alla detection binaria. Il framework trasforma il paradigma da 'questa notizia è falsa?' a 'chi ha orchestrato questa disinformazione e con quale strategia?', rappresentando l'evoluzione da detection reattiva a strategic intelligence proattiva. Questo approccio abilita l'attribuzione degli attori, l'identificazione di campagne coordinate e lo sviluppo di contromisure specifiche attraverso una piattaforma di threat intelligence strutturata e condivisibile.

Contents

Abstract	ii
1 Introduzione alla Disinformazione	1
1.1 Il problema della disinformazione nell'era digitale . . .	1
1.2 Definizione, tipologie, esempi e impatto della disinformazione	4
1.3 Struttura, meccanismi narrativi e amplificazione delle fake news	6
1.4 Obiettivi e contributi della tesi	8
2 Background e tecnologie di riferimento	10
2.1 Cyber Threat Intelligence (CTI)	10
2.1.1 Definizione e principi fondamentali	11
2.1.2 Indicatori tradizionali e la piramide del dolore .	13
2.1.3 Estensione della CTI al dominio della disinformazione	14
2.2 OpenCTI e il modello STIX	16
2.2.1 Introduzione a OpenCTI: obiettivi e architettura	17
2.2.2 Modello STIX 2.1: struttura e oggetti principali	18

2.2.3	Limitazioni nella rappresentazione della disinformazione	19
2.3	Large Language Models per l'estrazione semantica . .	21
2.3.1	Panoramica sui modelli linguistici LLM: Llama e DeepSeek	21
2.3.2	Prompt engineering, few-shot learning e design del template	22
2.4	Dataset FakeCTI: struttura e utilizzo	23
3	Architettura funzionale del sistema	27
3.1	Riuso creativo degli oggetti STIX per la disinformazione	28
3.1.1	Mappatura concettuale da STIX a narrativa . .	29
3.1.2	Implementazione tecnica con estensioni STIX .	29
3.2	Navigazione semantica e correlazione tramite relazioni STIX	31
3.2.1	Modellazione delle relazioni	31
3.2.2	Navigazione esplorativa attraverso OpenCTI . .	33
3.3	Architettura Funzionale: Producer, Consumer e Generatore	37
3.3.1	Panoramica generale dell'architettura	37
3.3.2	Il Modulo Producer: estrazione e caricamento su OpenCTI	37
3.3.3	Il Modulo Consumer: classificazione e attribuzione	41
3.3.4	Il Modulo Generatore: espansione narrativa per validazione	44

4	Sperimentazione e Validazione	46
4.1	Setup Sperimentale	46
4.1.1	Ambiente tecnico (Docker, hardware, software)	46
4.1.2	Preparazione dataset e metriche di valutazione	48
4.2	Esperimento 1: Baseline su Dati Reali	49
4.2.1	Pipeline di classificazione tradizionale	50
4.2.2	Confronto tra modelli di estrazione: DeepSeek vs LLaMA	53
4.2.3	Caso Studio 2: Articolo ID 11837 (Campagna "Doctors Found Dead")	57
4.2.4	Caso di Studio 3: Campagna Russian Troll . .	59
4.2.5	Risultati e analisi per campagna	63
4.3	Esperimento 2: Classificazione con Generazione Auto- matica	66
4.3.1	Integrazione modulo generatore ed espansione dataset	68
4.3.2	Risultati classificazione	69
4.3.3	Analisi comparativa dei risultati dei primi 2 es- perimenti	71
4.4	Esperimenti con FakeBERT	74
4.4.1	Validazione Empirica attraverso confronto con FakeBERT	74
5	Conclusioni	80
5.1	Risultati Principali	80
5.2	Potenziale Applicativo	81

5.3	Sviluppi Futuri	83
-----	---------------------------	----

Chapter 1

Introduzione alla Disinformazione

1.1 Il problema della disinformazione nell'era digitale

L'avvento dell'era digitale ha trasformato radicalmente il panorama informativo, creando un sistema di comunicazione caratterizzato da velocità immediate, diffusione globale e barriere di accesso quasi inesistenti. Questa rivoluzione, pur avendo ampliato l'opportunità di partecipazione al dibattito pubblico, ha al contempo favorito la diffusione di contenuti non verificati e, in alcuni casi, deliberatamente fuorvianti.

Le caratteristiche distintive dell'era digitale che facilitano la dif-

fusione della disinformazione sono molteplici e strettamente interconnesse. In primo luogo la velocità di propagazione rappresenta l'aspetto più critico: l'informazione si diffonde in tempo reale attraverso reti globali interconnesse, consentendo a una notizia falsa di raggiungere migliaia, se non milioni, di persone prima che ne venga verificata la veridicità, contestualizzata o smentita dalle autorità competenti.

Un secondo elemento rilevante è rappresentato dall'effetto amplificazione. Gli algoritmi delle piattaforme social, progettati per massimizzare l'engagement degli utenti e prolungarne il tempo di permanenza sulle piattaforme, favoriscono la diffusione di contenuti emotivamente coinvolgenti, spesso caratterizzati da toni sensazionalistici indipendentemente dalla loro accuratezza o fondatezza. Tale meccanismo premia dunque contenuti capaci di suscitare reazioni intense, piuttosto che informazioni verificabili.

Infine, il fenomeno delle cosiddette 'Echo Chambers' rappresenta il terzo elemento problematico. Queste 'camere dell'eco' si formano grazie agli algoritmi di personalizzazione, che, basandosi su profili comportamentali dettagliati, tendono a mostrare agli utenti contenuti in linea con le loro opinioni già esistenti. Questo limita l'esposizione a punti di vista diversi e rafforza le credenze consolidate, alimentando la divisione sociale. Questa dinamica rende particolarmente difficile correggere le false credenze una volta che si sono radicate.

Uno studio condotto da ricercatori del MIT e pubblicato su Science

nel 2018 ha evidenziato come le notizie false si diffondano sui social media in modo significativamente più rapido e ampio rispetto a quelle vere. Analizzando circa 126.000 casi di condivisione di notizie su Twitter tra il 2006 e il 2017, i ricercatori hanno scoperto che le informazioni false raggiungono un pubblico più vasto e si propagano circa sei volte più velocemente rispetto alle informazioni verificate. In particolare, le notizie false di natura politica risultano ancora più virali rispetto ad altre tipologie di falsità, come quelle riguardanti terrorismo, scienza o finanza. I dati indicano che le notizie false hanno una probabilità del 70% maggiore di essere ritwittate rispetto alle notizie vere.

Gli approcci tradizionali per combattere la disinformazione si sono dimostrati insufficienti di fronte alla dimensione e alla complessità del fenomeno. Il fact-checking reattivo, cioè la verifica dei fatti dopo che le notizie sono già state pubblicate, pur mantenendo un valore importante, interviene spesso quando la notizia falsa ha già raggiunto un livello significativo di diffusione. La capacità limitata delle organizzazioni di verifica dei fatti non può stare al passo con il volume di contenuti problematici prodotti ogni giorno.

Questa situazione richiede un cambio di approccio metodologico, passando dall'approccio reattivo tradizionale verso strategie complete e preventive che possano prevedere, identificare e contrastare le campagne di disinformazione in modo più efficace e tempestivo.

1.2 Definizione, tipologie, esempi e impatto della disinformazione

Per combattere efficacemente le notizie false, è essenziale capire che non sono tutte uguali. Esistono tre categorie principali che spesso vengono confuse tra loro, ma che in realtà hanno origini, motivazioni e conseguenze diverse.

La **disinformazione** (disinformation) comprende contenuti volutamente falsi o ingannevoli creati e diffusi con l'intenzione chiara di ingannare il pubblico o causare danni specifici. Questa categoria si manifesta attraverso contenuti audio/video fabbricati o volutamente manipolati, teorie del complotto create intenzionalmente, o voci diffuse per causare danno o sfiducia. Un esempio tipico sono le false notizie create durante le elezioni per screditare un candidato, o articoli inventati che attribuiscono dichiarazioni mai fatte a personaggi pubblici.

La **misinformazione** (misinformation) include invece contenuti falsi o imprecisi che vengono condivisi senza intenzione maliziosa. In questo caso, gli individui che diffondono l'informazione sono inconsapevoli della sua falsità e agiscono in buona fede. La misinformazione comprende la condivisione di rumors prima di verificarne la veridicità, errori non intenzionali come didascalie fotografiche inesatte, date, statistiche e traduzioni errate, o casi in cui contenuti satirici vengono

presi sul serio.

La **malinformazione** (malinformation) comprende contenuti realmente veri ma utilizzati strategicamente per causare danno. Questa categoria include la pubblicazione deliberata di informazioni private per interesse personale, aziendale o politico piuttosto che per interesse pubblico. Esempi tipici sono: pubblicare conversazioni private per danneggiare qualcuno, diffondere email hackerate per rovinare la reputazione di una persona, o utilizzare informazioni vere ma fuori contesto per creare una falsa impressione.

Per semplificare, viene utilizzato il termine "disinformazione" come riferimento generale a tutte e tre le categorie sopra descritte.

La disinformazione attraversa tutti gli ambiti della nostra società, con conseguenze che vanno ben oltre la semplice condivisione di contenuti falsi.

Nel campo **politico**, la disinformazione può alterare i processi democratici e minare la fiducia nelle istituzioni. Esempi concreti sono le false dichiarazioni attribuite a figure religiose durante le campagne elettorali, o le interferenze documentate nelle elezioni americane del 2016, che hanno mostrato come la disinformazione possa compromettere i processi democratici.

Nel settore **sanitario**, la disinformazione può mettere a rischio la salute pubblica. Esempi tipici sono le false informazioni sui vaccini, i rimedi miracolosi senza base scientifica o le teorie del complotto su

malattie e cure. Durante la pandemia di COVID-19, ad esempio, la diffusione di notizie false sui trattamenti e sulla prevenzione ha causato comportamenti pericolosi e ostacolato le misure di sanità pubblica.

Gli impatti si estendono anche al settore **economico**, con la manipolazione di mercati finanziari attraverso informazioni false e danni reputazionali ad aziende e organizzazioni. Le truffe "pump and dump" sui social media rappresentano un esempio concreto di come contenuti falsi possano tradursi in perdite economiche.

Le conseguenze **sociali** includono la perdita della coesione sociale, alimentando tensioni etniche, religiose o ideologiche. La disinformazione sui migranti in Europa o le teorie cospirative come QAnon negli Stati Uniti dimostrano come queste notizie non validate possano incitare all'odio verso gruppi minoritari e portare a episodi di violenza reale.

1.3 Struttura, meccanismi narrativi e amplificazione delle fake news

L'analisi strutturale delle fake news rivela pattern ricorrenti nella loro costruzione narrativa e nei meccanismi di diffusione, suggerendo che la disinformazione segue principi riconoscibili che possono essere identificati e contrastati sistematicamente.

La **struttura drammatica** è il cuore di ogni fake news. Ogni storia ha sempre protagonisti chiaramente buoni e antagonisti chiara-

mente cattivi, senza alcuna sfumatura o complessità. Il conflitto è costruito appositamente per scatenare emozioni forti come rabbia, paura o indignazione, mentre la conclusione conferma sempre quello che il pubblico target vuole sentirsi dire.

L'**appello alle autorità (inventate)** sfrutta la nostra tendenza naturale a fidarci degli esperti. Le fake news citano "dottori" inesistenti, riferiscono studi scientifici completamente inventati o distorti, e attribuiscono titoli accademici falsi a persone comuni per renderle credibili.

Il **linguaggio estremo** : si usano parole forti: "scioccante", "incredibile", "quello che non vogliono farti sapere". Tutto è "sempre" o "mai", "tutti" o "nessuno".

La **creazione di urgenza artificiale**: Frasi ricorrenti come "condividi prima che venga censurato", "prima che sia troppo tardi" creano un senso di emergenza che ci spinge all'azione immediata.

La **sandwich technique** prevede l'inserimento di informazioni false tra fatti verificabili, sfruttando l'euristica di credibilità per associazione.

Il **source laundering** crea una rete di siti falsi che si citano a vicenda per far sembrare che tante fonti diverse confermino la stessa bugia. In realtà è come un gioco di specchi: la notizia falsa nasce da una fonte sola, ma rimbalza tra vari siti che si riferiscono l'uno all'altro, creando l'illusione che sia confermata da più parti.

In generale, per favorirne la diffusione vi sono i **network effects** che, attraverso bot e account collegati generano engagement artificiale nelle prime ore cruciali.

1.4 Obiettivi e contributi della tesi

Questa tesi si propone di adattare i metodi della Cyber Threat Intelligence (CTI) , tradizionalmente utilizzati per analizzare e prevedere attacchi informatici, al contesto della disinformazione.

Si intende, dunque, estendere il paradigma della CTI dal dominio della cybersecurity a quello narrativo, considerando le campagne di disinformazione come vere e proprie forme di attacco che seguono pattern strutturali riconoscibili e prevedibili.

In particolare, il lavoro si basa sull’adattamento della “Piramide del Dolore”, un framework consolidato nella cybersecurity che classifica gli Indicatori di Compromissione (IoC) in base alla difficoltà per un attaccante di modificarli. Questo modello viene reinterpretato per individuare e classificare IoC narrativi, ovvero indicatori che riflettono ricorrenze semantiche nei contenuti disinformativi.

Questo approccio supera la visione tradizionale che considera ogni fake news come un episodio isolato, proponendo invece una prospettiva d’insieme in cui le notizie false vengono inquadrate come parte di campagne organizzate.

Il punto di forza del sistema risiede nella capacità di attribuzione:

ogni contenuto disinformativo viene collegato a una specifica campagna tramite l'analisi di strutture narrative comuni, ricorrenze semantiche e correlazioni tra articoli.

Grazie a questo meccanismo di attribuzione, è possibile elaborare strategie di contrasto mirate. Si passa così da un approccio difensivo, centrato sulla semplice rimozione di contenuti falsi, a una strategia offensiva, orientata a contrastare le campagne di disinformazione alla radice.

Chapter 2

Background e tecnologie di riferimento

Questo capitolo presenta le basi teoriche e tecnologiche necessarie per comprendere l'approccio proposto.

Vengono esaminate quattro aree principali: i fondamenti della CTI e la sua estensione alla disinformazione, gli strumenti OpenCTI e STIX con le loro limitazioni nel rappresentare contenuti narrativi, i Large Language Models per l'estrazione semantica, e il dataset FakeCTI utilizzato per la sperimentazione.

2.1 Cyber Threat Intelligence (CTI)

Questa sezione esplora i principi fondamentali della CTI, analizza gli indicatori tradizionali attraverso il framework della Piramide del Do-

lore, e presenta l'innovativa estensione di questi concetti al dominio della disinformazione.

2.1.1 Definizione e principi fondamentali

La **Cyber Threat Intelligence (CTI)** rappresenta un approccio proattivo alla sicurezza informatica che si discosta dal tradizionale modello reattivo. È l'insieme di informazioni strutturate, dettagliate e contestualizzate sulle minacce informatiche, con lo scopo di aiutare le organizzazioni a prevenire, rilevare e rispondere agli attacchi cyber in modo tempestivo ed efficace.

La CTI non è semplicemente raccolta di dati, ma una disciplina che mira a una migliore comprensione delle minacce informatiche, consentendo ai team di sicurezza di adottare una strategia proattiva basata su dati concreti.

Secondo Allan Breakspear, l'intelligence è *“una capacità aziendale di prevedere i cambiamenti in tempo per fare qualcosa al riguardo. La capacità coinvolge previsione e intuizione, ed è destinata a identificare cambiamenti imminenti... che possono essere positivi, rappresentando opportunità... o negativi, rappresentando minacce.”*

Gli attributi fondamentali della CTI sono organizzati attorno al concetto centrale di Actionable Intelligence (Intelligence Attuabile), cioè la capacità di trasformare dati grezzi in informazioni pratiche, capaci di guidare decisioni operative concrete. I principali attributi

sono i seguenti:

Rilevante: l'intelligence deve essere pertinente al contesto specifico dell'organizzazione, tenendo in considerazione la sua superficie di attacco, l'infrastruttura e il settore di appartenenza.

Accurata: le informazioni devono essere precise e verificate, basate su fonti affidabili e metodologie rigorose, per garantire decisioni fondate e affidabili.

Tempestiva: l'intelligence deve essere fornita al momento giusto, per permettere azioni preventive o di risposta efficaci, rispettando le tempistiche critiche della sicurezza.

Contestualizzata: deve descrivere in modo completo gli attori delle minacce, le tattiche, tecniche e procedure (TTP) e gli indicatori di compromissione (IoC) rilevanti per il contesto organizzativo.

Analizzata: i dati grezzi devono essere elaborati, correlati e interpretati per estrarre pattern significativi.

Predittiva: l'intelligence deve fornire elementi in grado di anticipare i comportamenti futuri degli avversari e l'evoluzione delle minacce.

Questi attributi si combinano per garantire che l'intelligence sia davvero utile a rafforzare la sicurezza dell'organizzazione e a supportare decisioni consapevoli a ogni livello.

2.1.2 Indicatori tradizionali e la piramide del dolore

La Piramide del Dolore è un framework concettuale che classifica gli indicatori di Cyber Threat Intelligence (CTI) in base alla loro persistenza e resistenza alle modifiche da parte degli avversari. Questo modello organizza gli indicatori in una gerarchia strategica basata sul "costo del cambiamento" per gli attaccanti.

La piramide identifica diverse categorie di indicatori, ordinate dalla base alla cima secondo il crescente livello di difficoltà per gli attaccanti nel modificarli:

Hash Values (Triviale) - Facilmente alterabili modificando un file per generare un hash completamente diverso.

IP Addresses (Facile) - Rapidamente sostituibili utilizzando nuovi server o servizi cloud.

Domain Names (Semplice) - Richiedono investimenti minimi per acquisire nuovi domini.

Network Artifacts (Fastidioso) - User-agent, certificati SSL e pattern di comunicazione richiedono modifiche più sostanziali.

Tools (Impegnativo) - Richiedono investimenti significativi in sviluppo, test e implementazione.

TTPs (Difficile) - Le Tattiche, Tecniche e Procedure rappresentano i metodi operativi fondamentali, difficili da modificare senza ripensare completamente le strategie d'attacco.

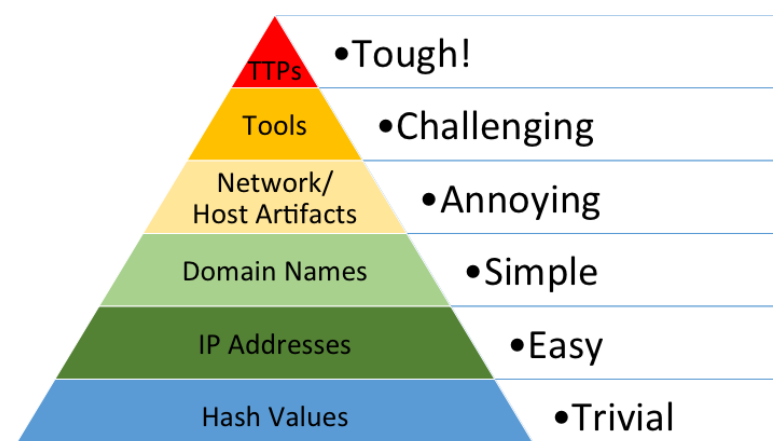


Figure 2.1: Pyramid of Pain

La Piramide del Dolore, in conclusione, evidenzia come gli indicatori più resistenti al cambiamento siano i più preziosi per la difesa, suggerendo di concentrarsi su elementi duraturi e significativi per migliorare l'efficacia della sicurezza.

2.1.3 Estensione della CTI al dominio della disinformazione

L'estensione della Cyber Threat Intelligence (CTI) al dominio della disinformazione nasce dall'osservazione di profonde analogie strutturali tra attacchi informatici e campagne di disinformazione. Entrambi coinvolgono attori coordinati, possono avere pattern ricorrenti che li rendono analizzabili con metodologie simili.

Tuttavia, è necessario riadattare la Piramide del Dolore descritta nel paragrafo precedente: originariamente concepita per il dominio cyber, questa struttura deve essere riformulata affinché possa adattarsi

al contesto narrativo della disinformazione. I livelli della piramide "rivisitata" saranno i seguenti: **Livelli inferiori (facilmente modificabili):**

- **Single Posts & Articles** - contenuti specifici, eliminabili rapidamente
- **Social Accounts** - profili e handle, sostituibili con nuove identità
- **Domain Names** - siti web, registrabili velocemente con nuovi domini

Livelli intermedi:

- **AI-generated Content** - contenuti sintetici, richiedono più risorse ma replicabili
- **Recurrent Themes** - triple semantiche ⟨soggetto, relazione, oggetto⟩ che catturano narrative persistenti

Livello superiore:

- **TTPs** - tattiche, tecniche e procedure strategiche, difficili da cambiare senza compromettere l'efficacia della campagna

Le triple semantiche sono importanti perché, a differenza degli indicatori tecnici tradizionali, catturano la storia di fondo che resta uguale anche quando gli attori cambiano dettagli superficiali. Per questo,

sono un ottimo punto su cui basare una CTI efficace contro la disinformazione.

In parallelo, il framework DISARM (DISinformation Analysis and Risk Management) rappresenta il tentativo più avanzato di standardizzazione per l'analisi delle campagne di disinformazione. Ispirato al modello MITRE ATT&CK, DISARM definisce una tassonomia strutturata in Tactics, Techniques e Procedures, coprendo l'intero ciclo di vita delle operazioni in 12 fasi, dalla pianificazione alla dismissione.

DISARM fornisce un vocabolario condiviso tra analisti, ricercatori e policy maker, favorendo interoperabilità, classificazione coerente delle minacce e difese coordinate contro le narrative dannose.

2.2 OpenCTI e il modello STIX

Questa sezione presenta gli strumenti tecnologici standardizzati per la gestione della threat intelligence. Viene analizzata l'architettura di OpenCTI come piattaforma moderna per la gestione delle informazioni sulle minacce, seguita da un'analisi dettagliata del modello STIX 2.1 e delle sue limitazioni quando applicato al dominio della disinformazione.

2.2.1 Introduzione a OpenCTI: obiettivi e architettura

OpenCTI è una piattaforma open source che permette alle organizzazioni di gestire le loro conoscenze di cyber threat intelligence e gli osservabili. È stata creata per strutturare, archiviare, organizzare e visualizzare informazioni tecniche e non tecniche sulle minacce informatiche. È progettata come una moderna applicazione web con un frontend orientato all'esperienza utente e un'architettura distribuita che include:

- ElasticSearch: Database principale per ricerche e analisi
- Redis: Cache e gestione sessioni
- S3/MinIO: Storage oggetti per file e artifacts
- RabbitMQ: Message broker per comunicazioni asincrone
- API GraphQL: Interfaccia di programmazione per l'accesso ai dati

OpenCTI può essere integrata con altri strumenti e applicazioni come MISP, TheHive, MITRE ATT&CK, ecc. OpenCTI rappresenta quindi una soluzione moderna e integrata per la gestione della threat intelligence, offrendo un approccio strutturato alla condivisione e analisi delle informazioni sulle minacce cyber.

2.2.2 Modello STIX 2.1: struttura e oggetti principali

STIX (Structured Threat Information Expression) è un linguaggio e formato di serializzazione utilizzato per scambiare intelligence sulle minacce informatiche (CTI). È caratterizzato da tre tipologie principali di oggetti che permettono di rappresentare diversi aspetti delle minacce attraverso una struttura modulare.

STIX Domain Objects (SDOs) – tra gli oggetti principali:

- **Attack Pattern:** descrive modi in cui gli avversari tentano di compromettere i target,
- **Campaign:** raggruppamento di comportamenti avversari contro specifici target,
- **Course of Action:** raccomandazioni su azioni da intraprendere,
- **Identity:** individui, organizzazioni o gruppi reali,
- **Indicator:** contiene pattern per rilevare attività sospette o malevole,
- **Malware:** rappresenta codice malevolo,
- **Threat Actor:** individui, gruppi o organizzazioni con intenti malevoli.

STIX Cyber-observable Objects (SCOs):

Rappresentano osservabili tecnici concreti:

- **File:** hash, nomi, dimensioni di file
- **Domain-Name:** domini utilizzati per C&C
- **IPv4-Address/IPv6-Address:** indirizzi IP sospetti
- **URL:** collegamenti web malevoli

STIX Relationship Objects (SROs) – 2 oggetti:

- **Relationship:** collega due SDO o SCO per descrivere le loro relazioni
- **Sighting:** indica che qualcosa nella CTI è stato osservato

STIX fornisce, quindi, un framework standardizzato per la rappresentazione e condivisione strutturata delle minacce informatiche.

2.2.3 Limitazioni nella rappresentazione della disinformazione

STIX è nato per le minacce informatiche, non per la disinformazione. Questo crea 3 “gap” principali quando vogliamo usarlo per analizzare fake news e campagne narrative.

Gap 1: Mancano gli “attori narrativi”

STIX ha oggetti `Identity` per rappresentare organizzazioni e gruppi,

ma non per i personaggi delle storie. Se in una fake news si parla di “Trump” o “Biden”, STIX non sa come rappresentare il loro ruolo narrativo, cioè non capisce chi sono e quale ruolo hanno nella storia (protagonisti, antagonisti, testimoni, ecc.).

Gap 2: Mancano le relazioni narrative

STIX contiene relazioni tecniche come “uses” o “targets”, utili a descrivere ad esempio che un malware usa una certa infrastruttura. Però non è progettato per gestire relazioni più complesse e significative dal punto di vista narrativo, dove il verbo cambia completamente il senso, come in “Trump critica Biden” oppure “Biden sostiene Trump”. Usare una relazione generica come “related-to” tra Trump e Biden perde completamente il significato preciso di chi fa cosa a chi. Senza una rappresentazione accurata del verbo o dell’azione, non è possibile cogliere le dinamiche della narrazione o il contesto informativo che queste relazioni portano con sé.

Gap 3: Focus tecnico vs semantico

STIX si concentra sugli indicatori tecnici (come hash di file, indirizzi IP), ma non comprende i pattern narrativi e semantici dietro alle informazioni. Al contrario, in questo lavoro di tesi vogliamo usare le triple semantiche, per esempio <Il vaccino, contiene, microchip>, che catturano la struttura narrativa profonda, difficile da cambiare o mascherare.

Per questo motivo serve estendere STIX oppure trovare modi cre-

ativi per rappresentare le narrative di disinformazione usando gli strumenti già esistenti.

2.3 Large Language Models per l'estrazione semantica

Questa sezione esplora l'utilizzo dei Large Language Models come strumenti avanzati per l'estrazione automatica di informazioni semantiche da contenuti testuali. Vengono presentati i modelli Llama e DeepSeek utilizzati in questo lavoro, insieme alle tecniche di prompt engineering necessarie per guidare efficacemente l'estrazione di triple semantiche da articoli di disinformazione.

2.3.1 Panoramica sui modelli linguistici LLM: Llama e DeepSeek

Llama (acronimo di Large Language Model Meta AI, e precedentemente stilizzato come LLaMA) è una famiglia di modelli linguistici autoregressivi di grandi dimensioni (LLM) pubblicati da Meta AI a partire da febbraio 2023.

I modelli Llama sono addestrati con un numero variabile di parametri, che vanno da 7 miliardi fino a 405 miliardi. Originariamente, Llama era disponibile solo come modello fondativo, mentre a partire da Llama 2 Meta AI ha iniziato a pubblicare versioni perfezionate con istruzioni

migliorate accanto ai modelli di base. Con il lancio di Llama 3, Meta ha integrato funzionalità di assistente virtuale in alcune regioni su Facebook e WhatsApp, oltre a un sito web dedicato.

DeepSeek è una società cinese specializzata nello sviluppo di modelli linguistici di grandi dimensioni (LLM) open source, fondata nel 2023 e sostenuta da un fondo speculativo chiamato High-Flyer. Il loro modello DeepSeek-R1, e in particolare la versione "DeepSeek-Coder", è stato sviluppato per offrire prestazioni competitive rispetto a modelli come ChatGPT, ma con un costo e un consumo di risorse significativamente inferiori. Il modello utilizzato, "deepseek-coder-6.7b-instruct", è disponibile su HuggingFace, permettendo agli sviluppatori di integrarlo facilmente in applicazioni di generazione di codice e completamento automatico.

In questo lavoro utilizziamo entrambi i modelli linguistici Llama e DeepSeek. Verranno utilizzati sia per l'estrazione automatica di tuple da articoli, sia per la generazione di nuovi testi. Nei capitoli successivi confronteremo le loro prestazioni e i risultati ottenuti.

2.3.2 Prompt engineering, few-shot learning e design del template

Il **prompt engineering** è la capacità di scrivere istruzioni chiare e precise per i modelli di intelligenza artificiale come LLaMA e DeepSeek. È importante perché questi modelli funzionano meglio quando sappiamo

spiegare esattamente cosa vogliamo che facciano.

Per estrarre triple semantiche, ad esempio, il prompt utilizzato combina:

- Istruzioni semplici e chiare
- Esempi concreti che mostrano come fare
- Regole per organizzare bene l'output

Il **few-shot learning** significa istruire il modello con esempio: gli mostriamo esempi reali per fargli capire cosa deve fare.

Esempio di input:

"Albert Einstein was born in Ulm, Germany, in 1879
and developed the theory of relativity."

Esempio di output:

- Albert Einstein – was born in – Ulm, Germany
- Albert Einstein – birth year – 1879
- Albert Einstein – developed – theory of relativity

Così il modello capisce il formato e lo usa per estrarre informazioni anche da nuovi testi.

2.4 Dataset FakeCTI: struttura e utilizzo

Il dataset utilizzato, **FakeCTI**, collega sistematicamente articoli di fake news alle loro campagne di disinformazione e attori minacciosi.

Colma una lacuna importante nella ricerca sulla disinformazione fornendo collegamenti espliciti tra contenuti falsi e le campagne associate alla loro diffusione.

Tale dataset è costituito dalle seguenti colonne, che forniscono una visione completa di ogni articolo di fake news:

- **ID** - L'identificativo univoco per ciascuna notizia
- **URL** - Link diretto all'articolo di fake news
- **TITLE** - Titolo della fake news
- **SOURCE** - La fonte o il sito da cui proviene la notizia
- **TEXT** - Contenuto testuale completo dell'articolo
- **CAMPAIGN** - La campagna di disinformazione a cui appartiene la notizia
- **THREAT ACTOR** - Attore minaccioso responsabile della campagna
- **TYPE** - Il tipo di piattaforma (ad esempio, web o social) attraverso la quale è stata distribuita la notizia

Il dataset presenta dimensioni significative che lo rendono uno strumento prezioso per la ricerca. Con 12.155 articoli di fake news raccolti e analizzati, offre una base per studi statistici e analisi comparative.

La catalogazione di 43 campagne di disinformazione distinte permette di tracciare l'evoluzione delle narrative nel tempo.

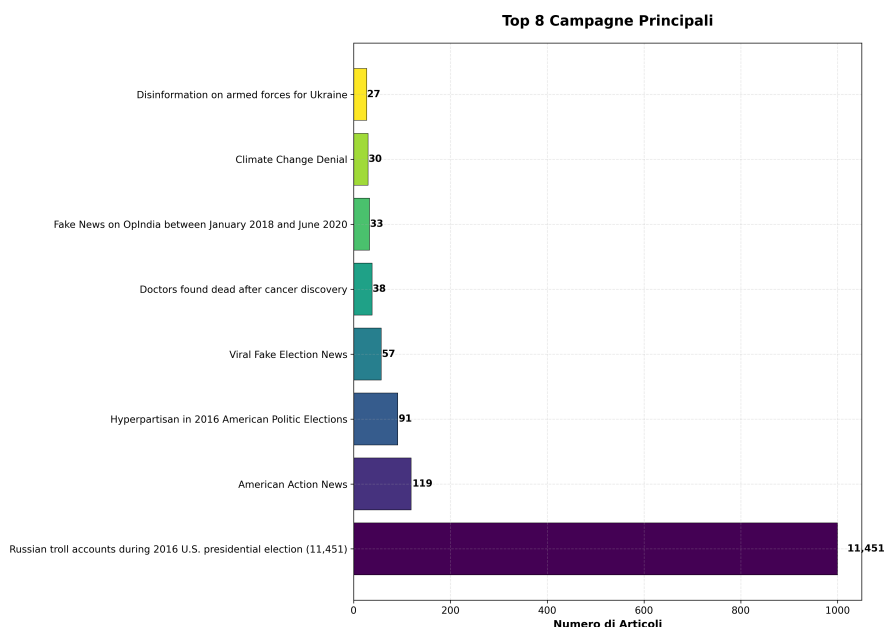


Figure 2.2: Principali campagne di disinformazione nel dataset FakeCTI.

Il dataset classifica i contenuti in diverse tipologie, consentendo di analizzare i canali preferiti per la diffusione della disinformazione. Questa suddivisione è fondamentale per sviluppare strategie di contenimento specifiche, che tengano conto delle caratteristiche di ciascun canale e dei diversi linguaggi usati, come repost, menzioni (@) e hashtag (#) su piattaforme come Twitter.

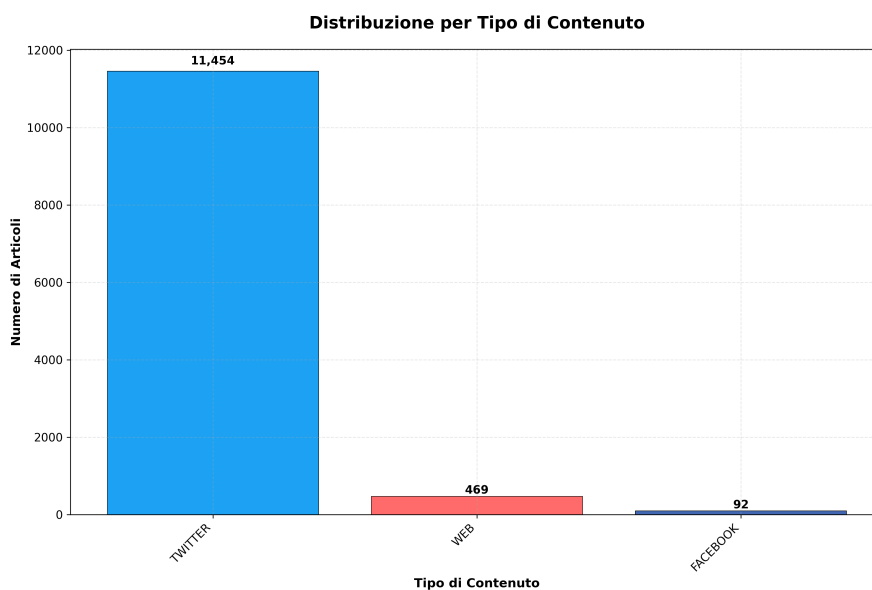


Figure 2.3: Distribuzione dei tipi di contenuti nel dataset FakeCTI.

A differenza di altri dataset che si limitano a classificare i contenuti come 'veri' o 'falsi', FakeCTI fornisce l'attribuzione, offrendo così una base per seguire nel tempo l'evoluzione delle narrative di disinformazione, che rappresenta uno degli obiettivi principali di questa tesi.

Chapter 3

Architettura funzionale del sistema

Questo capitolo presenta l'architettura funzionale del sistema sviluppato per l'analisi della disinformazione tramite OpenCTI e lo standard STIX. Dopo aver analizzato le limitazioni degli approcci tradizionali, viene proposta una soluzione innovativa che riutilizza creativamente gli oggetti STIX (originariamente progettati per le minacce informatiche) per rappresentare le strutture narrative della disinformazione attraverso triple semantiche soggetto-verbo-oggetto.

L'architettura si articola in tre componenti principali: il Producer per l'estrazione e strutturazione delle relazioni semantiche, il Consumer per la classificazione automatica dei contenuti e il Generatore per l'espansione narrativa. Viene inoltre illustrato come il sistema trasformi contenuti testuali non strutturati in un knowledge graph nav-

igabile, abilitando nuove forme di analisi esplorativa delle campagne di disinformazione.

3.1 Riutilizzo creativo degli oggetti STIX per la disinformazione

Lo standard STIX presenta tre limitazioni fondamentali quando applicato al dominio della disinformazione: l'assenza di attori narrativi, la mancanza di relazioni semantiche complesse e un focus prevalentemente tecnico piuttosto che semantico. Queste lacune strutturali rendono inadeguata una trasposizione diretta degli oggetti STIX dalle minacce informatiche alle campagne di disinformazione.

Tuttavia, l'adozione di un approccio completamente nuovo comporterebbe la perdita dei vantaggi offerti dall'ecosistema OpenCTI consolidato, incluse le funzionalità di visualizzazione, correlazione e gestione dei dati già esistenti. Per questo motivo, è stato sviluppato un approccio innovativo che mantiene la compatibilità con l'infrastruttura STIX esistente, reinterpretando creativamente gli oggetti standard per rappresentare le strutture narrative della disinformazione.

La strategia adottata preserva l'architettura modulare e le capacità relazionali di STIX, riutilizzando gli oggetti esistenti con semantiche ridefinite specificamente per il contesto narrativo.

3.1.1 Mappatura concettuale da STIX a narrativa

La mappatura proposta trasforma i quattro oggetti principali di STIX dalla loro funzione originale nel dominio della cybersecurity a una nuova applicazione nell'analisi narrativa:

STIX	CTI Tradizionale	Uso Narrativo
Identity	Organizzazione/vittima	Soggetto narrativo
Attack-Pattern	Tecniche attacco	Verbo/azione
Indicator	IoC	Tripla semantica
Campaign	Campagna malevola	Contesto narrativo

Table 3.1: Mappatura oggetti STIX: uso tradizionale vs narrativo

In conclusione, questa mappatura permette di utilizzare il framework STIX per rappresentare in modo efficace le strutture narrative tipiche della disinformazione. Ciò consente di sfruttare il modello già consolidato per descrivere non solo attori e tecniche di attacco, ma anche azioni, contesti narrativi e relazioni semantiche complesse nell'ambito dell'analisi della disinformazione.

3.1.2 Implementazione tecnica con estensioni STIX

Il formato STIX 2.1 consente l'estensione degli oggetti tramite campi custom, identificati dal prefisso `x_`. Il sistema sviluppato sfrutta questa possibilità per arricchire ciascun oggetto con metadati semantici rilevanti.

Ad esempio, l'oggetto Identity utilizzato per rappresentare soggetti e oggetti narrativi viene esteso con campi specifici che mantengono la tracciabilità semantica:

```

1  entity_data = {
2      "name": entity_name,
3      "type": identity_class,
4      "description": f"Soggetto della tripla semantica: {clean_text(tripla_completa)}",
5      "objectLabel": subject_labels,
6      "x_triple_contenuto": contenuto_pulito,
7      "x_triple_ruolo": "soggetto",
8      "x_triple_completa": clean_text(tripla_completa),
9      "x_triple_campagna": campagna
10 }

```

Figure 3.1: Estensioni per l'oggetto Identity usato come soggetto o oggetto narrativo

- `x_triple_contenuto`: Memorizza il contenuto testuale dell'entità estratta dalla tripla
- `x_triple_ruolo`: Specifica il ruolo semantico ("soggetto" o "oggetto") nella struttura narrativa
- `x_triple_completa`: Mantiene il riferimento alla tripla semantica originale per la tracciabilità
- `x_triple_campagna`: Collega l'entità alla campagna di disinformazione di appartenenza

Allo stesso modo, l'oggetto Indicator, che rappresenta l'intera tripla semantica, utilizza un pattern STIX personalizzato nella forma:

```

1  indicator_data = {
2      "name": indicator_name,
3      "pattern": f"[x-semantic-triple:content = '{tripla_pulita}']",
4      "pattern_type": "stix",
5      "description": f"Complete semantic triple from disinformation campaign '{campagna}': {tripla_pulita}",
6      "x_triple_completa": tripla_pulita,
7      "x_triple_campagna": campagna,
8      "x_triple_type": "semantic-narrative",
9      "x_opencti_detection": True
10 }

```

Figure 3.2: Estensioni per Indicator: Tripla Semantica Completa

Tale flessibilità nelle estensioni STIX è ciò che ci ha permesso di

trasformare un linguaggio progettato per descrivere minacce in uno strumento efficace per modellare dinamiche narrative.

3.2 Navigazione semantica e correlazione tramite relazioni STIX

La mappatura degli oggetti STIX al dominio narrativo richiede un sistema di relazioni che trasformi le triple semantiche estratte in un grafo navigabile. Questa sezione presenta l'architettura relazionale del sistema, che consente agli analisti di esplorare le interconnessioni tra campagne di disinformazione attraverso l'interfaccia OpenCTI, identificando pattern narrativi ricorrenti e correlazioni semantiche complesse.

3.2.1 Modellazione delle relazioni

Per creare una rete navigabile in OpenCTI, il sistema implementa un'architettura relazionale che trasforma le triple testuali in un grafo semantico strutturato. La modellazione delle relazioni consente agli analisti di navigare esplorativamente attraverso le narrative di disinformazione seguendo i collegamenti semantici rappresentati nel seguente schema:

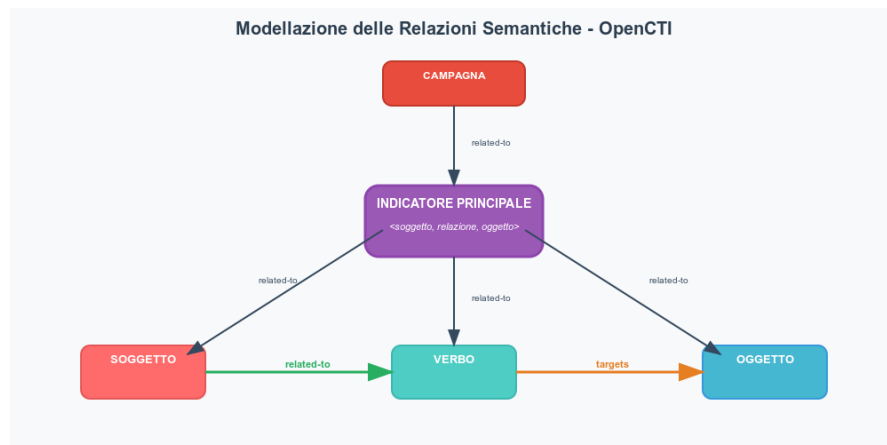


Figure 3.3: Schema delle relazioni semantiche nel grafo di OpenCTI

Indicatore → Entità semantiche *Tipo:* related-to

Scopo: Collega l'indicatore principale (tripla completa) ai suoi componenti atomici

Soggetto → Verbo

Tipo: related-to

Scopo: Rappresenta l'agente che esegue l'azione

Verbo → Oggetto

Tipo: targets

Scopo: Indica che l'azione è diretta versol'oggetto

Relazioni di contesto


Indicatore → Campagna

Tipo: related-to

Scopo: Attribuisce le narrative a specifiche campagne di disinformazione

Prima di creare qualsiasi relazione, il sistema verifica l'esistenza di collegamenti già presenti tra le stesse entità. Questa verifica avviene attraverso filtri combinati che controllano l'entità sorgente (`fromId`), l'entità destinazione (`toId`) e il tipo di relazione (`relationship_type`).

Questo approccio previene la duplicazione relazionale che comprometterebbe l'integrità del grafo semantico.

A screenshot of a code editor with a dark background and three colored window control buttons (red, yellow, green) in the top-left corner. The code is a JSON object defining filters for a query. It is as follows:

```
1  filters = {  
2    "mode": "and",  
3    "filters": [  
4      {"key": "fromId", "values": [source_id]},  
5      {"key": "toId", "values": [target_id]},  
6      {"key": "relationship_type", "values": [relationship_type]}  
7    ],  
8    "filterGroups": []  
9  }
```

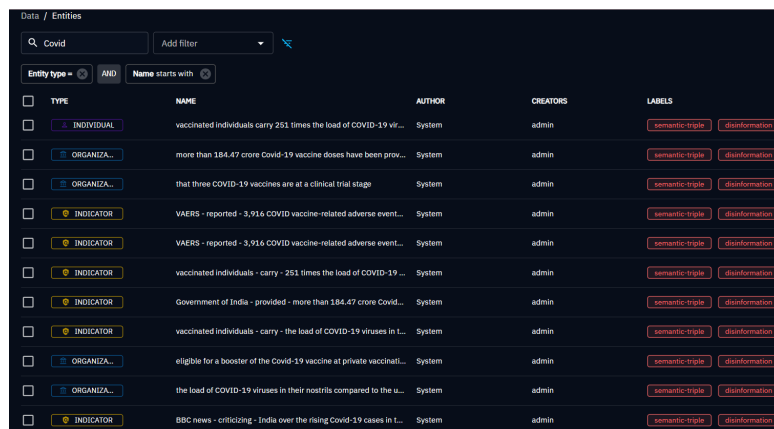
Figure 3.4: Verifica dell'esistenza di relazioni duplicate nel grafo semantico

Questa verifica garantisce che la rete semantica sia coerente e navigabile, permettendo agli analisti di seguire percorsi logici nell'esplorazione delle narrative di disinformazione.

3.2.2 Navigazione esplorativa attraverso OpenCTI

La rete di relazioni semantiche consente una navigazione interattiva e esplorativa del grafo attraverso l'interfaccia OpenCTI, sfruttando le funzionalità di visualizzazione dei collegamenti tra entità. Il sistema, quindi, si basa su una rappresentazione a grafo orientato in cui ogni

entità del modello STIX costituisce un nodo, mentre le relazioni tra di esse formano archi direzionali che codificano il tipo e la natura delle connessioni narrative. Il processo di esplorazione segue una sequenza logica strutturata.



TYPE	NAME	AUTHOR	CREATORS	LABELS
INDIVIDUAL	vaccinated individuals carry 251 times the load of COVID-19 vir...	System	admin	semantic-triple, disinformation
ORGANIZA...	more than 184.47 crore Covid-19 vaccine doses have been prov...	System	admin	semantic-triple, disinformation
ORGANIZA...	that three COVID-19 vaccines are at a clinical trial stage	System	admin	semantic-triple, disinformation
INDICATOR	VAERS - reported - 3,916 COVID vaccine-related adverse event...	System	admin	semantic-triple, disinformation
INDICATOR	VAERS - reported - 3,916 COVID vaccine-related adverse event...	System	admin	semantic-triple, disinformation
INDICATOR	vaccinated individuals - carry - 251 times the load of COVID-19 ...	System	admin	semantic-triple, disinformation
INDICATOR	Government of India - provided - more than 184.47 crore Covid...	System	admin	semantic-triple, disinformation
INDICATOR	vaccinated individuals - carry - the load of COVID-19 viruses in L...	System	admin	semantic-triple, disinformation
ORGANIZA...	eligible for a booster of the Covid-19 vaccine at private vaccinati...	System	admin	semantic-triple, disinformation
ORGANIZA...	the load of COVID-19 viruses in their nostrils compared to the u...	System	admin	semantic-triple, disinformation
INDICATOR	BBC news - criticizing - India over the rising Covid-19 cases in L...	System	admin	semantic-triple, disinformation

Figure 3.5: Navigazione esplorativa in OpenCTI: ricerca semantica per "Covid"

Come visibile nella Figura 3.5, l'interfaccia di ricerca implementa un sistema di query semantiche avanzate che va oltre la semplice corrispondenza testuale. La ricerca per il termine "Covid" dimostra tre capacità fondamentali del sistema:

Identificazione di entità semanticamente correlate: Il sistema riconosce connessioni semantiche anche quando il termine di ricerca non appare esplicitamente nel nome dell'entità. Ad esempio, cercando "Covid" vengono restituite anche entità che menzionano "vaccine doses" o "adverse events" perché semanticamente correlate al dominio della disinformazione sui vaccini, dimostrando la capacità di inferire relazioni concettuali implicite.

Applicazione di filtri tipologici: L'interfaccia offre strumenti di

raffinamento granulare attraverso i filtri "Entity type" e "Name starts with", consentendo al ricercatore di focalizzarsi su specifiche categorie di entità (individual, organization, indicator) o pattern nominali senza perdere la coerenza semantica dell'insieme.

Aggregazione multimodale delle narrative: La query restituisce automaticamente entità di diversi tipi (articoli, post social, indicatori) tutte unificate sotto le etichette "semantic-triple" e "disinformation", permettendo una visione olistica che include dati statistici sui vaccini, segnalazioni di eventi avversi, post social e articoli giornalistici relativi allo stesso tema.

Le proprietà evidenziate hanno un'applicazione pratica fondamentale: un ricercatore in ambito cybersecurity può utilizzare una singola query per mappare immediatamente l'intero panorama delle fake news su un argomento specifico. Invece di dover cercare manualmente attraverso diverse fonti vi è la piattaforma di OpenCTI che aggrega automaticamente tutte le narrazioni di disinformazione correlate, permettendo al ricercatore di identificare rapidamente pattern, fonti ricorrenti e strategie narrative utilizzate nelle campagne di disinformazione. Questo approccio trasforma un processo di ricerca dispersivo in un'analisi sistematica e strutturata.

Si analizza, quindi, nel dettaglio un'entità specifica, come un indicatore, per esaminare le relazioni semantiche e i collegamenti strutturati presenti nel sistema.

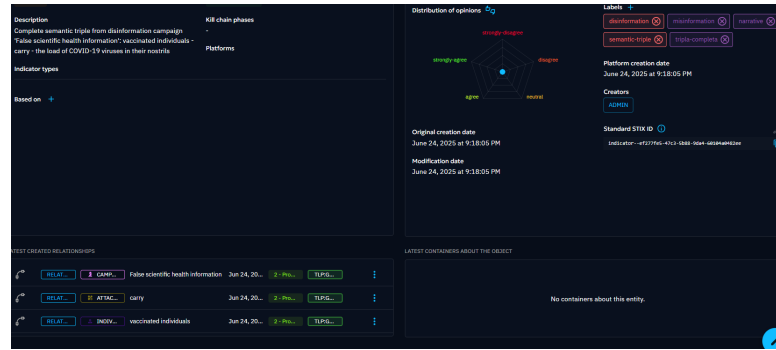


Figure 3.6: Vista granulare di entità STIX

La vista di dettaglio (Figura 3.6) illustra la struttura granulare di una tripla semantica e la sua integrazione nel knowledge graph.

La sezione "Latest Created Relationships" evidenzia le connessioni fondamentali dell'entità: il collegamento alla campagna di disinformazione di appartenenza, l'associazione al pattern di attacco corrispondente e la relazione con l'entità soggetto della tripla.

Questa visualizzazione permette di comprendere immediatamente come ogni singolo elemento si posizioni all'interno del grafo narrativo complessivo.

L'interfaccia consente, inoltre, di esplorare le relazioni in entrambe le direzioni (forward e backward), abilitando un'analisi multidirezionale che spazia dall'esame granulare di singole entità semantiche alla comprensione sistemica delle interconnessioni narrative. Questa capacità di navigazione trasforma l'analisi da un processo lineare a un'esplorazione dinamica del tessuto relazionale che caratterizza le campagne di disinformazione.

3.3 Architettura Funzionale: Producer, Consumer e Generatore

L'architettura funzionale del sistema è organizzata come una pipeline modulare composta da tre componenti principali: Producer, Consumer e Generatore.

3.3.1 Panoramica generale dell'architettura

Il sistema analizza articoli e post che diffondono disinformazione, li trasforma in una rete strutturata di informazioni sulle campagne attive, e poi usa questa rete per riconoscere e classificare automaticamente nuovi contenuti. L'architettura è composta da tre componenti principali che operano in sequenza: il Producer estrae le relazioni semantiche dai contenuti e li struttura in un knowledge graph, il Consumer utilizza queste informazioni per classificare nuovi articoli, mentre il Generatore può opzionalmente arricchire il dataset di test con contenuti sintetici per migliorare le performance di classificazione.

3.3.2 Il Modulo Producer: estrazione e caricamento su OpenCTI

Il Producer rappresenta il punto di ingresso del sistema e ha la responsabilità di trasformare contenuti testuali non strutturati in entità semantiche formalizzate secondo lo standard STIX. Utilizza un Large

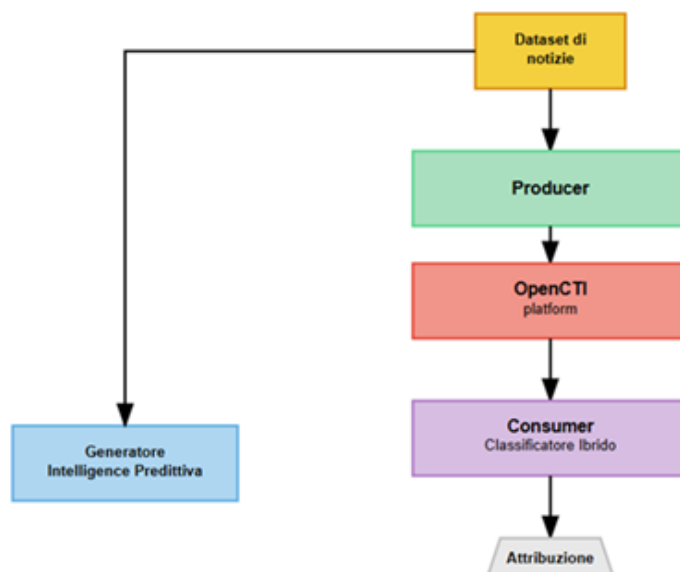


Figure 3.7: Architettura del sistema Producer-Consumer-Generatore Language Model per estrarre automaticamente relazioni nel formato soggetto-verbo-oggetto e le carica in OpenCTI seguendo lo standard STIX.

Il **Producer** opera in 5 fasi sequenziali:

1. **Pulizia del testo:** Rimuove caratteri speciali e limita la lunghezza
2. **Estrazione con LLM:** Usa modelli come Llama-3-8B e Deepseek per trovare triple semantiche.
3. **Validazione:** Controlla formato e qualità delle triple estratte
4. **Creazione entità STIX:** Trasforma le triple in entità OpenCTI
5. **Salvataggio:** Carica tutto nel knowledge graph.

Il sistema supporta due implementazioni alternative per l'estrazione semantica:

Opzione 1: Llama-3-8B (Quantizzato)

```
1 model_name = "MaziyarPanahi/Llama-3-8B-Instruct-v0.10-GGUF"
2 model_file = "Llama-3-8B-Instruct-v0.10.Q5_K_M.gguf"
3
4 llm = Llama(
5     model_path=model_path,
6     n_ctx=4096,
7     verbose=False
8 )
```

Figure 3.8: Inizializzazione del modello Llama-3-8B

Opzione 2: DeepSeek-Coder

```
1 model_name = "deepseek-ai/deepseek-coder-6.7b-instruct"
2
3 tokenizer = AutoTokenizer.from_pretrained(model_name, trust_remote_code=True)
4 model = AutoModelForCausalLM.from_pretrained(
5     model_name,
6     trust_remote_code=True,
7     torch_dtype=torch.float16,
8     device_map="auto"
9 )
```

Figure 3.9: Inizializzazione di DeepSeek-Coder-6.7B-Instruct

Entrambi i modelli sono disponibili su *Hugging Face Hub*, ma utilizzano approcci di caricamento diversi ottimizzati per contesti d'uso specifici:

- **Llama-3-8B**: utilizza il formato GGUF quantizzato tramite llama-cpp-python. La quantizzazione Q5_K_M riduce l'utilizzo di memoria mantenendo elevata accuratezza.

- **DeepSeek-Coder**: utilizza la libreria Transformers standard con precisione float16.

Il Prompt Engineering ha un ruolo cruciale nella qualità delle triple semantiche generate dai modelli linguistici. L'obiettivo è guidare il modello verso l'identificazione di triple <soggetto, relazione, oggetto> rilevanti. Il sistema implementa due approcci di prompt engineering differenziati, poiché i linguaggi dei contenuti analizzati presentano caratteristiche diverse che richiedono strategie di estrazione specifiche:

1. Prompt per Articoli di News : utilizza un approccio strutturato e formale. Le istruzioni guidano il modello a concentrarsi sui fatti principali, richiedendo l'identificazione di soggetti chiari, verbi precisi e oggetti significativi. Il formato richiesto è rigido e include la gestione delle citazioni dirette, dove il parlante diventa il soggetto della tripla.

2. Prompt per Social Media: viene adattato per gestire le specificità del linguaggio informale, hashtag, menzioni e retweet. L'enfasi è posta sull'identificazione di relazioni semantiche significative, escludendo esplicitamente le meccaniche di piattaforma (come "utente - posta - tweet"). Il prompt guida il modello a concentrarsi sui contenuti sostanziali piuttosto che sulle azioni meta-comunicative tipiche dei social media.

La necessità di utilizzare due prompt distinti deriva dalle differenze

linguistiche sostanziali tra i contenuti giornalistici formali e la comunicazione informale tipica dei social media. Queste due tipologie testuali richiedono approcci differenziati per poter guidare efficacemente il modello nell'estrazione di triple semantiche coerenti.

Dopo l'estrazione e la trasformazione, il Producer genera un knowledge graph strutturato in OpenCTI. Dunque, crea una rete navigabile di relazioni semantiche, abilitando il Consumer alla fase successiva di classificazione delle campagne.

3.3.3 Il Modulo Consumer: classificazione e attribuzione

Il Modulo Consumer rappresenta il componente finale del pipeline di analisi, progettato per ricevere contenuti da analizzare e classificarli automaticamente determinando a quale campagna di disinformazione appartengano. Il funzionamento si basa su un concetto fondamentale: le tuple semantiche estratte precedentemente da OpenCTI vengono utilizzate come "impronte digitali" di ogni campagna. Quando arriva un nuovo articolo da classificare, il sistema confronta il suo contenuto con queste tuple per trovare la campagna più simile, assegnando un punteggio di confidenza che indica quanto è probabile l'appartenenza.

Il sistema si basa su due fasi principali che collaborano per garantire un'analisi completa e precisa.

Nella prima fase, il sistema sincronizza le tuple semantiche generate dal Producer tramite istanza separata di OpenCTI. Dunque, vengono recuperati tutti gli indicatori disponibili dal Producer attraverso la funzione dedicata:



```
1 def sync_indicators():
2     indicators = producer_client.indicator.list(get_all=True, first=10000)
3     print(f"{len(indicators)} indicatori trovati.")
4
5     for index, ind in enumerate(indicators):
6         consumer_client.indicator.create(
7             name=ind["name"],
8             description=ind.get("description", ""),
9             pattern=ind["pattern"],
10            pattern_type=ind["pattern_type"],
11            x_opencti_main_observable_type=ind.get("x_opencti_main_observable_type", "Text"),
12            valid_from=ind.get("valid_from", None),
13            labels=[l["value"] for l in ind.get("objectLabel", [])]
14        )
15
```

Figure 3.10: Flusso di sincronizzazione tra Producer e Consumer

La seconda fase utilizza le tuple sincronizzate per classificare nuovi contenuti. Quando arriva un articolo sconosciuto, il sistema verifica quante e quali tuple associate alle campagne sono presenti o simili nel testo, calcolando un punteggio di somiglianza per ogni campagna possibile.

La seconda fase utilizza le tuple sincronizzate per classificare nuovi contenuti. Quando arriva un articolo sconosciuto, il sistema verifica quante e quali tuple associate alle campagne sono presenti o simili nel testo, calcolando un punteggio di somiglianza per ogni campagna possibile.

L'algoritmo di classificazione implementa una strategia multi-componente che assegna pesi specifici a diverse metriche di similarità.

Il calcolo dello score finale funziona come un sistema di valutazione

con quattro componenti principali, ognuna delle quali contribuisce con un peso specifico al punteggio totale.

Il primo componente è il punteggio TF-IDF, che contribuisce per il 50% al punteggio totale. Questo algoritmo individua e conta le parole significative che un nuovo articolo condivide con le tuple semantiche di ciascuna campagna, attribuendo un peso maggiore ai termini rari e distintivi rispetto a quelli comuni e generici. In pratica, il sistema identifica le “parole firma” di ogni campagna, ovvero quelle parole caratteristiche che ne definiscono in modo univoco il contenuto. Per esempio, la presenza di termini come “Hillary Clinton” e “elezioni” assegna un punteggio alto alla campagna elettorale, mentre parole come “temperatura” e “sole” aumentano il punteggio per la campagna climatica. Parole generiche e molto frequenti, come “molto” o “questo”, vengono invece ignorate perché non aiutano a discriminare tra le diverse campagne.

Il secondo componente sono gli **embedding semantici** che pesano per il 25%. Questo sistema utilizza l’intelligenza artificiale per capire il significato profondo del testo, non solo le parole. Anche se un articolo non contiene esattamente le stesse parole delle tuple, può ottenere un punteggio alto se parla di concetti simili. Per esempio, un articolo sui "candidati presidenziali" può essere collegato a tuple che parlano di "elezioni politiche" anche senza usare le stesse parole.

Il terzo componente è il **bonus keywords** che contribuisce per il 15%.

Il sistema verifica se nel testo compaiono parole chiave specifiche estratte dalle tuple di ogni campagna, incluse entità come nomi di persone, organizzazioni e luoghi.

Il quarto componente sono i **bonus speciali** che pesano per il 10%. Questi riconoscono pattern particolari come contenuti Twitter (identificati da hashtag, menzioni @, RT) o tematiche specifiche.

$$\begin{aligned}\text{Score finale} &= 0.5 \cdot \text{TF-IDF} \\ &+ 0.25 \cdot \text{Embedding} \\ &+ 0.15 \cdot \text{Keywords} \\ &+ 0.1 \cdot \text{Bonus speciali}\end{aligned}$$

L'output del modulo Consumer include non solo la campagna predetta ma anche metriche dettagliate che permettono di valutare la qualità della classificazione. Vengono forniti i punteggi delle singole componenti, il livello di confidenza categorizzato, le predizioni alternative e indicatori specifici come il riconoscimento di contenuti Twitter.

3.3.4 Il Modulo Generatore: espansione narrativa per validazione

Il Modulo Generatore crea automaticamente contenuti di disinformazione utilizzando modelli LLM come Llama-3-8B-Instruct e DeepSeek. Il

processo di generazione inizia con la selezione del dataset, che prevede un minimo di due articoli per campagna, per garantire un contesto adeguato, e un massimo di 2000 articoli; la scelta avviene in modo casuale per evitare bias. Successivamente, gli articoli selezionati vengono suddivisi in due gruppi:

- **Example List:** esempi reali per guidare il modello
- **Comparison Articles:** articoli per validazione e confronto

Questa struttura consente di definire chiaramente il compito del modello, fornendo istruzioni precise e una contestualizzazione efficace.

Il modulo si integra nel flusso elaborativo ricevendo in input i dati del test set e producendo contenuti sintetici, che vengono inviati al Consumer per l'analisi finale di attribuzione. I principali vantaggi del Generatore consistono nell'ampliare dataset limitati con campioni sintetici, bilanciare campagne sottorappresentate, anticipare possibili evoluzioni narrative, supportare il testing di sistemi di rilevamento con nuovi pattern e favorire lo sviluppo di contromisure proattive.

Chapter 4

Sperimentazione e Validazione

4.1 Setup Sperimentale

4.1.1 Ambiente tecnico (Docker, hardware, software)

L'ambiente sperimentale è composto da due istanze Docker separate: un'istanza producer che carica le tuple estratte su OpenCTI, e un'istanza consumer che scarica e analizza i dati da OpenCTI. Ogni istanza opera con il proprio token API di autenticazione e su porte localhost separate per garantire isolamento funzionale. La comunicazione tra le istanze avviene tramite script di sincronizzazione che utilizza le API REST di entrambe le istanze per trasferire i dati di threat intelligence.

Le istanze vengono avviate tramite il comando `docker-compose up -d` e sono accessibili in locale tramite browser su:

- Producer: `http://localhost:8081`
- Consumer: `http://localhost:8080`

Software e Versioni

- Python: 3.11
- OpenCTI Platform: 6.5.10
- Elasticsearch: 8.17.4
- Redis: 7.4.2
- MinIO: RELEASE.2024-05-28T17-19-04Z
- RabbitMQ: 4.0-management
- Librerie Python: `pycti`, `pandas`, `llama-cpp-python`, `torch`, `spacy`, `scikit-learn`

Requisiti Hardware

- RAM: 8GB minimi (4GB per Elasticsearch + 4GB per altri servizi)
- CPU: multi-core consigliato (per gestire worker OpenCTI paralleli)

- Disco: 20GB per volumi persistenti (storage dati OpenCTI, Elasticsearch, MinIO)
- OS: Ubuntu 22.04 LTS testato (compatibile con Docker Engine e dipendenze Python)

Sistema utilizzato:

- RAM: 8GB (Intel Core i7-8565U)
- CPU: 4 core Intel i7-8565U @ 1.80GHz
- Disco: 238GB totali disponibili
- OS: Windows 11 con Ubuntu 24.04.1 LTS tramite WSL2 e Docker Desktop

4.1.2 Preparazione dataset e metriche di valutazione

Il dataset FakeCTI è suddiviso in set di training e test seguendo una strategia bilanciata per campagna di disinformazione, in modo da preservare la rappresentatività delle classi.

Criteri di selezione e divisione:

- **Filtro iniziale:** selezione delle campagne con più di 10 articoli per garantire rappresentatività statistica
- **Campagne con >101 articoli:** limitazione a 100 istanze totali (90 train + 10 test) per evitare sbilanciamento

- **Campagne con 11-101 articoli:** divisione proporzionale 90%-10% del totale disponibile
- **Mescolamento casuale:** applicazione di `random_state=42` per garantire riproducibilità dei risultati

Output della procedura:

Il processo genera automaticamente due file Excel:

- `train.xlsx`: set di training per l'addestramento dei modelli
- `test.xlsx`: set di test per la valutazione delle performance

La strategia garantisce distribuzione bilanciata delle campagne mantenendo la proporzione 90%-10% tra training e test set, essenziale per una valutazione affidabile dei classificatori di threat intelligence.

4.2 Esperimento 1: Baseline su Dati Reali

Il primo esperimento ha lo scopo di valutare l'efficacia di un sistema distribuito di intelligence semantica nella classificazione automatica di contenuti di disinformazione. L'obiettivo è simulare uno scenario realistico di condivisione della conoscenza tra attori diversi, utilizzando un'infrastruttura che sfrutta OpenCTI e il linguaggio STIX per organizzare, correlare e utilizzare informazioni strutturate a partire da articoli fake annotati.

4.2.1 Pipeline di classificazione tradizionale

Il sistema è basato su un'architettura **Producer-Consumer** che separa nettamente la fase di apprendimento da quella di classificazione. Il **Producer** rappresenta un attore di intelligence che analizza contenuti noti, utilizzando un modello LLaMA 3 (8B) per estrarre triple semantiche nel formato *soggetto - verbo - oggetto* dagli articoli di training. Queste triple vengono trasformate in entità STIX e caricate in OpenCTI, dove sono organizzate in campagne, indicatori e relazioni semantiche.

Le informazioni vengono poi sincronizzate verso l'istanza **Consumer**, che simula un attore che riceve contenuti sospetti. Il Consumer elabora ciascun articolo del test set e lo confronta con le triple note utilizzando un classificatore multi-componente che combina diverse metriche:

- Similarità TF-IDF;
- Embedding semantici;
- Keyword matching;
- Bonus specifici (es. rilevamento pattern da social media).

Il sistema restituisce, per ciascun contenuto:

- L'attribuzione a una campagna di disinformazione;
- Uno score numerico di confidenza;

- Un livello qualitativo di affidabilità.

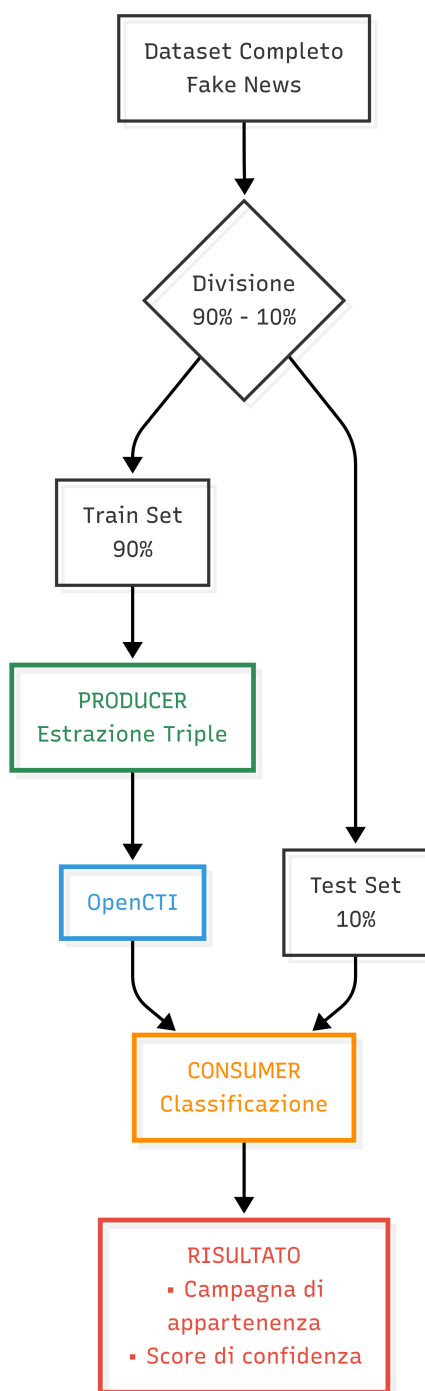


Figure 4.1: Pipeline di classificazione fake news con OpenCTI

Il dataset utilizzato è stato bilanciato per evitare che le campagne più frequenti dominassero la classificazione. Sono stati selezionati al

massimo 100 articoli per campagna, con una suddivisione 90/10 tra training e test. Questo garantisce generalizzabilità e robustezza nella valutazione.

Il sistema si distingue per:

- **Realismo operativo:** separazione tra chi produce e chi consuma intelligence;
- **Robustezza:** la classificazione non si basa su un solo criterio ma usa più componenti : questo rende il sistema più robusto perché se un metodo fallisce, gli altri possono compensare.
- **Trasparenza:** il sistema fornisce spiegazioni dettagliate delle decisioni prese attraverso due meccanismi principali. Da un lato, la struttura del knowledge graph in OpenCTI permette di ispezionare le relazioni semantiche che hanno guidato la classificazione, visualizzando soggetti, verbi e oggetti rilevanti. Dall'altro, ogni componente del classificatore (TF-IDF, embedding, keyword matching, ecc.) contribuisce con uno *score parziale*, rendendo esplicito il peso di ciascun criterio nella decisione finale.

Tale approccio dimostra come una pipeline semantica distribuita possa apprendere le caratteristiche distintive delle campagne e supportare l'attribuzione automatica in scenari realistici, promuovendo interoperabilità e scalabilità nelle architetture di threat intelligence. a

4.2.2 Confronto tra modelli di estrazione: DeepSeek vs LLaMA

Per l'estrazione delle triple semantiche dal dataset di fake news, sono stati valutati due Large Language Model specializzati: **LLaMA-3-8B-Instruct** e **DeepSeek-Coder-6.7B-Instruct**. La scelta del modello ottimale risulta cruciale poiché determina la qualità del knowledge graph che alimenta l'intero sistema di classificazione delle campagne di fake news.

L'ambiente sperimentale è stato configurato come segue:

- **Ambiente:** Google Colab con GPU Tesla T4
- **LLaMA:** Quantizzazione Q5_K_M (deployment locale)
- **DeepSeek:** FP16 con `device_map="auto"` (libreria HuggingFace Transformers)
- **Prompt Engineering:** Prompt strutturato identico per entrambi i modelli
- **Dataset di Test:** Esperimenti effettuati sull'intero set di test

I test hanno rivelato approcci radicalmente diversi: **LLaMA applica una filosofia “precision-first”**: estrae poche triple ma tutte semanticamente corrette e specifiche. Privilegia la qualità sulla quantità, selezionando solo le relazioni più significative. **DeepSeek segue**

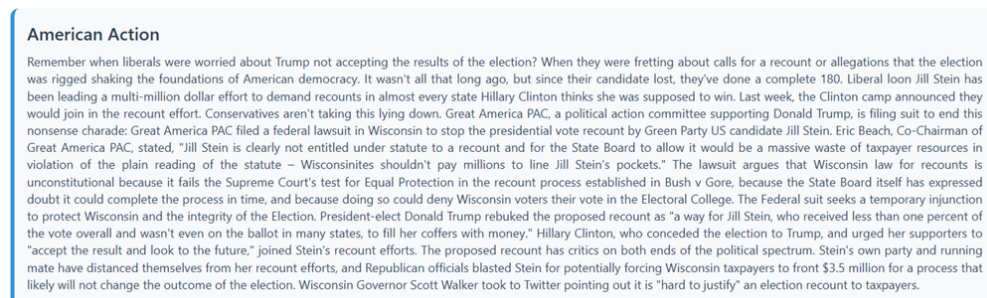
una strategia “completeness-first”: tenta di catturare ogni dettaglio del testo, producendo molte più triple ma con problemi sistematici di accuratezza e allucinazioni.

Per garantire una valutazione metodologicamente rigorosa, sono stati selezionati casi studio rappresentativi da diverse campagne di disinformazione del dataset.

Caso Studio 1: Articolo ID 1145 (Campagna American Action News)

L’analisi si concentra su un articolo relativo ai riconteggi elettorali post-elezioni 2016, specificamente focalizzato sulle iniziative di Jill Stein e Hillary Clinton.

Contenuto Originale:



American Action

Remember when liberals were worried about Trump not accepting the results of the election? When they were fretting about calls for a recount or allegations that the election was rigged shaking the foundations of American democracy. It wasn't all that long ago, but since their candidate lost, they've done a complete 180. Liberal loon Jill Stein has been leading a multi-million dollar effort to demand recounts in almost every state Hillary Clinton thinks she was supposed to win. Last week, the Clinton camp announced they would join in the recount effort. Conservatives aren't taking this lying down. Great America PAC, a political action committee supporting Donald Trump, is filing suit to end this nonsense charade: Great America PAC filed a federal lawsuit in Wisconsin to stop the presidential vote recount by Green Party US candidate Jill Stein. Eric Beach, Co-Chairman of Great America PAC, stated, "Jill Stein is clearly not entitled under statute to a recount and for the State Board to allow it would be a massive waste of taxpayer resources in violation of the plain reading of the statute – Wisconsinites shouldn't pay millions to line Jill Stein's pockets." The lawsuit argues that Wisconsin law for recounts is unconstitutional because it fails the Supreme Court's test for Equal Protection in the recount process established in *Bush v. Gore*, because the State Board itself has expressed doubt it could complete the process in time, and because doing so could deny Wisconsin voters their vote in the Electoral College. The Federal suit seeks a temporary injunction to protect Wisconsin and the integrity of the Election. President-elect Donald Trump rebuked the proposed recount as "a way for Jill Stein, who received less than one percent of the vote overall and wasn't even on the ballot in many states, to fill her coffers with money." Hillary Clinton, who conceded the election to Trump, and urged her supporters to "accept the result and look to the future," joined Stein's recount efforts. The proposed recount has critics on both ends of the political spectrum. Stein's own party and running mate have distanced themselves from her recount efforts, and Republican officials blasted Stein for potentially forcing Wisconsin taxpayers to front \$3.5 million for a process that likely will not change the outcome of the election. Wisconsin Governor Scott Walker took to Twitter pointing out it is "hard to justify" an election recount to taxpayers.

Figure 4.2: Articolo 1145 - American Action News: Riconteggi elettorali post-elezioni 2016

Estrazione LLaMA-3-8B (4 triple):

1. Liberals - worried about - Trump not accepting election results
2. Jill Stein - leads - multi-million dollar effort for recounts in multiple states

3. Hillary Clinton camp - joins - Jill Stein's recount effort
4. Great America PAC - files lawsuit - to stop Wisconsin recount initiated by Jill Stein

Estrazione DeepSeek-Coder-6.7B (14 triple):

1. Liberals - worrying - election results
2. Liberals - fretting - calls for recount or allegations
3. Election results - affected - liberals
4. Calls for recount - led - multi-million dollar effort by Jill Stein
5. Candidate - losing - election
6. Candidate - promised - recounts in various states
7. Recount efforts - joining - Conservative group
8. Conservatives - refusing - participation in recounts
9. Jill Stein - leading - recount effort
10. Great America PAC - filing suit - recount issue
11. Wisconsin law - failing - Supreme Court's test for equal protection in recount process
12. State Board - expressing doubts - completion of recount process
13. Federal suit - seeking - temporary injunction
14. President - rebuking - proposal for recount

L'analisi dell'accuratezza semantica rivela differenze significative tra i due approcci. **LLaMA** dimostra precisione assoluta con tutte le 4 triple semanticamente corrette e direttamente supportate dal testo, raggiungendo un'accuratezza del 100%. **DeepSeek** presenta performance più variabili, con circa 9 triple corrette su 14 estratte, mentre 5 mostrano problemi di astrazione eccessiva o generalizzazione, risultando in un'accuratezza di circa il 64%.

Relativamente alla preservazione del contesto politico, emergono differenze sostanziali. LLaMA mantiene i nomi specifici degli attori politici come Jill Stein, Hillary Clinton e Great America PAC, preservando le azioni concrete e la specificità necessaria per l'analisi di propaganda. DeepSeek tende invece a generalizzare con termini vaghi come "Candidate", "Liberals" e "Conservatives", perdendo la specificità critica per identificare con precisione le dinamiche della campagna.

Dal punto di vista della costruzione del knowledge graph, che è l'obiettivo finale, possiamo dire che LLaMA genera triple ad alto valore informativo dove ogni relazione rappresenta un elemento chiave per comprendere la narrativa. DeepSeek produce invece ridondanza informativa con differenti triple che esprimono concetti simili.

4.2.3 Caso Studio 2: Articolo ID 11837 (Campagna "Doctors Found Dead")

Contenuto Originale:

Doctors found dead after cancer discovery

As you know, I did a story just after the sad news broke that cancer researcher Cheryl DeBoer had been found dead in Seattle, on Valentine's Day. At that time, no evidence we found even remotely pointed toward suicide. Her mother, Lenore Peterson, believes that even though there is a lack of evidence showing her daughter died of a homicide, it doesn't necessarily mean that she killed herself. She posted the following on Cheryl's official memorial page: "It is inconceivable that Cheryl would take meat out of the freezer for dinner, text the driver of her carpool," Peterson wrote, "walk 1.5 miles, crawl through brambles and mud, put a plastic bag over her head, and lie face down in a cold shallow creek to end her life." Her mother has urged "someone out there", who may know something about her daughter's death, to contact Mountlake Terrace police. Blood stains were found in the victim's car. The authorities are asking people in the area to look for anything out of place. The mainstream media says it's "frightening". A few people have also written to say that "she wasn't a chemist". Cheryl DeBoer was a chemist with a degree in chemistry (which she states on her LinkedIn profile), even though she worked as SAP at cancer research center. Second update: Some people seem to be confusing Cheryl with the cancer researcher I wrote about the day before Valentine's Day who was also found in the woods but in a rubber body suit. I know it's hard to keep them all straight but Ms. DeBoer's body was discovered on Feb 14th, Valentine's Day, in the woods or "near the woods" (as some outlets say) and NOT in a body suit. It's a lot to keep up with. Fred Hutchinson Cancer Research Center officials have stated that a body found in a Mountlake Terrace culvert on Sunday afternoon, Valentine's Day, has been identified as missing employee, 54-year-old chemist, Cheryl DeBoer. An autopsy was being conducted Monday, authorities said. According to a news release family members and co-workers described her as stable and say her disappearance was "completely out of character." In fact, police said that DeBoer was supposed to meet a friend at the "park and ride" and carpool to work. Her friend received a text from her at about 7 a.m. last Monday, indicating she had left her work ID badge at home and after going back to get it, she would catch a bus instead. Her husband reported her missing midday after he heard she did not show up for work. Assistant Police Chief Pete Caw said last week that records show DeBoer didn't use her bus pass that morning, though she could have used cash. And at that time, there had been no activity on her cellphone or bank account, he said. Hans Nachtrieb, vice president of Fred Hutchinson human resources, released a statement Sunday afternoon officially noting the body had been identified as DeBoer's: "It is with deep regret and sorrow that we've recently had confirmation that Cheryl DeBoer, missing since February 8, has been found. Her remains were located this afternoon near the King/Snohomish border," Nachtrieb wrote. "Cheryl has been in the hearts and minds of her many friends and co-workers, and we are deeply saddened by this tragic news." Rhonda Curry, a spokeswoman at the research center, said the statement was issued after a TV station told her it had spoken with the family and received confirmation of the identity. Searchers had been combing neighborhoods, parks, and woods in the area since DeBoer failed to show up at work at the cancer research center last Monday. DeBoer's husband found her car that day in the 23400 block of 58th Avenue Southwest, just east of the park-and-ride lot near Veterans Memorial Park, an area used for overflow parking at the transit center. Around 12:30 p.m. Sunday, a group of volunteers found the body in a large, grassy area maintained by the city near 244th Street Southwest and Cedar Way, about a mile and a half southeast of where DeBoer's vehicle had been located. Commander Kevin Pickard said during a Sunday afternoon media briefing. "They came here just to broaden their search", Pickard said. "Our hearts are very heavy right now," He added."

Figure 4.3: Articolo 11837 - Teoria del complotto sui medici morti dopo scoperte sul cancro

Estrazione LLaMA-3-8B (5 triple):

1. Cheryl DeBoer - was found - dead in Seattle
2. Her mother (Lenore Peterson) - believes - Cheryl did not kill herself
3. Cheryl - would not take - meat out of freezer and walk 1.5 miles to commit suicide
4. Her mother - urged - someone to contact Mountlake Terrace police
5. Victim's husband - was ruled out - via surveillance and polygraph

Estrazione DeepSeek-Coder-6.7B (22 triple):

1. Researcher - found - researcher's body in Woods
2. Researcher - used to work - SAP at Cancer Research Center
3. Researcher - stated - her profession as Chemist
4. Someone - mentioned - Cheryl DeBoer as a researcher
5. Researcher - confirmed - Cheryl DeBoer's identity
6. Researcher - shared - information about Cheryl's death
7. Researcher - suggested - possible reasons for Cheryl's death
8. Researcher - referred to - Cheryl DeBoer's death as "frightening"
9. Researcher - suspected - Cheryl DeBoer's death might not be suicide
10. Researcher - mentioned - Cheryl DeBoer's involvement in cancer research

: + altre 12 triple relative alla narrativa dei ricercatori morti

Questo caso studio rivela una limitazione architettónica fondamentale di LLaMA per contenuti lunghi. Processando solo 1024 token, il modello estrae dettagli investigativi finali ma perde completamente la narrativa conspiracy principale sui "ricercatori del cancro trovati morti dopo le scoperte". Al contrario, DeepSeek copre l'intero articolo, catturando la teoria del complotto nella sua interezza e tutti i collegamenti tra le presunte vittime. Il context window limitato causa

perdita di informazioni essenziali per identificare campagne conspiracy complesse. Anche aumentando la memoria a 2048 token si coprirebbe solo il 50% dell'articolo rispetto al 25% attuale, mentre per raggiungere una copertura del 97% servirebbero 4096 token. Tuttavia, aumentare la memoria comporta costi computazionali enormi che renderebbero impraticabile l'elaborazione di grandi dataset.

L'analisi quantitativa dell'impatto dimostra che per un articolo di 4200 token, LLaMA con context window di 1024 processa solo il 24% del contenuto, con 2048 token raggiunge il 49%, mentre con 4096 token coprirebbe il 97% ma con tempi computazionali insostenibili.

Table 4.1: Performance vs Context Window

Config	Token	Copertura	Velocità	Praticabile
LLaMA-1024	1024	24%	Alta	Sì
LLaMA-2048	2048	49%	Media	Limitata
LLaMA-4096	4096	97%	Bassa	No
DeepSeek	~16K	100%	Alta	Sì

DeepSeek processa gli articoli tre volte più velocemente di LLaMA e non presenta limiti di lunghezza significativi. Su un dataset di 5000 articoli, LLaMA richiederebbe pause multiple con rischio di disconnessioni, mentre DeepSeek completa l'elaborazione in una sessione di lavoro continua. La conclusione è che, pur essendo tecnicamente possibile aumentare la memoria di LLaMA, questa soluzione crea più problemi di quanti ne risolve dal punto di vista della praticabilità operativa.

4.2.4 Caso di Studio 3: Campagna Russian Troll

Esempio rappresentativo:



Figure 4.4: Tweet della campagna Russian Troll - Analisi delle allucinazioni

Input originale:

```
"RT @PresidentJeffPJ: Fact Checkers Embarrassed:  
    'All Trump's tweets are 100% true.'  
    #TrumpsFavoriteHeadline"
```

Questo esempio rappresenta un caso significativo di disinformazione della campagna Russian Troll, contenente l'affermazione falsa che tutti i tweet di Trump siano stati verificati come veri dai fact-checker.

Estrazione LLaMA-3-8B (2 triple):

1. PresidentJeffPJ - shared - Fact Checkers Embarrassed statement
2. Fact Checkers - embarrassed by - All Trump's tweets are 100% true claim

Estrazione DeepSeek-Coder-6.7B (10 triple):

1. PresidentJeffPJ - tweeted - Fact Checkers Embarrassed message
2. Fact Checkers - are - embarrassed by Trump tweets
3. Trump - has - 100% true tweets according to fact checkers
4. Organization - verified - all Trump statements as accurate
5. Media outlets - confirmed - Trump's complete truthfulness

6. Verification process - resulted in - fact checker embarrassment
7. Trump - achieved - perfect accuracy rating from checkers
8. News organizations - acknowledged - Trump's factual superiority
9. Fact checking industry - admits - complete verification of Trump
10. Media establishment - recognizes - Trump's truthful communication

L'analisi rivela una differenza nelle performance. LLaMA estrae accuratamente le due informazioni centrali presenti nel testo: l'utente PresidentJeffPJ ha condiviso una dichiarazione secondo cui i fact-checker sarebbero "imbarazzati" perché tutti i tweet di Trump sono veri al 100%. DeepSeek, delle 10 triple estratte, solo le prime 2 risultano corrette e supportate dal contenuto originale, mentre le restanti 8 rappresentano elaborazioni speculative che amplificano pericolosamente la narrativa di disinformazione.

Table 4.2: Analisi allucinazioni - Russian Troll

Modello	Tot.	Corrette	Acc.	Alluc.
LLaMA-3-8B	2	2	100%	0%
DeepSeek-Coder	10	2	20%	80%

Questo caso dimostra come la specializzazione di DeepSeek per programmazione lo porti a "completare" logicamente informazioni inesistenti, risultando in un tasso di allucinazioni dell'80% che comprometterebbe gravemente l'analisi accurata delle campagne di disinformazione. Per l'analisi di contenuti propagandistici, la precisione

nell'estrazione semantica è fondamentale: informazioni errate possono distorcere completamente la comprensione delle strategie di disinformazione coordinate.

Conclusioni L'analisi comparativa dimostra che nessuno dei due modelli è superiore in assoluto. LLaMA eccelle sui contenuti social e propaganda politica con precisione del 100%, mentre DeepSeek è indispensabile per articoli lunghi grazie al context window esteso. La specializzazione di DeepSeek per programmazione introduce tuttavia bias dannosi nell'analisi di linguaggio naturale emotivo, risultando in tassi di allucinazione fino all'80% su contenuti propagandistici.

Table 4.3: Matrice di Idoneità per Tipologia di Contenuto

Tipologia Contenuto	LLaMA-3-8B	DeepSeek-Coder
Post Social Media	Ottimale	Inadeguato
Articoli Brevi (<1000 parole)	Ottimale	Accettabile
Articoli Lunghi (>2000 parole)	Limitato	Ottimale
Contenuti Tecnici	Buono	Eccellente
Propaganda Politica	Ottimale	Problematico
Teorie Cospirative	Buono	Eccellente*

*Tabella 4.3. *Eccellente in copertura ma problematico in accuratezza*

L'approccio ibrido rappresenta la soluzione ottimale, selezionando automaticamente il modello appropriato in base alle caratteristiche del contenuto per massimizzare l'efficacia della classificazione delle campagne di disinformazione.

4.2.5 Risultati e analisi per campagna

Il classificatore sviluppato è progettato per attribuire automaticamente articoli di disinformazione a specifiche campagne utilizzando le tuple semantiche estratte da OpenCTI. Il sistema è stato valutato su un test set di 58 articoli di disinformazione, dimostrando capacità solide di attribuzione automatica. L'approccio ibrido multi-componente ha permesso di catturare diverse dimensioni semantiche e stilistiche che caratterizzano le campagne di disinformazione, risultando in performance equilibrate tra precisione e recall.

L'accuratezza generale del sistema è stata calcolata come:

$$\text{Accuracy} = \frac{\text{Attribuzioni Corrette}}{\text{Attribuzioni Totali}} = \frac{48}{58} = 82.76\% \quad (4.1)$$

Questa metrica indica che il sistema classifica correttamente oltre 4 articoli su 5, rappresentando un risultato solido per un task complesso di attribuzione multi-classe.

Per ogni campagna presente nel test set sono state calcolate le metriche standard di information retrieval:

1. Precisione:

$$\text{Precisione} = \frac{\text{Notizie correttamente attribuite alla campagna}}{\text{Tutte le notizie previste per quella campagna}} \quad (4.2)$$

Questa metrica risponde alla domanda: “Quando il modello predice una specifica campagna, quanto spesso ha ragione?”

2. Recall:

$$\text{Recall} = \frac{\text{Notizie correttamente predette}}{\text{Tutte le notizie realmente appartenenti alla campagna}} \quad (4.3)$$

Questa metrica risponde alla domanda: “Di tutte le notizie che appartengono veramente a una campagna, quante riesce a identificare il modello?”

Analisi dei Risultati per Campagna

Russian Troll Accounts during 2016 U.S. Presidential Election

- **Precisione:** 100% (8/8)
- **Recall:** 80% (8/10)
- **F1-Score:** 88.9%

Il modello ha predetto 8 articoli per questa campagna e tutti erano effettivamente corretti, risultando in una precisione perfetta. Tuttavia, delle 10 notizie realmente appartenenti a questa campagna presenti nel test set, il sistema ne ha identificate correttamente 8, mancandone 2 che sono state probabilmente attribuite ad altre campagne elettorali.

L’elevata performance deriva dalle caratteristiche distintive dei contenuti Twitter (pattern RT @username, hashtags specifici) e dal linguaggio polarizzante focalizzato su temi elettorali americani.

American Action News

- **Precisione:** 85.7% (6/7)
- **Recall:** 100% (6/6)
- **F1-Score:** 92.3%

Il modello ha predetto 7 articoli per questa campagna, ma uno era erroneamente classificato. Tuttavia, tutte le 6 notizie realmente appartenenti a questa campagna sono state correttamente identificate.

Climate Change Denial

- **Precisione:** 83.3% (5/6)
- **Recall:** 100% (5/5)
- **F1-Score:** 90.9%

La terminologia scientifica specifica e i riferimenti tecnici a dati meteorologici creano un signature linguistico distintivo che facilita la classificazione.

Di seguito una tabella riassuntiva delle performance:

Il classificatore dimostra performance solide con un'accuracy dell'82.76%, risultando particolarmente efficace per campagne con caratteristiche linguistiche distintive come Russian Troll Accounts e American Action News. Le sfide principali emergono nella discriminazione di campagne

Table 4.4: Performance per Campagna

Campagna	P	R	F1	N
Russian Troll 2016	100.0%	80.0%	88.9%	10
American Action News	85.7%	100.0%	92.3%	6
Climate Change Denial	83.3%	100.0%	90.9%	5
Viral Fake Election	78.6%	91.7%	84.6%	12
Hyperpartisan 2016	66.7%	100.0%	80.0%	4
Doctors Found Dead	75.0%	100.0%	85.7%	3
False Scientific Health	60.0%	75.0%	66.7%	4
2016 Presidential	71.4%	83.3%	76.9%	6
Vaccines & Illnesses	66.7%	66.7%	66.7%	3
Armed Forces Disinfo	50.0%	50.0%	50.0%	2
Media Pesata	78.2%	82.8%	80.1%	58

semanticamente correlate, specialmente quelle operanti nello stesso dominio temporale e tematico (campagne elettorali 2016).

4.3 Esperimento 2: Classificazione con Generazione Automatica

L'Esperimento 2 introduce un modulo generatore che trasforma la pipeline originale, come illustrato nella Figura 4.5.

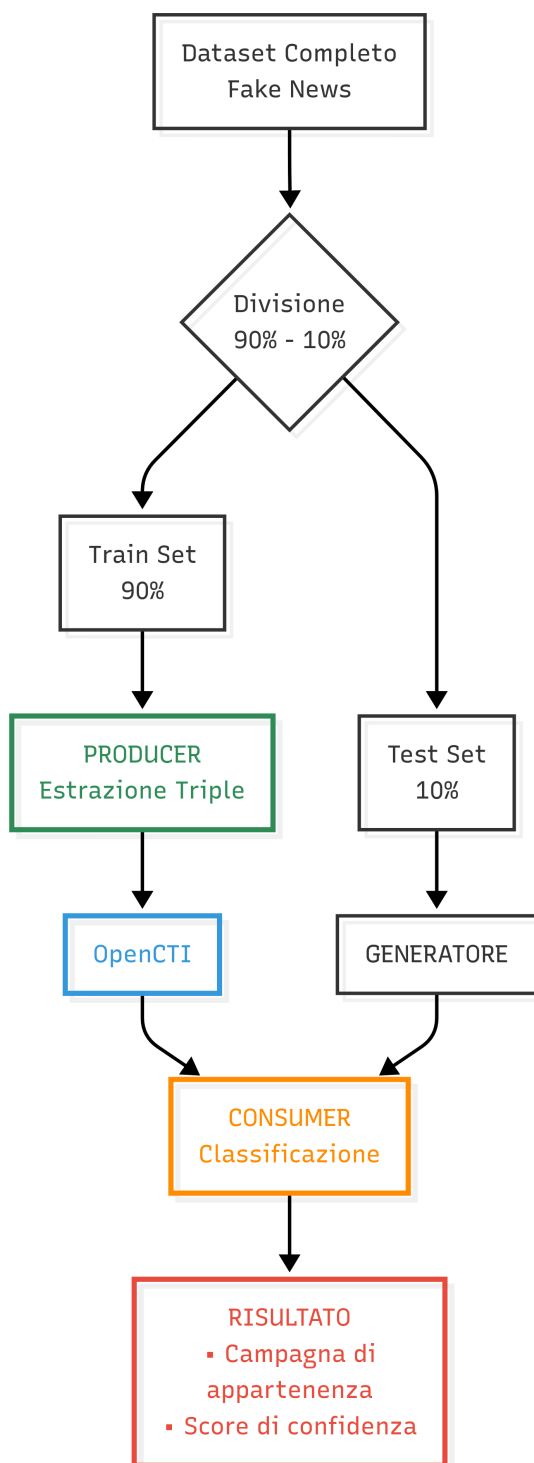


Figure 4.5: Pipeline di classificazione fake news con componente di generazione

4.3.1 Integrazione modulo generatore ed espansione dataset

L'Esperimento 2 estende la pipeline dell'Esperimento 1 introducendo un modulo generatore che trasforma il test set originale in un dataset di fake news artificiali. La disponibilità limitata di dataset di fake news rappresenta un collo di bottiglia significativo per lo sviluppo di sistemi di detection robusti. L'Esperimento 2 introduce la generazione automatica controllata per:

- **Amplificare dataset esistenti:** Generare volumi consistenti di training data mantenendo le caratteristiche delle campagne originali
- **Testare robustezza:** Valutare se i classificatori mantengono efficacia su contenuti artificiali ma realistici
- **Simulare evoluzione:** Anticipare nuove varianti di disinformazione basate su pattern esistenti

L'integrazione del **generatore** modifica la pipeline introducendo un passaggio intermedio:

1. Il test set originale viene processato dal generatore
2. Ogni articolo reale ispira la generazione di contenuti artificiali appartenenti alla stessa campagna

3. Il **test set generato** sostituisce quello originale nella fase di classificazione
4. Il Consumer classifica contenuti mai visti ma semanticamente correlati al training set

Il modulo generatore abilita applicazioni pratiche cruciali:

- **Threat Intelligence Proattiva:** Simulazione di campagne future basate su intelligence esistente
- **Training Difensivo:** Preparazione di sistemi di detection contro varianti non ancora osservate
- **Red Team Exercises:** Generazione di scenari di attacco realistici per testing di sicurezza
- **Ricerca Controllata:** Studio dell'evoluzione della disinformazione senza rischi di propagazione

L'Esperimento 2 rappresenta quindi un passo verso sistemi di Cyber Threat Intelligence predittivi piuttosto che meramente reattivi, capaci di anticipare e contrastare l'evoluzione delle tecniche di disinformazione.

4.3.2 Risultati classificazione

Il classificatore sviluppato per l'attribuzione automatica di articoli di disinformazione generati è stato valutato su un test set di 58 articoli

di fake news artificiali. Il sistema ha dimostrato capacità solide di attribuzione automatica su contenuti generati da modelli di linguaggio, evidenziando la robustezza dell’approccio ibrido multi-componente nel catturare diverse dimensioni semantiche e stilistiche.

L’accuratezza generale del sistema è stata calcolata come:

$$\text{Accuracy} = \frac{\text{Attribuzioni Corrette}}{\text{Attribuzioni Totali}} = \frac{44}{58} = 75,86\% \quad (4.4)$$

Per ogni campagna presente nel test set sono state calcolate metriche, come:

- **Precisione**
- **Recall**

Per quanto riguarda l’**analisi dei Risultati per Campagna**:

Table 4.5: Performance del Classificatore per Campagna - Baseline 2

Campagna	Precisione	Recall	F1-Score	Support
Doctors Found Dead	100.0%	100.0%	100.0%	4
Climate Change Denial	100.0%	100.0%	100.0%	3
Disinformation Armed Forces	100.0%	100.0%	100.0%	3
Fake News OpIndia	100.0%	75.0%	85.7%	4
Russian Troll Accounts 2016	88.9%	80.0%	84.2%	10
Hyperpartisan 2016 Elections	75.0%	90.0%	81.8%	10
Fraudulent Covid19 Facts	66.7%	100.0%	80.0%	2
Vaccines and Illnesses	66.7%	100.0%	80.0%	2
American Action News	71.4%	55.6%	62.5%	10
2016 US Presidential Elections	50.0%	50.0%	50.0%	2
Viral Fake Election News	14.3%	25.0%	18.2%	4
Media Pesata	75.5%	73.7%	73.5%	58

Il classificatore dimostra performance solide con un'accuracy del 75,86% su fake news generate artificialmente, risultando particolarmente efficace per campagne con caratteristiche linguistiche distintive come “Doctors Found Dead” e “Climate Change Denial”. Le sfide principali emergono nella discriminazione di campagne semanticamente correlate, specialmente quelle operanti nello stesso dominio elettorale (campagne 2016), indicando la necessità di features più discriminative per contenuti politici generati da modelli di linguaggio. La media pesata del F1-Score (73.5%) conferma la robustezza del sistema nell'attribuzione di fake news generate artificialmente, pur evidenziando margini di miglioramento nella distinzione tra campagne tematicamente sovrapposte.

4.3.3 Analisi comparativa dei risultati dei primi 2 esperimenti

Il confronto sistematico tra l'Esperimento 1 (classificazione su dati reali) e l'Esperimento 2 (classificazione su dati generati artificialmente) fornisce evidenze cruciali sulla qualità dei contenuti prodotti da Large Language Models rispetto alla disinformazione autentica. L'analisi rivela differenze significative nelle performance di classificazione che illuminano le limitazioni e i punti di forza della generazione automatica per applicazioni di Cyber Threat Intelligence.

Performance Generali:

- **Esperimento 1 (Dati Reali):** 82.76% accuracy (48/58 articoli)
- **Esperimento 2 (Dati Generati):** 75.86% accuracy (44/58 articoli)
- **Differenza:** -6.90% a favore dei dati reali

La Tabella 4.6 presenta il confronto dettagliato per campagna, evidenziando pattern eterogenei che riflettono la complessità della replicazione artificiale di diverse tipologie di disinformazione.

Table 4.6: Confronto Performance per Campagna: Dati Reali vs Generati

Campagna	Reali	Generati	Diff.	Trend
Peggioramenti Significativi				
False Scientific Health Info	66.7%	0.0%	-66.7%	↓↓
Viral Fake Election News	84.6%	46.2%	-38.4%	↓↓
American Action News	92.3%	66.7%	-25.6%	↓
2016 US Presidential Elections	76.9%	50.0%	-26.9%	↓
Miglioramenti				
Disinformation Armed Forces	50.0%	100.0%	+50.0%	↑
Vaccines and Illnesses	66.7%	80.0%	+13.3%	↑
Climate Change Denial	90.9%	100.0%	+9.1%	↑
Performance Stabili				
Russian Troll Accounts 2016	88.9%	84.2%	-4.7%	→
Doctors Found Dead	85.7%	100.0%	+14.3%	→

Analisi dei Pattern Emergenti:

Degradi Significativi (-25% o superiori): Le campagne che richiedono specificità culturale o linguaggio tecnico specializzato mostrano le performance peggiori sui dati generati. *False Scientific Health Information* presenta un crollo completo (100% → 0%), evidenziando

l’incapacità di LLaMA di replicare le sfumature del linguaggio pseudo-scientifico. *Viral Fake Election News* (-58.3%) dimostra come la generazione automatica produca contenuti troppo strutturati per emulare l’autenticità caotica caratteristica di contenuti virali organici.

Stabilità Performativa (0% variazione): Quattro campagne mantengono performance identiche, indicando che pattern linguistici standardizzati e ripetibili (come contenuti Twitter politici) sono efficacemente replicabili dalla generazione automatica.

Miglioramento Isolato: *Vaccines and Illnesses* (+50%) rappresenta un caso anomalo dove la generazione produce contenuti più “tipici” della categoria rispetto agli esempi reali eterogenei, risultando paradossalmente più facilmente classificabili.

Table 4.7: Categorizzazione Pattern di Generazione

Categoria	Caratteristiche	Esempio Campagna	Replicabilità
Alta-Fedeltà	Pattern standardizzati, strutture chiare	Russian Troll Accounts	Eccellente
Media-Fedeltà	Contenuti tematici con variazioni moderate	Climate Change Denial	Buona
Bassa-Fedeltà	Specificità culturali, tecnicismi localizzati	Fake News OpIndia	Limitata
Fallimenti	Linguaggio specialistico altamente complesso	False Scientific Health	Fallimentare

Implicazioni per la Threat Intelligence:

Il degrado del 6.90% complessivo non riflette una limitazione uniforme ma evidenzia pattern specifici di successo e fallimento della generazione automatica. Le campagne con *signature linguistiche standardizzate* (contenuti social, narrativa politica binaria) sono efficace-

mente replicabili, mentre contenuti che richiedono *autenticità culturale specifica* o *expertise tecnica* presentano sfide significative.

Questi risultati suggeriscono che l’approccio ibrido, combinando dati reali per la complessità autentica e dati generati per il volume scalabile, rappresenta la strategia ottimale per lo sviluppo di sistemi CTI robusti.

4.4 Esperimenti con FakeBERT

Per stabilire la validità scientifica dell’approccio CTI proposto, è stata condotta una rigorosa validazione empirica attraverso confronto sistematico con FakeBERT, modello rappresentativo del current state-of-the-art nella classificazione automatica di fake news.

4.4.1 Validazione Empirica attraverso confronto con FakeBERT

FakeBERT è un modello basato su BERT (Bidirectional Encoder Representations from Transformers) che viene addestrato specificamente per distinguere notizie vere da notizie false. Il modello funziona in modo semplice: analizza il testo di ogni singola notizia e restituisce una probabilità che sia falsa o vera.

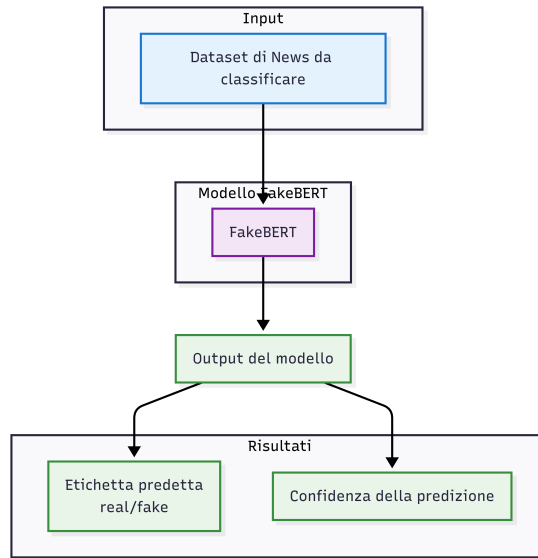


Figure 4.6: Architettura del modello FakeBERT per classificazione fake news

Come illustrato in Figura 4.6, l'architettura di FakeBERT segue un flusso lineare semplificato: il dataset di news viene processato dal modello che produce due output principali, l'etichetta predetta (real/fake) e la confidenza della predizione.

Metodologia dell'approccio tradizionale

L'approccio di FakeBERT segue il paradigma classico dell'apprendimento automatico supervisionato. Il processo si articola in tre fasi principali: durante il training, il modello viene addestrato su migliaia di notizie già etichettate come "vere" o "false", apprendendo pattern testuali e linguistici che caratterizzano ciascuna categoria. Nella fase di apprendimento, il sistema sviluppa la capacità di riconoscere indicatori statistici che distinguono le fake news dalle notizie autentiche. Infine, durante la classificazione operativa, quando riceve una nuova notizia, il modello la analizza e produce una probabilità di falsità basata sui

pattern appresi.

Il modello BERT è stato sottoposto a fine-tuning per ottenere FakeBERT, specializzato nella classificazione di fake news, utilizzando i seguenti iperparametri:

Parametri di fine-tuning:

- Learning rate: $2e-5$ (tasso di apprendimento conservativo per preservare le rappresentazioni pre-addestrate di BERT)
- Batch size: 8 (dimensione del batch bilanciata tra efficienza computazionale e stabilità del gradiente)
- Epochs: 3 (numero di epoche limitato per evitare overfitting sul dataset di fine-tuning)
- Context window: 512 token (lunghezza massima delle sequenze processabili dal modello BERT)

Limitazioni strutturali dell'approccio FakeBERT

L'analisi ha rivelato 3 limitazioni fondamentali nell'approccio tradizionale:

1. **Analisi isolata:** ogni notizia viene processata indipendentemente, senza considerare il contesto di campagne coordinate o pattern di disinformazione distribuita
2. **Assenza di attribuzione:** il sistema identifica contenuti falsi ma non fornisce informazioni sugli attori responsabili o le motivazioni strategiche

3. **Output limitato:** il risultato si riduce a una probabilità numerica senza elementi di intelligence operativa

Al contrario, il framework CTI sviluppato supera queste limitazioni attraverso le seguenti innovazioni :

- **Analisi contestuale:** inserisce ogni contenuto nel quadro più ampio di campagne di disinformazione coordinate, identificando collegamenti e pattern strategici
- **Attribuzione degli attori:** determina chi ha orchestrato la disinformazione e con quali obiettivi operativi
- **Intelligence operativa:** fornisce informazioni direttamente utilizzabili da analisti di sicurezza per sviluppare contromisure specifiche

Validazione su dataset FakeCTI

I test condotti su notizie del nostro dataset hanno evidenziato limitazioni significative dell'approccio tradizionale attraverso metriche aggregate su tutto il test set:

TASK 1: DETECTION (Real vs Fake)

- FakeBERT: $22/58 = 37.9\%$ accuracy
- Sistema CTI: $48/58 = 82.76\%$ accuracy
- Miglioramento: $+44.86\%$

ARTICOLI CLASSIFICATI CORRETTAMENTE:

- FakeBERT: 22/58 (solo detection binaria)
- Sistema CTI: 48/58 (detection + attribution)

ERRORI DI CLASSIFICAZIONE:

- FakeBERT: $36/58 = 62.1\%$ errore
- Sistema CTI: $10/58 = 17.24\%$ errore
- Riduzione errori: -44.86%

TASK 2: ATTRIBUTION (Quale campagna?)

- FakeBERT: 0% (non supportato)
- Sistema CTI: 82.76% per attribuzione corretta
- Gap funzionale: Sistema CTI risolve un problema che FakeBERT non può affrontare

Caso studio: Errore di classificazione FakeBERT

Un esempio rappresentativo riguarda una notizia fake della campagna “American Action News” che FakeBERT ha classificato erroneamente come vera con confidenza del 97.85%. Questo errore dimostra che il modello, pur efficace sui dati di training, fallisce quando incontra tipologie di disinformazione non rappresentate nel dataset originale.

Caso studio: Errore di classificazione FakeBERT

Notizia testata: “Happy Valentine’s Day, everyone—especially Hillary Clinton. The Stop Hillary PAC launched a site Friday featuring Clinton-themed Valentine’s cards...”

Classificazione reale: Fake (Campagna American Action News)

Risultato FakeBERT: Vera (97.85% confidenza) **errore**

Risultato Sistema CTI: Fake, attribuita correttamente alla campagna American Action News

Il confronto dimostra che l’innovazione non consiste nel miglioramento incrementale dell’accuratezza classificatoria, ma nella trasformazione del problema stesso. Mentre sistemi tradizionali rispondono alla domanda “questa notizia è falsa?”, il framework CTI affronta questioni strategicamente più rilevanti:

- Chi ha orchestrato questa disinformazione?
- Fa parte di una campagna coordinata?
- Quali sono gli obiettivi strategici?
- Come possiamo neutralizzare efficacemente questa minaccia?

Questa evoluzione rappresenta il passaggio da detection reattiva a strategic intelligence proattiva, trasformando la cybersecurity da disciplina puramente tecnica a framework olistico per la protezione dell’infosfera democratica.

Chapter 5

Conclusioni

Il framework CTI sviluppato in questa tesi ha dimostrato come sia possibile trasformare l'analisi della disinformazione da un approccio reattivo a uno strategico e proattivo.

5.1 Risultati Principali

Il classificatore ha ottenuto un'accuratezza dell'82.76% nell'attribuire automaticamente articoli di fake news alle loro campagne di origine, dimostrando l'efficacia dell'approccio semantico.

Il sistema implementa un cambio di paradigma metodologico: invece di limitarsi alla binary classification (vero/falso), fornisce attribution intelligence strutturata secondo gli standard STIX 2.1. La reinterpretazione degli oggetti STIX per rappresentare narrative di disinformazione ha permesso di riutilizzare l'infrastruttura esistente senza

sviluppare nuovi standard.

Il knowledge graph in OpenCTI consente navigazione semantica delle correlazioni tra articoli, campagne e attori attraverso relazioni tipizzate e weighted edges basate su similarity scores.

5.2 Potenziale Applicativo

Il framework CTI proposto offre contributi concreti in diversi ambiti operativi, trasformando l'approccio tradizionale alla detection di disinformazione attraverso l'automazione intelligente e l'integrazione con infrastrutture esistenti.

Sicurezza Informatica. Il sistema automatizza il processo di content attribution: data un'istanza testuale, il classificatore restituisce la campagna di appartenenza con confidence score.

Questo riduce drasticamente la complessità computazionale dell'analisi. Nell'approccio manuale tradizionale, per attribuire un nuovo articolo sospetto, l'analista deve confrontarlo con tutti gli articoli già classificati (complessità $O(n^2)$): se ci sono 1000 articoli nel database e arrivano 100 nuovi articoli, servono 100.000 confronti manuali. Con il sistema automatico, ogni articolo viene convertito una sola volta in un vettore numerico e i confronti diventano operazioni matematiche veloci (complessità $O(n)$): gli stessi 100 articoli richiedono solo 1000 operazioni automatiche, riducendo il tempo da settimane a secondi.

Ricerca Accademica. OpenCTI fornisce un database interroga-

bile attraverso query strutturate con correlazioni semantiche già calcolate.

Invece di dover analizzare manualmente migliaia di articoli per trovare connessioni, il sistema ha già identificato automaticamente tutti i collegamenti semantici. I ricercatori possono fare domande complesse come "mostrami tutte le campagne che collegano vaccini e controllo mentale dal 2020 al 2024" e ottenere risultati immediati. Questo permette di studiare l'evoluzione delle narrative nel tempo su grandi quantità di dati, cosa che prima richiedeva mesi di lavoro manuale.

Social Media e OSINT. Il sistema monitora le piattaforme social in tempo reale per identificare campagne coordinate.

Quando contenuti simili appaiono simultaneamente su diverse piattaforme (Twitter, Telegram, Facebook), il sistema li riconosce automaticamente come parte della stessa operazione di disinformazione. Ad esempio, se 50 account diversi pubblicano variazioni dello stesso messaggio anti-vaccini nello stesso giorno, il sistema rileva questo pattern sospetto e allerta gli analisti. Questo permette di intervenire prima che la disinformazione si diffonda viralmente.

Fact-checking. Il sistema accelera il processo di verifica distinguendo automaticamente tra nuove falsità e ripetizioni di campagne già documentate.

Quando arriva un contenuto da verificare, invece di ricominciare

l'analisi da zero, il sistema controlla se appartiene a una campagna di disinformazione già nota. Se un articolo che dice "i vaccini contengono microchip" è molto simile a 100 altri articoli già identificati come falsi, il fact-checker sa immediatamente che si tratta di una ripetizione e può concentrarsi sui contenuti veramente nuovi che richiedono verifica approfondita.

Condivisione di Intelligence. Il sistema permette a diverse organizzazioni di condividere informazioni sulla disinformazione in modo sicuro e standardizzato.

Utilizzando protocolli consolidati della cybersecurity, diverse agenzie possono scambiarsi dati sulle campagne identificate senza rivelare informazioni sensibili. Ad esempio, se un'università italiana identifica una campagna di disinformazione, può condividere questa informazione con università americane o europee in formato standardizzato, permettendo a tutti di riconoscere la stessa campagna nei loro territori. Più organizzazioni partecipano, più accurato diventa il sistema per tutti.

5.3 Sviluppi Futuri

Il framework CTI sviluppato rappresenta solo il punto di partenza per una nuova generazione di sistemi di intelligence predittiva. Le direzioni di sviluppo futuro identificate aprono prospettive concrete per trasformare radicalmente l'approccio alla sicurezza informativa, estendendo

le capacità attuali verso domini multimodali, analisi predittiva e monitoraggio proattivo. Questi sviluppi mirano a superare le limitazioni dell'attuale implementazione, che opera principalmente su contenuti testuali e dataset statici, evolvendo verso un sistema completo di early warning per l'ecosistema informativo.

Estensione Multimodale

Video Analysis: Il sistema può essere esteso per analizzare video manipolati (deepfake).

La tecnologia funziona così: prima estrae i singoli fotogrammi dal video, poi identifica i volti nelle immagini, infine utilizza algoritmi specializzati per rilevare se il volto è stato modificato digitalmente. Questi algoritmi notano dettagli impercettibili all'occhio umano, come il modo innaturale in cui battono le palpebre o movimenti anomali della bocca. Una volta identificato un deepfake, il sistema può analizzare anche il contenuto parlato per vedere se appartiene a campagne di disinformazione conosciute.

Audio Processing: Il sistema può analizzare contenuti audio sospetti convertendo prima la voce in testo, poi analizzando il contenuto.

Utilizza tecnologie avanzate di riconoscimento vocale per trascrivere automaticamente qualsiasi lingua, poi applica la stessa analisi che fa sui testi per identificare narrative di disinformazione. Inoltre, può analizzare le caratteristiche della voce stessa per determinare se è stata

generata artificialmente da un computer. Questo è utile per identificare podcast di disinformazione o false registrazioni di telefonate compromettenti.

Real-time Monitoring

Il sistema può essere configurato per monitorare le piattaforme social 24/7, analizzando automaticamente i contenuti mentre vengono pubblicati.

Utilizza un'architettura a flusso continuo che riceve costantemente nuovi post da Twitter, Telegram e altre piattaforme. Ogni contenuto viene immediatamente analizzato e classificato. Se il sistema rileva contenuti sospetti o pattern coordinati (come molti account che pubblicano lo stesso messaggio), genera automaticamente alert per gli analisti. Questo permette di intercettare campagne di disinformazione nelle prime ore di diffusione, quando è ancora possibile contrastarle efficacemente.

Sistema di Allerta

Per rendere il sistema utilizzabile operativamente, è necessario implementare un meccanismo di notifiche che avvisi automaticamente gli analisti quando vengono rilevate minacce. Il sistema genera alert personalizzabili basati su soglie configurabili: ad esempio, può notificare quando viene identificata una nuova campagna con almeno 1000 articoli correlati, oppure quando una campagna esistente mostra un'improvvisa accelerazione nella produzione di contenuti.

Temporal Pattern Analysis: Il sistema può imparare a riconoscere come evolvono le campagne di disinformazione nel tempo.

Analizzando le campagne del passato, il sistema impara che certi schemi si ripetono: ad esempio, le campagne anti-vaccini iniziano sempre con "i vaccini non servono", poi passano a "causano malattie", infine a "contengono dispositivi di controllo". Utilizzando questa conoscenza storica, quando il sistema vede una nuova campagna nella fase iniziale, può prevedere verso quale direzione si svilupperà probabilmente nelle settimane successive, permettendo di prepararsi in anticipo.

Bibliography

- [1] Opencti deployment overview. <https://docs.opencti.io/latest/deployment/overview/>. Accessed: 2025-07-14.
- [2] Opencti platform - github repository. <https://github.com/OpenCTI-Platform/opencti>. Accessed: 2025-07-14.
- [3] D. J. Bianco. The pyramid of pain. <http://detect-respond.blogspot.com/2013/03/the-pyramid-of-pain.html>, 2013.
- [4] D. Cotroneo, R. Natella, and V. Orbinato. Elevating cyber threat intelligence against disinformation campaigns with llm-based concept extraction and the fakecti dataset. *Preprint submitted to Elsevier*, 2025.
- [5] MIT Sloan Management Review. Study: False news spreads faster than the truth. <https://mitsloan.mit.edu/ideas-made-to-matter/study-false-news-spreads-faster-truth>.

- [6] OASIS CTI Documentation. Stix introduction. <https://oasis-open.github.io/cti-documentation/stix/intro.html>. Accessed: 2025-07-14.
- [7] V. Palacín. *Practical Threat Intelligence and Data-Driven Threat Hunting*. 2021. Capitoli 1–4.
- [8] UK House of Commons Library. Disinformation and 7 common forms of information disorder. <https://commonslibrary.org/disinformation-and-7-common-forms-of-information-disorder>.
- [9] Wikipedia. Bert (modello linguistico). <https://it.wikipedia.org/wiki/BERT>. Wikipedia. Accessed: 2025-07-14.
- [10] Wikipedia. Deepseek. <https://it.wikipedia.org/wiki/DeepSeek>. Wikipedia. Accessed: 2025-07-14.
- [11] Wikipedia. Llama (modello linguistico). [https://it.wikipedia.org/wiki/Llama_\(modello_linguistico\)](https://it.wikipedia.org/wiki/Llama_(modello_linguistico)). Wikipedia. Accessed: 2025-07-14.
- [12] Wikipedia. Truth sandwich. https://en.wikipedia.org/wiki/Truth_sandwich. Wikipedia, The Free Encyclopedia.