CORSO DI DOTTORATO DI RICERCA IN BIG DATA MANAGEMENT

XXXV-ciclo

COORDINATORE: PROF: VALERIO ANTONELLI

Abstract Tesi di Dottorato

Candidato: Domenico Alfano

Tutor: Prof. Domenico Parente

La trasformazione digitale ha innescato un cambiamento epocale nel modo in cui i dati vengono generati e raccolti, inaugurando un'era di informazioni senza precedenti. Questo fenomeno, spesso descritto come un'autentica "esplosione dei dati", ha aperto le porte a un universo ricco di informazioni utili a numerosi campi del sapere. Tuttavia, ha anche introdotto notevoli sfide nel trattamento e nell'analisi di questo immenso volume di dati. Una delle sfide più rilevanti è la capacità di individuare schemi non ordinari o deviazioni, un campo che si è affermato come ambito di ricerca cruciale nell'analisi dati.

In questo contesto, l'Anomaly Detection, ovvero la rilevazione delle anomalie, assume un ruolo centrale. Tale approccio è fondamentale per distinguere i dati rilevanti all'interno di un mare di informazioni.

Questo lavoro si propone di esaminare e approfondire l'importanza del Machine Learning nell'affrontare le complessità legate alla rilevazione di anomalie. L'attenzione si concentra su tre aspetti principali: l'analisi delle irregolarità statistiche, l'indagine sulla profondità semantica dei dati, e l'identificazione di errori.

Il secondo capitolo della tesi, infatti, delinea i fondamenti teorici dell'Anomaly Detection, offrendo una visione approfondita delle definizioni, delle sfide e delle metodologie chiave in questo campo. Dall'importanza di comprendere la natura delle irregolarità nei dati all'applicazione di algoritmi avanzati di Machine Learning e Natural Language Processing.

Questo capitolo fornisce la base concettuale necessaria per comprendere l'ambito della nostra ricerca. Il nucleo della ricerca si sviluppa nei successivi tre capitoli, ciascuno dedicato a una prospettiva specifica dell'Anomaly Detection, offrendo applicazioni pratiche delle metodologie attraverso casi di studio specifici. Attraverso l'analisi di risultati ottenuti in contesti del mondo reale, si vuole dimostrare l'efficacia delle proposte metodologiche in situazioni applicative concrete fornendo un quadro critico degli approcci tradizionali e delle ultime innovazioni nell'Anomaly Detection. Attraverso questa analisi, si vuole posizionare il lavoro all'interno del contesto attuale, identificando le lacune nella letteratura esistente e delineando l'originalità dei contributi.

Nel terzo capitolo, viene affrontata l'analisi delle irregolarità statistiche, esaminando come le tecniche di Machine Learning e di Natural Language Processing possano identificare pattern anomali nei dati.

Attraverso l'uso di modelli e tecniche di apprendimento supervisionato, si superano le limitazioni dei metodi tradizionali, migliorando la precisione e l'affidabilità della rilevazione di anomalie nel campo delle criptovalute. Le criptovalute, con la loro crescente popolarità come veicolo di investimento, hanno portato ad un aumento significativo delle opportunità e, purtroppo, delle truffe.

Tra le frodi più diffuse si evidenzia il Pump and Dump (P&D), una tattica utilizzata da speculatori senza scrupoli per manipolare artificialmente il valore di una criptovaluta.

Il Pump and Dump è una frode ben orchestrata che coinvolge la manipolazione del mercato attraverso la

diffusione di informazioni false al fine di aumentare artificialmente il prezzo di una criptovaluta. In questo contesto, gli operatori fraudolenti accumulano gradualmente un numero significativo di token di una determinata criptovaluta. Successivamente, promuovono attivamente questa criptovaluta, diffondendo notizie false o esagerate su di essa attraverso canali online come social media e chat di gruppo.

Questa fase di promozione è chiamata “pumping”. Una volta che il prezzo della criptovaluta è stato artificialmente gonfiato a causa della manipolazione del mercato, i truffatori procedono a vendere rapidamente i token accumulati a investitori ignari, causando una rapida caduta del prezzo. Questa fase di vendita è nota come “dumping”. Gli investitori che hanno acquistato la criptovaluta basandosi sulle informazioni false durante la fase di pumping subiscono perdite significative quando il prezzo crolla improvvisamente.

La rilevazione di schemi Pump and Dump rappresenta una sfida complessa, ma negli ultimi anni sono stati sviluppati approcci innovativi, spesso basati sull’analisi avanzata del linguaggio naturale e dei social media. Uno dei metodi più promettenti per individuare schemi Pump and Dump è l’analisi dei social media e delle chat online, dove gli operatori fraudolenti spesso coordinano le loro attività. In particolare, piattaforme come Telegram sono diventate luoghi privilegiati per la pianificazione e l’esecuzione di P&D. Il Natural Language Processing si presenta come una risorsa cruciale nella lotta contro il Pump and Dump. Questa disciplina consente l’analisi avanzata del linguaggio naturale utilizzato nei messaggi online, consentendo di identificare pattern e tendenze che potrebbero essere indicative di attività fraudolente.

Attraverso il NLP, è possibile estrarre opinioni e stati d’animo dagli utenti dei social media e delle chat. L’identificazione di segnali linguistici che suggeriscono manipolazioni intenzionali, come l’uso di espressioni iperboliche o la diffusione di informazioni contraddittorie, può contribuire alla rilevazione precoce di P&D. I truffatori spesso creano gruppi pubblici specifici per diffondere informazioni false e coordinare il Pump and Dump. Il NLP può essere impiegato per monitorare questi gruppi, identificando modelli comportamentali sospetti e individuando l’uso di linguaggio fuorviante. Il costante progresso nella tecnologia e nello sviluppo di algoritmi avanzati consente di affinare le tecniche di rilevazione. Gli algoritmi di apprendimento automatico possono essere addestrati su grandi quantità di dati storici, identificando i tratti distintivi di schemi Pump and Dump e adattandosi continuamente alle nuove strategie adottate dai truffatori. In conclusione, questo studio dimostra la possibilità di identificare uno schema di Pump and Dump non appena inizia, sottolineando che gli scambi finanziari hanno capacità di rilevamento superiori grazie all’accesso a dati più dettagliati.

Questi dati includono informazioni approfondite sulle operazioni, i loro volumi e gli attori coinvolti nei Pump and Dump. Lo studio ha inoltre dimostrato l’importanza delle informazioni dai social media nel rilevamento dei P&D nelle criptovalute.

Grazie all'utilizzo di caratteristiche provenienti dai social media, sono stati superati gli approcci precedenti e raggiunti risultati che si posizionano nelle prime posizioni dello stato dell'arte.

Il quarto capitolo si concentra sulla valutazione della ricchezza semantica nei dati, introducendo il ruolo della comprensione semantica e su come possa arricchire il processo di identificazione delle anomalie, considerando contesti e relazioni che vanno oltre le semplici irregolarità statistiche. Attraverso questa prospettiva, si vuole cogliere la complessità e la profondità delle informazioni contenute nei dati, migliorando ulteriormente le capacità di rilevazione di anomalie. L'attenzione in questo caso è sul testo legale, dove le anomalie sono considerate una ricchezza semantica da sfruttare per segmentare correttamente il testo.

Nel complesso mondo dei documenti legali, la segmentazione del testo svolge un ruolo di fondamentale importanza. I documenti legali sono tipicamente costituiti da una serie di segmenti che, pur essendo semanticamente coerenti e legati da un filo logico comune, possono variare notevolmente in termini di lunghezza e tematiche trattate. Questa caratteristica strutturale li rende particolarmente complessi e spesso difficili da navigare, soprattutto per chi non è esperto nel campo legale.

La segmentazione del testo entra in gioco proprio per affrontare questa sfida: essa permette di suddividere documenti testuali prolissi e continui in unità più piccole, definite segmenti, che sono più facili da gestire e analizzare.

Nell'ambito legale, dove il tempo è spesso un fattore critico e l'accuratezza è imperativa, una segmentazione efficace del testo può rappresentare un notevole vantaggio. Per i professionisti del diritto, avere la capacità di identificare rapidamente e con precisione le sezioni rilevanti di un documento può significare una migliore comprensione del caso e una pianificazione strategica più efficiente. Inoltre, la segmentazione del testo si rivela un passo cruciale all'interno delle pipeline di Natural Language Processing legali, soprattutto nelle fasi preliminari di elaborazione di un documento, quando è necessario preparare il testo per operazioni più complesse come il Riconoscimento di Entità Nominate (NER) o il riassunto del testo.

Tuttavia, la segmentazione dei documenti legali non è priva di sfide. La terminologia legale, per sua natura, è complessa e piena di specificità che possono variare notevolmente da un contesto all'altro. Inoltre, la lunghezza dei documenti legali può essere estesa, con testi che spesso si estendono per centinaia o addirittura migliaia di pagine. Questi fattori, combinati con il potenziale rumore introdotto da testi manoscritti, scannerizzati o in qualche modo alterati, rendono la segmentazione un compito particolarmente arduo. In risposta a queste sfide, sono emersi vari approcci neurali alla segmentazione del testo, alcuni dei quali sono stati specificamente progettati per adattarsi al contesto legale.

Questi metodi utilizzano tecnologie avanzate di intelligenza artificiale per analizzare i documenti e identificare i vari segmenti in modo efficiente e preciso.

Nonostante questi progressi, molti di questi metodi non incorporano una funzionalità cruciale: la classificazione dei segmenti all'interno della stessa attività di segmentazione del testo. In altre parole, mentre sono capaci di suddividere il testo in segmenti, non categorizzano questi segmenti in base al loro contenuto o alla loro funzione all'interno del documento. Inoltre, molti approcci presentano anche un'altra difficoltà: analizzano l'intero documento e per evitare il problema del token limit, utilizzano tecniche come il chunking che riducono la precisione del modello stesso.

Per superare questi limiti, questo lavoro di tesi introduce un nuovo approccio basato sul rilevamento di "break-point". Questo metodo innovativo non solo identifica i segmenti all'interno di un documento legale, ma classifica anche le linee di testo che segnalano l'inizio di un nuovo segmento.

La base di questo metodo è un modello basato su transformer, una forma avanzata di rete neurale, che è stato pre-addestrato su testi pre-elaborati provenienti dall'Archivio Nazionale della Giurisprudenza Italiana. Questo addestramento preliminare fornisce al modello una solida base di conoscenza sulla struttura e sul linguaggio dei documenti legali, consentendogli di adattarsi e reagire in modo più efficace ai vari tipi di testo che potrebbe incontrare. L'approccio proposto va oltre la semplice segmentazione del testo, integrando sia il rilevamento dei break-point che la classificazione delle sezioni.

Ciò si traduce in una soluzione più completa e versatile per la segmentazione dei documenti legali, che è robusta anche di fronte a perturbazioni testuali e rumore. Infatti, il modello si è dimostrato molto più affidabile di un motore a regole costruito specificamente sulla struttura testuale dei documenti, mostrando una notevole capacità di adattamento e flessibilità.

Importante è sottolineare che, sebbene il modello presentato sia focalizzato sui documenti legali in lingua italiana, la metodologia e le tecniche impiegate possono essere adattate ed estese per affrontare le sfide della segmentazione del testo in documenti legali di altre lingue. Ciò apre la strada a un vasto campo di applicazioni, permettendo di adattare l'approccio a diversi contesti linguistici e giuridici.

Nel quinto capitolo, si considerano le anomalie come possibili errori, analizzando attentamente gli errori nei dati e farli emergere come vere e proprie anomalie. Questa prospettiva mira a migliorare la precisione del processo di rilevazione, fornendo una metodologia più robusta per identificare pattern anomali legati a informazioni errate. In questo caso, l'attenzione si è focalizzata sui sistemi di Optical Character Recognition.

I sistemi di Optical Character Recognition, comunemente noti come OCR, rappresentano una delle innovazioni più significative nel campo del trattamento dell'informazione digitale. L'OCR, che letteralmente significa "riconoscimento ottico dei caratteri", è una tecnologia che permette di convertire diversi tipi di documenti in dati testuali digitali. Questa trasformazione rende il testo leggibile da macchine e quindi suscettibile di essere editato, ricercato e archiviato elettronicamente rivoluzionando il modo in cui gestiamo

e archiviamo le informazioni, rendendo i dati facilmente accessibili, modificabili e ricercabili. Il concetto di OCR non è recente; le sue radici possono essere rintracciate agli inizi del XX secolo, con lo sviluppo delle prime macchine per la lettura dei caratteri. Questi primi dispositivi erano rudimentali e limitati nel riconoscere solo un ristretto set di font e numeri. Tuttavia, con l'avvento dell'era digitale e lo sviluppo dell'informatica, l'OCR ha subito una rapida evoluzione.

I moderni sistemi OCR utilizzano algoritmi complessi e tecniche di apprendimento automatico per identificare caratteri in quasi ogni tipo di font e in diverse lingue, con un'accuratezza sempre maggiore.

Seppur l'avanzamento delle tecnologie OCR, spesso i risultati non sono perfetti e richiedono un ulteriore passaggio chiamato post-processing. Questo processo mira a migliorare la qualità del testo riconosciuto, identificando e correggendo gli errori che si sono verificati durante la scansione e il riconoscimento iniziali.

Il post-processing OCR si divide in due parti principali: la rilevazione degli errori e la correzione degli errori. Entrambe le fasi sono cruciali per garantire l'accuratezza del documento finale.

Il rilevamento degli errori implica l'identificazione delle parti del testo che potrebbero essere state riconosciute in modo errato. Questo può variare da semplici errori di battitura a problemi più complessi come la confusione tra caratteri simili.

Per la correzione degli errori, una volta identificati gli errori, il sistema tenta di correggerli. Ciò può avvenire attraverso diversi metodi, come il confronto con un database di parole corrette o l'analisi del contesto per determinare la parola più probabile. In questo caso, lo studio mira a catturare gli errori derivanti dall'OCR, sviluppando un modello specifico per la rilevazione di queste anomalie.

È stato necessario costruire un data set contenente errori tipici di un sistema OCR, utilizzando la tecnica di "Data Augmentation", comunemente impiegata nell'apprendimento automatico e nel deep learning, per aumentare la quantità e la diversità dei dati di addestramento. Questa pratica è essenziale in presenza di un data set limitato, poiché aiuta a prevenire l'overfitting e a migliorare la capacità di generalizzazione del modello su nuovi dati. La data augmentation testuale aumenta artificialmente la quantità di dati testuali disponibili, migliorando le performance di modelli di machine learning e intelligenza artificiale, specialmente in scenari con dati limitati. Le tecniche comuni includono parafrasi, back translation, introduzione di errori ortografici o grammaticali intenzionali, sostituzione di sinonimi, inserimento casuale di parole, eliminazione casuale di parole e riorganizzazione delle frasi.

Nello specifico di questo studio, la tecnica della perturbazione è stata utilizzata per simulare le condizioni reali in cui i professionisti legali lavorano con documenti testuali digitalizzati tramite OCR.

Il dataset risultante è stato utilizzato per addestrare un classificatore basato su Transformer, pre-addestrato

su testi provenienti dall'Archivio Nazionale della Giurisprudenza Italiana. È stato scelto ITALIAN-LEGAL-BERT per le sue capacità di catturare le sfumature del linguaggio giuridico. Questo studio fa parte di un lavoro più ampio che mira a risolvere il problema della rilevazione degli errori dei sistemi di OCR e a correggere tali errori. Tale lavoro, denominato REMOAC (A Retroactive Explainable Method for OCR Anomalies Correction in Legal Domain), si avvale di metodi all'avanguardia per affrontare le complesse sfide presentate dagli errori OCR in questo ambito specializzato. L'obiettivo primario di REMOAC è migliorare significativamente le prestazioni dei classificatori di testi legali, correggendo gli errori introdotti durante la fase di scansione OCR.

Infine, nel capitolo sei, le conclusioni e le prospettive future sintetizzano i risultati ottenuti, riflettono sull'importanza delle scoperte nel contesto più ampio dell'analisi dei dati e suggeriscono direzioni per la ricerca futura. In questo modo, la tesi fornisce un contributo significativo all'evoluzione dell'Anomaly Detection, integrando in modo coerente le potenzialità del Machine Learning e del Natural Language Processing.