UNIVERSITÀ DEGLI STUDI DI SALERNO

DIPARTIMENTO DI SCIENZA AZIENDALI - MANAGEMENT & INNOVATION SYSTEMS

Dottorato di ricerca in Big Data Management

XXXV Ciclo

Tesi di Dottorato

Machine Learning per l'Anomaly Detection: Analisi di Irregolarità, Ricchezza Semantica ed Errori nei Dati

Tutor:
Ch.mo Prof.
Domenico Parente

Candidato:
Domenico Alfano
Mat. 8801500027

Coordinatore:
Ch.mo Prof.
Valerio Antonelli

ANNO ACCADEMICO 2021/2022

RINGRAZIAMENTI

In questo momento importante che sancisce il completamento degli studi, ho la possibilità di esprimere la mia più profonda gratitudine e il mio sincero apprezzamento a tutti coloro che hanno reso possibile questo traguardo.
Innanzitutto, vorrei esprimere la mia immensa gratitudine ai miei cari genitori. La vostra incondizionata fiducia, il vostro amore, il sostegno e la pazienza infinita hanno costituito la mia roccaforte nei momenti di difficoltà.
Avete sempre creduto in me, anche quando ho dubitato di me stesso, e per questo vi sarò eternamente grato.
A mio fratello, un pilastro fondamentale della mia vita, per la sua comprensione, il suo incoraggiamento e per essere stato sempre al mio fianco. I tuoi consigli, il tuo supporto e la tua presenza hanno avuto un impatto inestimabile sul mio percorso.
Un ringraziamento speciale va alla mia amata, la mia futura sposa, la cui presenza nella mia vita è stata una fonte continua di felicità e ispirazione. Il tuo amore e il tuo sostegno non hanno solo illuminato i miei giorni, ma hanno anche contribuito in modo significativo al mio successo accademico e personale.
Ai miei amici di una vita, i Micheloni, quest'anno compagni di un'avventura incredibile come l'emozionante esperienza di essere maestri di festa del Giglio del Sarto. La vostra amicizia, il lavoro di squadra e i momenti condivisi sono stati e saranno un prezioso rifugio e una fonte di allegria durante la mia vita.

Per finire, vorrei anche esprimere il mio ringraziamento alla mia azienda, Eustema, che ha reso possibile questo percorso formativo. Il supporto, la flessibilità e le opportunità fornite dall'azienda hanno giocato un ruolo cruciale nel consentirmi di bilanciare il mio impegno accademico con le responsabilità lavorative. Non posso trascurare di menzionare il sostegno morale e l'incoraggiamento ricevuti da parte di tutti i miei colleghi.

Cari Roberto, Emilio, Vincenzo e Donato, la vostra guida e i vostri consigli hanno arricchito la mia esperienza professionale e personale.

In conclusione, vorrei dedicare questo traguardo a tutti voi che avete camminato al mio fianco in questo viaggio. Senza il vostro sostegno, la vostra fiducia e il vostro amore, nulla di tutto ciò sarebbe stato possibile.

Grazie di cuore.

ABSTRACT

In questo lavoro, viene affrontata l'applicazione dell'Anomaly Detection e del Machine Learning in contesti diversificati, ossia nel settore finanziario delle criptovalute e nel dominio legale. Ogni studio fornisce una propria interpretazione dell'anomalia: come errori causati dai processi di riconoscimento ottico dei caratteri (OCR) nei documenti legali, come irregolarità nel mercato delle criptovalute e come ricchezza semantica nei documenti legali.

Nel primo studio, sono state esaminate le criptovalute sotto la lente delle irregolarità di mercato. Con il crescente interesse nel settore delle criptovalute, diventa fondamentale comprendere e prevedere i movimenti dei prezzi. Ci si è concentrato specificamente sugli eventi di Pump and Dump, fenomeni manipolativi in cui gli attori del mercato cercano di influenzare artificialmente il valore delle criptovalute. Attraverso un approccio innovativo che combina l'analisi dei social media e tecniche avanzate di Natural Language Processing, sono state sviluppate nuove metriche per identificare queste irregolarità. La metodologia proposta si distingue per la sua capacità di catturare segnali sottili e indicazioni di manipolazione del mercato, fornendo strumenti più precisi per la previsione dei prezzi e la rilevazione delle frodi.

Nel secondo studio, l'attenzione è spostata sul testo legale, dove le anomalie sono viste come una ricchezza semantica. I documenti legali sono notoriamente complessi e densi di informazioni. La sfida qui è segmentare efficacemente

questi documenti per estrarre le parti più rilevanti e significative. Adottando un modello Transformer pre-addestrato su un vasto corpus di giurisprudenza italiana, è stato sviluppato un metodo che rileva e classifica i segmenti di testo in base alla loro pertinenza e ricchezza informativa. Questo approccio rappresenta un significativo passo avanti nella comprensione automatica dei testi legali, permettendo ai professionisti del settore di accedere rapidamente alle informazioni cruciali.

Il terzo studio si concentra sul rilevamento e la correzione di anomalie OCR in documenti legali. Gli errori OCR possono distorcere significativamente il contenuto e il significato dei testi, specialmente in un ambito sensibile come quello legale. La sfida in questo caso è lo sviluppo di un metodo che rileva gli errori OCR; questo è particolarmente utile nel contesto legale, dove la tracciabilità e l'accuratezza delle informazioni sono fondamentali.

In conclusione, questi tre studi illustrano il potenziale trasformativo dell'Anomaly Detection in vari ambiti. Ogni ricerca affronta sfide uniche e presenta soluzioni innovative, dimostrando l'importanza di interpretare e utilizzare le anomalie in modi diversi a seconda del contesto. Insieme, questi lavori forniscono nuove prospettive e strumenti per professionisti nei campi finanziario e legale, evidenziando l'importanza della tecnologia avanzata nell'analisi dei dati e nella presa di decisioni informate.

INDICE

CAPITOLO 1

INTRODUZIONE

L'era digitale ha portato ad una rapida espansione nella generazione e nell'accumulo di dati, creando un ambiente informativo senza precedenti. Questo fenomeno, noto come "esplosione dei dati", ha fornito una ricca fonte di informazioni per diverse discipline, ma ha anche introdotto sfide significative nella gestione e nell'analisi di questo vasto corpus di dati. In particolare, l'identificazione di modelli anomali o irregolarità è diventata un'area critica di ricerca nell'ambito dell'analisi dei dati.

L'Anomaly Detection, o rilevazione di anomalie, si colloca al centro di questo scenario, giocando un ruolo fondamentale nel filtrare segnali significativi da una massa di informazioni. Questa tesi si propone di esplorare e approfondire il ruolo cruciale del Machine Learning nell'affrontare le sfide dell'Anomaly Detection, focalizzandosi su tre prospettive principali: l'analisi delle irregolarità statistiche, l'esplorazione della ricchezza semantica nei dati e l'identificazione degli 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.

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.

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.

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.

CAPITOLO 2

FONDAMENTI TEORICI

2.1 Anomaly Detection

La rilevazione di anomalie è una parte importante dell'analisi dei dati che si concentra sull'identificazione di pattern che non corrispondono al comportamento normale atteso. Questi schemi, noti come anomalie, possono assumere vari nomi come valori anomali, osservazioni discordanti, eccezioni o aberrazioni, a seconda del contesto applicativo. In questo ambito, i termini più comuni sono "anomalie" e "valori anomali", spesso usati in modo interscambiabile. La rilevazione di frodi in ambito finanziario, assicurativo e sanitario, sicurezza informatica, monitoraggio di sistemi critici di sicurezza e sorveglianza militare sono solo alcuni dei molti settori in cui questo processo può essere utilizzato.

L'importanza della rilevazione di anomalie deriva dalla sua capacità di trasformare le anomalie nei dati in informazioni significative e spesso cruciali, applicabili in diversi settori. Ad esempio, un modello di traffico di rete insolito potrebbe indicare un tentativo di furto di dati da un computer compromesso [1]. Similmente, un'immagine MRI anomala può suggerire la presenza di tumori maligni [2], mentre anomalie nei dati di transazioni con carte di credito

possono segnalare furti di identità o di carte di credito [3]. Anche nelle missioni spaziali, rilevare anomalie nei dati dei sensori delle navicelle spaziali può essere fondamentale per identificare guasti nei componenti [4].

Lo studio delle anomalie nei dati ha origini nella statistica, risalendo al secolo scorso [5]. Da allora, si è evoluto attraverso lo sviluppo di numerose tecniche, con contributi da diverse comunità di ricerca. Queste tecniche variano notevolmente; alcune sono state specificamente sviluppate per determinati ambiti applicativi, mentre altre sono più generali e trasversali.

Questo campo si è notevolmente ampliato grazie alla crescita esponenziale dei dati disponibili e all'avanzamento delle tecnologie di calcolo. L'analisi delle anomalie non è più limitata a contesti statistici tradizionali, ma si estende a settori come l'apprendimento automatico, il data mining e l'intelligenza artificiale. Queste tecnologie avanzate hanno migliorato significativamente l'accuratezza, la velocità e l'efficienza della rilevazione di anomalie.

Le anomalie nei dati sono schemi che non corrispondono a un comportamento normale ben definito. Per esemplificare, immaginiamo un set di dati bidimensionali dove la maggior parte delle osservazioni si raggruppa in due regioni normali, etichettate come N_1 e N_2. I punti che si trovano a una distanza significativa da queste regioni, sono considerati anomalie. Ciò che rende queste anomalie particolarmente importanti è il loro "interesse" o la loro rilevanza pratica, rendendole oggetto di attenzione per gli analisti.

La rilevazione di anomalie, seppur correlata, si distingue dalla rimozione [6] e dall'accomodamento del rumore [7]. Il rumore nei dati è definito come un fenomeno che non interessa l'analista, ma ostacola l'analisi. La rimozione del rumore si concentra sull'eliminazione di elementi indesiderati prima dell'analisi dei dati, mentre l'accomodamento del rumore si riferisce alla protezione di una stima del modello statistico da osservazioni anomale [8].

Un altro ambito collegato è la rilevazione di novità [9, 10, 11], che mira a identificare schemi emergenti o nuovi nei dati, come un argomento di discussione inedito in un gruppo di notizie. La differenza principale tra schemi nuovi e anomalie sta nel fatto che i primi, una volta identificati, vengono

generalmente incorporati nel modello normale.
Questa area non è immobile; continua a cambiare man mano che nascono nuovi problemi e nuove applicazioni. Ad esempio, con l'aumento dell'utilizzo dei dispositivi connessi ad Internet (IoT), la rilevazione di anomalie diventa fondamentale per garantire che questi dispositivi e le reti con cui sono collegati siano sicuri. Inoltre, le capacità di rilevazione di anomalie sono migliorate dall'integrazione dell'apprendimento automatico e dell'intelligenza artificiale, che consentono ai sistemi di imparare dai dati e di adattarsi a nuovi schemi di comportamento.
L'anomalia è definita astrattamente come un modello di dati che non segue il comportamento normale previsto. Un approccio diretto alla rilevazione delle anomalie consiste nel definire una regione di comportamento normale e considerare come anomale tutte le osservazioni fuori da questa regione. Tuttavia, questo approccio, apparentemente semplice, presenta diverse sfide significative:

- **Definizione di una Regione Normale**: Stabilire una regione che catturi tutti i possibili comportamenti normali è estremamente difficile. Inoltre, il confine tra comportamenti normali e anomali è spesso sfumato. Di conseguenza, un'osservazione anomala vicina al confine potrebbe in realtà essere normale e viceversa.

- **Azione di Avversari Malintenzionati**: Nel caso di anomalie derivanti da azioni malevole, come la frode con carta di credito o le intrusioni informatiche, gli avversari tendono ad adattare le loro azioni per mascherare le anomalie come comportamenti normali, complicando ulteriormente la definizione di normalità.

- **Evoluzione del Comportamento Normale**: In molti settori, il concetto di comportamento normale è in continua evoluzione, rendendo la definizione attuale potenzialmente inadeguata in futuro.

- **Diversità nei Contesti Applicativi**: La definizione di anomalia varia significativamente a seconda del dominio applicativo. Ad esempio,

una piccola variazione in un parametro medico può essere considerata anomala, mentre una variazione simile in un contesto finanziario potrebbe essere normale. Questa diversità rende difficile applicare tecniche sviluppate in un settore a un altro.

- **Mancanza di Dati Etichettati**: La disponibilità di dati etichettati per l'addestramento e la validazione dei modelli è spesso limitata, costituendo un ostacolo significativo per lo sviluppo e l'implementazione efficace delle tecniche di rilevazione delle anomalie.

- **Presenza di Rumore nei Dati**: Il rumore nei dati, che può apparire simile alle anomalie, rende difficile distinguere e rimuovere le vere anomalie.

Il problema della rilevazione delle anomalie, nella sua forma più comune, è lungi dall'essere semplice a causa della complessità e della varietà delle sfide che comporta. La maggior parte delle tecniche attuali risolve problemi specifici. Questi elementi includono la natura dei dati, la disponibilità dei dati etichettati, il tipo di anomalia da rilevare e il contesto applicativo.
Un aspetto cruciale nella rilevazione di anomalie è la natura dei dati di input. L'input è generalmente una raccolta di istanze di dati, che possono essere identificati come oggetti, record, punti, vettori, modelli, eventi, casi, campioni, osservazioni o entità [12]. Ogni istanza di dati è caratterizzata da una serie di attributi, noti anche come variabili, caratteristiche, funzionalità, campi o dimensioni. Questi attributi possono essere di vari tipi, come binari, categorici o continui, e ogni istanza di dati può essere uni-variata (con un solo attributo) o multi-variata (con più attributi). Nel caso di istanze di dati multi-variati, gli attributi possono essere tutti dello stesso tipo o una combinazione di tipi diversi.
La tipologia degli attributi è determinante per l'applicabilità delle tecniche di rilevazione di anomalie. Per esempio, nelle tecniche statistiche si utilizzano modelli differenti per dati continui e categorici. Analogamente, nelle tecniche basate sul vicino più vicino, la natura degli attributi influisce sulla scelta della

misura di distanza da utilizzare. In alcuni casi, al posto dei dati reali, potrebbe essere fornita una matrice di distanza (o similarità) che rappresenta le distanze reciproche tra le istanze. In questi scenari, le tecniche che richiedono le istanze di dati originali non sono applicabili, come molte tecniche statistiche e basate sulla classificazione.

I dati di input possono essere ulteriormente categorizzati in base alla presenza di relazioni tra le istanze di dati [12]. La maggior parte delle tecniche di rilevazione di anomalie esistenti si occupa di dati di record (o puntuali), dove non si assume alcuna relazione tra le istanze di dati. Tuttavia, in generale, le istanze di dati possono essere correlate tra loro, portando a diversi tipi di dati come dati di sequenza, dati spaziali e dati grafici.

Nei dati di sequenza, le istanze sono ordinate linearmente, come nei dati di serie temporali, sequenze genomiche o proteiche. Nei dati spaziali, ogni istanza è collegata alle sue istanze vicine, come nei dati del traffico veicolare o ecologici. Se i dati spaziali includono anche una componente temporale, sono definiti come dati spazio-temporali, ad esempio i dati climatici. Nei dati grafici, le istanze sono rappresentate come vertici in un grafico e connesse ad altri vertici tramite spigoli.

Questi tipi di dati presentano sfide uniche nella rilevazione di anomalie. Ad esempio, nei dati di sequenza, l'ordine degli eventi può essere fondamentale per identificare comportamenti anomali. Nei dati spaziali, la prossimità fisica o la relazione tra le istanze di dati può influenzare la percezione di ciò che è normale o anomalo. Nei dati grafici, la struttura del grafico stesso e le relazioni tra i nodi diventano cruciali per comprendere le anomalie.

Le anomalie possono manifestarsi in diverse forme e contesti, e la loro classificazione può essere una guida utile per la progettazione di sistemi di rilevazione più mirati. Possiamo distinguere tra:

- **Anomalie Puntuali:** questo è il tipo più semplice di anomalia, dove un'istanza singola di dati è considerata anomala rispetto al resto del set di dati. È il focus principale della maggior parte delle ricerche nel campo della rilevazione di anomalie. Un esempio classico è il rilevamento di

frodi con carte di credito, dove una transazione significativamente diversa dalle abitudini di spesa normali di un individuo viene identificata come anomalia puntuale. Questo tipo di anomalia è relativamente più semplice da identificare poiché si basa sull'analisi di singole istanze di dati senza la necessità di considerare il contesto o le relazioni tra i dati.

- **Anomalie Contestuali:** queste anomalie sono specifiche di un determinato contesto e non sarebbero considerate anomale al di fuori di tale contesto [13]. La rilevazione di anomalie contestuali richiede la definizione di attributi contestuali e comportamentali. Gli attributi contestuali, come il tempo o la posizione geografica, definiscono il contesto per l'istanza di dati, mentre gli attributi comportamentali descrivono aspetti non legati al contesto dell'istanza. Un esempio può essere una temperatura insolitamente alta per una specifica stagione e località, che sarebbe normale in un altro contesto temporale o geografico. Nel dominio delle transazioni con carta di credito, una spesa elevata potrebbe essere normale in un periodo come il Natale, ma considerata anomala in altri periodi dell'anno. La sfida qui sta nell'identificare correttamente il contesto e interpretare i dati all'interno di tale contesto.

- **Anomalie Collettive:** queste si verificano quando una collezione di istanze di dati correlate è anomala nel contesto dell'intero set di dati. Le singole istanze in sé potrebbero non essere anomale, ma la loro combinazione o il loro modello di occorrenza collettiva è anomalo. Questo tipo di anomalia è comune in set di dati dove esiste una relazione intrinseca tra le istanze, come nei dati di serie temporali o nei dati spaziali. Un esempio potrebbe essere una sequenza di transazioni finanziarie che, prese individualmente, sembrano normali, ma insieme indicano un comportamento di frode. La rilevazione di anomalie collettive richiede l'analisi di pattern e relazioni tra gruppi di istanze di dati, che è spesso più complessa rispetto all'analisi di singole istanze.

In ogni categoria, il rilevamento efficace di anomalie dipende dalla

comprensione dettagliata del contesto specifico e dalla natura dei dati. Ciò include la scelta di metodi statistici appropriati, l'uso di tecniche di apprendimento automatico e data mining, e la comprensione delle dinamiche specifiche del dominio applicativo. Ad esempio, mentre le anomalie puntuali possono essere identificate attraverso metodi statistici relativamente semplici, le anomalie contestuali e collettive richiedono un'analisi più sofisticata che tiene conto del contesto, della temporalità o delle relazioni spaziali tra le istanze di dati.

Inoltre, è importante notare che un'anomalia puntuale o collettiva può diventare un'anomalia contestuale quando analizzata in un contesto specifico. Questo significa che la stessa istanza di dati può essere interpretata in modi diversi a seconda del contesto in cui viene analizzata. La scelta di una tecnica di rilevazione di anomalie contestuali è quindi determinata dalla pertinenza delle anomalie contestuali nel dominio di applicazione target.

Un ulteriore aspetto cruciale nella rilevazione di anomalie è la natura e la disponibilità delle etichette associate a ciascuna istanza di dati, che indicano se l'istanza è normale o anomala. Ottenere dati accuratamente etichettati, che siano rappresentativi di tutti i tipi di comportamenti, è spesso estremamente costoso e complesso. Generalmente, l'etichettatura viene effettuata manualmente da esperti, richiedendo un notevole sforzo e risorse per la creazione di un set di dati di addestramento etichettati. In particolare, raccogliere un set di dati di istanze anomale che copra tutti i possibili tipi di comportamento anomalo è più difficile rispetto alla raccolta di etichette per comportamenti normali. Il comportamento anomalo può essere dinamico, con nuovi tipi di anomalie che emergono e per i quali non esistono dati di addestramento etichettati. In alcuni contesti, come la sicurezza del traffico aereo, le istanze anomale possono essere estremamente rare, ma tradursi in eventi catastrofici.

In base alla disponibilità di etichette, le tecniche di rilevazione di anomalie possono operare in uno dei seguenti tre modi:

- **Rilevazione di Anomalie Supervisionata**: Questo approccio

presuppone la disponibilità di un set di dati di addestramento che contenga istanze etichettate sia per la classe normale che per quella anomala. L'obiettivo è costruire un modello predittivo che differenzi le classi normali dalle anomale. Le nuove istanze di dati vengono confrontate con il modello per determinare a quale classe appartengono. I problemi principali in questo ambito includono la scarsità relativa di istanze anomale rispetto a quelle normali nel set di dati di addestramento e le difficoltà nell'ottenere etichette accurate e rappresentative, specialmente per la classe anomala [14, 15, 16, 17, 18]. Tali problemi sono affrontati in letteratura con varie strategie, come l'iniezione artificiale di anomalie in un set di dati normali per creare un set di addestramento etichettato [19, 20, 21].

- **Rilevazione di Anomalie Semi-Supervisionata**: In questo caso, si assume che il set di dati di addestramento contenga istanze etichettate solo per la classe normale. Non richiedendo etichette per la classe anomala, queste tecniche sono più applicabili rispetto a quelle supervisionate. L'approccio tipico è quello di costruire un modello basato sul comportamento normale e utilizzarlo per identificare anomalie nei dati di test. Esempi di applicazione includono la rilevazione di guasti in astronavi, dove gli scenari anomali possono significare incidenti difficili da modellare. Un numero limitato di tecniche assume la disponibilità solo di istanze anomale per l'addestramento [22, 23, 24], ma queste non sono comunemente usate a causa della difficoltà di ottenere un set di dati di addestramento che copra ogni possibile comportamento anomalo.

- **Rilevazione di Anomalie Non Supervisionata**: Queste tecniche non richiedono dati di addestramento, rendendole le più ampiamente applicabili. Operano sulla premessa implicita che le istanze normali siano molto più frequenti delle anomalie nei dati di test. Se questa assunzione non è valida, queste tecniche possono soffrire di un alto tasso di falsi allarmi. Molte tecniche semi-supervisionate possono essere adattate per

operare in modalità non supervisionata utilizzando un campione dei dati non etichettati come dati di addestramento. Questa adattabilità presuppone che i dati di test contengano poche anomalie e che il modello appreso durante l'addestramento sia robusto a queste poche anomalie.

Ogni modalità ha i suoi vantaggi e limitazioni. La rilevazione di anomalie supervisionata, pur essendo potenzialmente più accurata grazie all'uso di dati etichettati, soffre di problemi legati alla disponibilità e alla rappresentatività delle etichette, in particolare per le istanze anomale. Inoltre, il problema dell'equilibrio tra classi (normali vs. anomale) è un'altra sfida chiave, poiché le istanze anomale tendono ad essere molto meno frequenti rispetto a quelle normali.

La rilevazione di anomalie semi-supervisionata, d'altra parte, è più flessibile in quanto richiede etichette solo per la classe normale. Questo la rende adatta in scenari dove le anomalie sono difficili da etichettare o rare. Tuttavia, l'efficacia di queste tecniche dipende dalla capacità di costruire modelli accurati del comportamento normale e dalla loro robustezza nell'identificare deviazioni significative come anomalie.

Infine, la rilevazione di anomalie non supervisionata è la più versatile in termini di applicabilità, poiché non richiede alcun dato di addestramento etichettato. Questo la rende adatta in contesti dove l'etichettatura è impraticabile o dove i dati etichettati sono inesistenti. Tuttavia, la sfida principale in questo approccio è l'assunzione che le anomalie siano significativamente meno frequenti delle istanze normali, il che potrebbe non essere sempre vero e potrebbe portare a un elevato numero di falsi positivi.

Un aspetto fondamentale delle tecniche di rilevazione di anomalie è il modo in cui vengono segnalate le anomalie. Tipicamente, ci sono due principali metodi di output utilizzati in queste tecniche:

1. **Punteggi (Scores)**: Questo metodo implica l'assegnazione di un punteggio di anomalia a ciascuna istanza nei dati di test. Il punteggio riflette il grado di anomalia dell'istanza, permettendo di creare un elenco ordinato di anomalie. Questo approccio offre flessibilità nell'analisi: gli

analisti possono scegliere di esaminare solo le anomalie con i punteggi più alti o impostare una soglia di punteggio specifica per identificare quali istanze considerare anomale. Questo metodo è particolarmente utile in scenari dove l'anomalia è un concetto graduale piuttosto che binario e dove è importante misurare il grado di deviazione dalla norma.

2. **Etichette (Labels)**: In questo approccio, ogni istanza di test riceve un'etichetta binaria che indica se è normale o anomala. Questo metodo fornisce un risultato diretto e chiaro, utile in contesti dove è necessario un'identificazione definitiva e rapida delle anomalie. Tuttavia, a differenza del metodo basato sui punteggi, l'approccio basato sulle etichette non fornisce informazioni sul grado di anomalia di ciascuna istanza. Nonostante ciò, è possibile avere un certo grado di controllo su questo processo tramite la regolazione dei parametri interni della tecnica utilizzata.

Entrambi i metodi hanno i loro vantaggi e sono scelti in base alle esigenze specifiche del contesto di applicazione. Le tecniche basate sui punteggi offrono una maggiore flessibilità e gradazione nell'identificazione delle anomalie, mentre quelle basate sulle etichette forniscono una distinzione chiara e diretta tra comportamenti normali e anomali. La scelta tra questi due approcci dipende dalla natura dei dati, dalle esigenze analitiche e dalle preferenze dell'analista nel trattare e interpretare le anomalie all'interno del set di dati.
In sintesi, i concetti di base sull'Anomaly Detection coinvolgono la definizione chiara di anomalia, la scelta dell'approccio di rilevazione, l'adattamento ai dati specifici e la considerazione dell'evoluzione temporale dei pattern anomali. Questi concetti costituiscono le fondamenta su cui si basa qualsiasi approccio di successo all'Anomaly Detection.

2.2 Machine Learning per l'Anomaly Detection

L'uso del Machine Learning (ML) nella rilevazione di anomalie ha rivoluzionato il modo in cui affrontiamo la complessità dei dati contemporanei. Questo approccio avanzato consente ai sistemi di apprendere autonomamente modelli intricati e di adattarsi dinamicamente alle fluttuazioni nei dati nel corso del tempo. In questo paragrafo, esploreremo dettagliatamente i principi chiave che governano il Machine Learning applicato alla rilevazione di anomalie, concentrandoci sulla selezione del modello, la preparazione dei dati, la fase di addestramento, la gestione delle classi sbilanciate, la valutazione delle prestazioni, l'adattabilità temporale, la gestione delle nuove informazioni e la robustezza del modello.

Il punto di partenza cruciale è la selezione del modello di Machine Learning. Diversi algoritmi disponibili offrono vantaggi distinti, ma la scelta dipende fortemente dalla natura specifica dei dati e dalle caratteristiche delle anomalie che si cercano di identificare. Algoritmi come gli alberi decisionali sono noti per la loro interpretabilità, mentre support vector machines e reti neurali possono catturare relazioni più complesse. I metodi di clustering, come il DBSCAN, si concentrano sulla struttura dei dati e possono essere utili quando le anomalie formano cluster distinti. Una scelta oculata del modello costituisce il fondamento su cui si costruisce l'efficacia dell'intero sistema di rilevazione di anomalie.

La preparazione dei dati svolge un ruolo centrale nel garantire il successo del processo di rilevazione di anomalie. Questa fase abbraccia diverse attività, dalla normalizzazione per ridurre le differenze di scala tra variabili, alla gestione di valori mancanti, alla rimozione di outlier. L'ingegneria delle caratteristiche è particolarmente critica: selezionare le giuste feature può rendere il modello più sensibile alle deviazioni significative. La preparazione accurata dei dati è il punto di partenza per garantire che il modello di Machine Learning sia in grado di riconoscere pattern anormali e di adattarsi alle variazioni nei dati.

La fase di addestramento del modello rappresenta un'opportunità cruciale per trasmettere al sistema la comprensione dei comportamenti normali e anomali. Durante questa fase, il modello è esposto a un set di dati etichettati, che contiene esempi di entrambi i comportamenti. La quantità e la qualità dei dati di addestramento sono determinanti. In scenari in cui le anomalie sono rare, si devono adottare tecniche di campionamento bilanciato per garantire che il modello apprenda in modo equo da entrambe le classi. L'addestramento rappresenta il momento in cui il modello internalizza i pattern intrinseci dei dati, preparandosi a riconoscere deviazioni significative.

La gestione delle classi sbilanciate è una sfida comune nell'ambito della rilevazione di anomalie. Dato che le anomalie sono di solito una minoranza nei dati, il modello potrebbe inclinare verso la classe maggioritaria, generando falsi negativi. Tecniche come il campionamento under-sampling o over-sampling sono fondamentali per affrontare questa disparità. L'equilibrio tra le classi è essenziale per garantire che il modello possa identificare con successo le anomalie senza sacrificare la capacità di riconoscere i dati normali.

La valutazione delle prestazioni è un aspetto importante per determinare l'efficacia del modello nella rilevazione di anomalie. Mentre metriche come precision, recall, l'F1-score e l'area sotto la curva ROC sono ampiamente utilizzate, la scelta dipende dalle specifiche esigenze dell'applicazione. Ad esempio, in scenari in cui i falsi positivi sono costosi, la precision diventa una metrica prioritaria. L'utilizzo di approcci di validazione incrociata è essenziale per ottenere stime robuste delle prestazioni del modello. Un'analisi approfondita delle metriche è necessaria per comprendere appieno le capacità del modello e per identificare eventuali aree di miglioramento.

L'adattabilità temporale è un principio fondamentale quando si affronta la rilevazione di anomalie nel contesto dinamico dei dati temporali. I pattern anomali possono emergere e mutare nel tempo, richiedendo che il modello si adatti alle nuove dinamiche. Tecniche come l'apprendimento incrementale permettono al modello di incorporare dati più recenti senza dover essere completamente ricreato. Inoltre, l'uso di modelli ricorsivi, come le reti neurali

ricorrenti (RNN), è particolarmente utile nella cattura di relazioni sequenziali nei dati temporali, migliorando l'adattabilità del modello.

I modelli di rilevazione di anomalie basati su Machine Learning devono essere in grado di gestire nuovi dati e di adattarsi a nuovi pattern senza richiedere una completa riconfigurazione del sistema. La gestione delle nuove informazioni è cruciale per mantenere la rilevazione di anomalie allineata con l'evoluzione dei dati nel tempo. L'implementazione di strategie di aggiornamento continuo, che consentono al modello di adattarsi progressivamente alle nuove informazioni, è essenziale per garantire che il sistema rimanga robusto e reattivo alle variazioni nei dati.

La robustezza del modello e la sua capacità di generalizzazione sono principi chiave per garantire che il sistema sia in grado di affrontare una varietà di scenari. Un modello robusto è in grado di gestire dati rumorosi o con variazioni inaspettate senza subire un degrado significativo delle prestazioni. La capacità di generalizzazione, d'altra parte, assicura che il modello possa essere applicato con successo a dati nuovi e sconosciuti. Tecniche come la regolarizzazione durante l'addestramento e l'utilizzo di dataset di validazione diversificati sono strumenti cruciali per migliorare la robustezza e la generalizzazione del modello.

Pertanto, i principi del Machine Learning per l'Anomaly Detection sono intrinsecamente collegati, e la comprensione approfondita di ciascun elemento è essenziale per sviluppare modelli efficaci e adattabili. La selezione accurata del modello, la preparazione dei dati scrupolosa, l'addestramento mirato, la gestione delle classi sbilanciate, la valutazione accurata, l'adattabilità temporale, la gestione delle nuove informazioni e la robustezza del modello convergono per formare un approccio complessivo alla rilevazione di anomalie che può essere applicato con successo in una vasta gamma di contesti. La consapevolezza di questi principi guida lo sviluppo di sistemi di rilevazione di anomalie avanzati e adattabili che possono affrontare le sfide mutevoli dei dati del mondo reale.

In questo lavoro di tesi, i modelli utilizzati nella rilevazione di anomalie

sono quelli basati su classificazione [12]: un modello, o classificatore, che apprende da un insieme di istanze di dati etichettati (fase di addestramento) e successivamente classifica le istanze di test in una delle classi utilizzando la conoscenza acquisita (fase di test)

Queste tecniche operano sotto un'assunzione generale: è possibile far apprendere un classificatore che distingua tra classi normali e anomale nello spazio delle caratteristiche dato.

A seconda delle etichette disponibili per la fase di addestramento, le tecniche di rilevazione di anomalie basate sulla classificazione si dividono in due grandi categorie: tecniche di rilevazione di anomalie multi-classe e mono-classe.

Le tecniche di rilevazione di anomalie multi-classe presuppongono che i dati di addestramento contengano istanze etichettate appartenenti a più classi normali [25, 26]. Si addestra un classificatore per distinguere ogni classe normale dalle altre classi. Un'istanza di test è considerata anomala se non viene classificata come normale da nessuno dei classificatori. Alcune tecniche in questa categoria associano un punteggio di confidenza alla previsione fatta dal classificatore. Se nessuno dei classificatori è sicuro nel classificare un'istanza di test come normale, l'istanza viene dichiarata anomala.

Le tecniche di rilevazione di anomalie mono-classe partono dal presupposto che tutte le istanze di addestramento appartengano a una sola classe. Imparano un confine discriminante attorno alle istanze normali utilizzando un algoritmo di classificazione mono-classe, come le SVM mono-classe [27]. Qualsiasi istanza di test che non rientra nel confine appreso viene dichiarata come anomala.

Tuttavia, esistono diversi algoritmi di classificazione per costruire i classificatori:

- **Reti Neurali**: Utilizzate sia in contesti multi-classe che mono-classe. In una configurazione multi-classe, una rete neurale viene addestrata sui dati normali per apprendere le diverse classi normali. In seguito, ogni istanza di test viene fornita in input alla rete neurale. Se la rete accetta l'input, è normale; se lo rifiuta, è un'anomalia [25].

- **Reti Bayesiane**: Utilizzate per la rilevazione di anomalie in contesti multi-classe. Un'approccio di base per dati categorici univariati usa una rete Bayesiana ingenua per stimare la probabilità a posteriori di osservare un'etichetta di classe, data un'istanza di test. La label con la maggiore probabilità a posteriori viene scelta come classe prevista per l'istanza di test.

- **Support Vector Machines (SVM)**: Applicate alla rilevazione di anomalie in contesti mono-classe. Tali tecniche utilizzano metodi di apprendimento mono-classe per SVM e apprendono una regione che contiene le istanze di dati di addestramento. Per ogni istanza di test, si determina se cade all'interno della regione appresa. Se cade all'interno, viene considerata normale; altrimenti, viene dichiarata anomala [28].

- **Motori a Regole**: Imparano regole che catturano il comportamento normale di un sistema. Un'istanza di test che non è coperta da nessuna regola è considerata un'anomalia. Queste tecniche sono state applicate sia in contesti multi-classe che mono-classe.

Il Natural Language Processing (NLP), un ramo dell'intelligenza artificiale dedicato all'interazione tra computer e linguaggio umano, si inserisce in modo cruciale nel contesto della rilevazione di anomalie. Questa disciplina avanzata apre nuove prospettive nell'analisi dei dati basati sul linguaggio naturale, introducendo una dimensione semantica che va oltre le tradizionali tecniche di rilevazione di anomalie basate su dati numerici. In questo paragrafo, esploreremo dettagliatamente il ruolo fondamentale del NLP, concentrandoci sulla sua capacità di comprendere il significato semantico, l'estrazione di informazioni rilevanti, la contestualizzazione dei dati, l'identificazione di pattern comportamentali anomali e l'applicabilità a dati non strutturati.
Un elemento distintivo del ruolo del NLP nella rilevazione di anomalie è la sua capacità di comprendere il significato semantico intrinseco nei testi. Mentre gli approcci tradizionali spesso si concentrano su caratteristiche numeriche o strutturali, il NLP si spinge oltre, esplorando il contenuto linguistico e

semantico per identificare non solo anomalie statistiche ma anche irregolarità concettuali. La sfida qui è trasformare il testo in un formato che il modello possa interpretare in modo efficace, sfruttando tecniche avanzate di vettorizzazione, come Word Embeddings o BERT (Bidirectional Encoder Representations from Transformers) [29], per rappresentare le parole in uno spazio semantico.

La rappresentazione efficace dei testi è cruciale. L'utilizzo di tecniche di vettorizzazione consente di tradurre il significato implicito delle parole in dati interpretabili per il modello. Le rappresentazioni distribuite, come Word Embeddings, catturano le relazioni semantiche tra le parole, mentre modelli più avanzati, come BERT, considerano il contesto delle parole all'interno di una frase. Questa vettorizzazione avanzata è essenziale per facilitare l'analisi del significato semantico e per identificare deviazioni significative nei dati basati sul linguaggio.

BERT è un lavoro innovativo pubblicato dai ricercatori di Google AI Language. Questo studio ha rivoluzionato il campo del Machine Learning, ottenendo risultati di punta in diversi task di elaborazione del linguaggio naturale, come risposta a domande e inferenza del linguaggio naturale.

La principale innovazione tecnica di BERT è l'uso dell'addestramento bidirezionale del Transformer, un modello di attenzione molto utilizzato, nella modellazione del linguaggio.

L'architettura di un modello basato su Transformer è stata introdotta per la prima volta nel paper "Attention is All You Need" da Vaswani [30] ed è diventata una delle architetture più importanti nel campo del Natural Language Processing (NLP) e dell'apprendimento automatico.

Un modello Transformer è progettato per catturare le relazioni a lungo raggio tra le parole in un testo o in una sequenza di dati. È noto per la sua efficienza nel trattare sequenze di lunghezze variabili e per la sua capacità di catturare le dipendenze a lungo raggio grazie all'attenzione autoregressiva.

L'architettura di base di un modello Transformer è la seguente:

- **Input Embedding**: L'input iniziale, come una sequenza di parole o

token, viene convertito in vettori di embedding. Questi vettori sono spesso inizializzati casualmente e quindi addestrati durante il processo di apprendimento. Gli embedding rappresentano il contesto semantico di ciascun token nella sequenza.

- **Posizione degli Embedding**: Per tener conto dell'ordine delle parole nella sequenza, viene aggiunta un'informazione sulla posizione agli embedding. Ciò è fondamentale perché il Transformer non ha una struttura intrinseca di sequenza e deve imparare la relazione tra i token basandosi sulle posizioni relative.

- **Encoder e Decoder**: Il modello Transformer è costituito da due parti principali: l'encoder e il decoder. L'encoder elabora l'input, mentre il decoder genera l'output. In alcune applicazioni, come la traduzione automatica, si usa un solo encoder-decoder, mentre in altre, come il BERT, si usa solo un encoder.

- **Multi-Head Self-Attention**: Questo è uno dei componenti chiave del Transformer. L'attenzione autoregressiva consente al modello di concentrarsi su diverse parti della sequenza di input in modo pesato. È chiamata "multi-head" perché il modello calcola l'attenzione da più "testine" separate, consentendo di catturare diverse informazioni contestuali contemporaneamente.

- **Strati di Feedforward**: Ogni "blocco" dell'encoder e del decoder comprende uno strato di feedforward dopo l'attenzione. Questo strato è costituito da reti neurali completamente connesse che operano indipendentemente su ciascuna posizione della sequenza.

- **Connettività residua e normalizzazione**: Per agevolare la propagazione del gradiente durante l'addestramento, vengono utilizzate connessioni residue e normalizzazione del batch all'interno di ciascun blocco.

- **Stacking dei blocchi**: Gli encoder e decoder sono costituiti da una pila di blocchi identici (generalmente 6, 12, 24 o più) che eseguono operazioni di attenzione multi-head e strati di feedforward in sequenza.

- **Uscita e Decodifica**: Nel caso di modelli di traduzione, il decoder produce una distribuzione di probabilità su tutte le possibili parole in uscita per ciascuna posizione nella sequenza di output. Durante la generazione, viene utilizzata la distribuzione di probabilità per campionare la parola successiva.

Questo approccio si distingue dalle metodologie precedenti, che analizzavano i testi sequenzialmente, da sinistra a destra o combinando le due direzioni. I risultati ottenuti dimostrano che un modello di linguaggio addestrato in modo bidirezionale può raggiungere una comprensione del contesto e del flusso linguistico più approfondita rispetto ai modelli monodirezionali. Gli autori dell'articolo [29] introducono una tecnica chiamata Masked LM (MLM), che rende possibile l'addestramento bidirezionale in modelli dove precedentemente non era fattibile. Grazie a questi progressi, BERT rappresenta un significativo passo avanti nel campo dell'NLP, stabilendo nuovi standard per il trattamento e l'analisi del linguaggio naturale.
L'impiego di BERT per specifici compiti nel campo del linguaggio naturale è notevolmente intuitivo e versatile.
Uno degli usi primari di BERT è nelle attività di classificazione, come la sentiment analysis. In questo contesto, il processo è simile alla classificazione della frase successiva, un compito standard per il quale BERT è stato originariamente ottimizzato. Qui, un ulteriore strato di classificazione viene posizionato sopra l'output del Transformer, che è il primo token della sequenza e serve come rappresentazione aggregata dell'intera sequenza di input. Questo strato di classificazione è addestrato per predire la categoria o il sentimento della frase o del documento fornito.
Nel campo del Question Answering (QA), BERT si rivela particolarmente efficace. In questa applicazione, il modello riceve una domanda in relazione

a una sequenza di testo e deve individuare la risposta all'interno di quella sequenza. L'approccio con BERT implica l'addestramento di due vettori aggiuntivi che identificano rispettivamente l'inizio e la fine della risposta nel testo. Questo permette al modello di localizzare con precisione la porzione di testo che risponde alla domanda posta, sfruttando la sua capacità di comprendere il contesto in modo bidirezionale.

Un altro importante impiego di BERT è nel Riconoscimento di Entità Nominate (NER). In questo scenario, il modello riceve una sequenza di testo e deve identificare e classificare entità specifiche, come persone, organizzazioni, date e altri elementi chiave. Utilizzando BERT, si può addestrare un modello NER alimentando il vettore di output di ogni token in uno strato di classificazione. Questo strato è incaricato di prevedere l'etichetta NER appropriata per ciascun token, permettendo così al modello di riconoscere e classificare accuratamente le entità nel testo.

Durante il processo di fine-tuning di BERT per queste applicazioni, si mantiene la maggior parte dei iperparametri come nell'addestramento originale di BERT. Tuttavia, ci sono alcuni iperparametri che necessitano di essere ottimizzati specificatamente per ogni compito. Questa fase di fine-tuning del modello è cruciale per massimizzare la sua efficacia su un particolare compito di linguaggio naturale.

Il team dietro BERT ha dimostrato l'efficacia di questo approccio, raggiungendo risultati di punta in una serie di task linguistici complessi e sfidanti, illustrando come la flessibilità e l'adattabilità di BERT lo rendano uno strumento prezioso nell'ambito dell'intelligenza artificiale e dell'elaborazione del linguaggio naturale. Attraverso queste varie applicazioni, BERT non solo stabilisce nuovi standard nel campo, ma apre anche la strada a futuri sviluppi nell'analisi e nell'interpretazione automatizzata del linguaggio umano.

Oltre a comprendere il significato semantico, il NLP eccelle nell'estrazione di informazioni rilevanti dai testi. Questo processo implica l'identificazione di entità, relazioni e concetti chiave, arricchendo la comprensione dei dati e contribuendo all'individuazione di anomalie legate a informazioni specifiche.

Ad esempio, in contesti come la rilevazione di frodi finanziarie, il NLP può analizzare testi associati alle transazioni per individuare comportamenti sospetti, parole chiave indicative o relazioni non usuali tra gli elementi.
La comprensione del contesto rappresenta un altro elemento distintivo del ruolo del NLP. La capacità di interpretare il significato in un contesto più ampio consente di discernere tra comportamenti normali e anomalie che potrebbero essere altrimenti sfuggite all'attenzione. Questo approccio contestuale è particolarmente rilevante in scenari in cui le anomalie dipendono dalle circostanze o dal contesto circostante. Il NLP può essere utilizzato per analizzare il contesto che circonda un testo, fornendo al modello una visione più completa delle informazioni e facilitando la rilevazione di anomalie basate su contesti specifici.
Il NLP, attraverso l'analisi dei pattern comportamentali anomali nei testi, offre un'ulteriore dimensione alla rilevazione di anomalie. Nei contesti in cui le interazioni umane sono documentate attraverso testi, come chat, recensioni, e-mail o documenti, il NLP può individuare comportamenti linguistici inusuali o utilizzare l'analisi del sentiment per identificare deviazioni dal normale. Ad esempio, nelle operazioni di sicurezza informatica, il NLP può essere impiegato per analizzare i log delle attività degli utenti, individuando l'uso di linguaggio sospetto o la presenza di parole chiave indicative di attività non autorizzate.
È cruciale sottolineare che l'efficacia del NLP nella rilevazione di anomalie dipende dalla qualità dei dati di addestramento e dalla corretta configurazione dei modelli. La diversità dei dati linguistici e la complessità delle strutture linguistiche richiedono un approccio attento all'elaborazione e all'interpretazione dei dati. La pulizia e la preparazione accurata dei dati, insieme a una comprensione approfondita delle specificità linguistiche del dominio di applicazione, sono essenziali per garantire che il modello di NLP possa apprendere in modo efficace e identificare anomalie in modo affidabile.
In conclusione, il ruolo del Natural Language Processing nella rilevazione di anomalie è multidimensionale e ricco di sfaccettature. La sua capacità di comprendere il significato semantico, estrarre informazioni rilevanti,

contestualizzare i dati, analizzare pattern comportamentali anomali e affrontare dati non strutturati offre un contributo significativo all'efficacia dei sistemi di rilevazione di anomalie. L'integrazione del NLP in queste applicazioni non solo arricchisce la comprensione dei dati basati sul linguaggio naturale ma apre anche nuove possibilità per affrontare le sfide complesse della rilevazione di comportamenti anomali in scenari reali.

CAPITOLO 3

ANOMALIE COME IRREGOLARITÀ: LE CRIPTOVALUTE ED IL FENOMENO PUMP AND DUMP

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, la rilevazione dei Pump and Dump rappresenta una tappa fondamentale nella tutela degli investitori e nella preservazione dell'integrità del mercato delle criptovalute. Attraverso l'impiego di tecniche avanzate come il NLP e l'analisi dei social media, è possibile sviluppare approcci robusti per identificare e prevenire queste frodi, contribuendo così a garantire una partecipazione equa e informata agli investimenti in criptovalute.
Lo scopo di questo studio è quindi quello di applicare le innovazioni nel campo del NLP per creare nuove metriche basate sui social media che possano contribuire al rilevamento dei P&D nelle criptovalute. Si vuole dimostrare che, prendendo in considerazione queste metriche generate, è possibile ottenere risultati significativamente migliori rispetto ai metodi di rilevamento dei P&D esistenti.

3.1 Stato dell'Arte

La crescente popolarità delle criptovalute come piattaforme di investimento ha portato a un'espansione significativa del loro utilizzo. Tuttavia, questo aumento di interesse è accompagnato da una serie di sfide, tra cui la complessità tecnica, rendendo le criptovalute un terreno fertile per le attività fraudolente. In particolare, il fenomeno noto come Pump and Dump (P&D) si è rivelato una truffa diffusa.
Il P&D è una tattica ingannevole in cui gli speculatori cercano di manipolare il mercato alzando artificialmente il valore di una specifica criptovaluta attraverso la diffusione di informazioni false. L'obiettivo finale è quello di vendere rapidamente le monete acquisite a un prezzo gonfiato a investitori ignari. Come mostrato in Figura 3.1, questa pratica si sviluppa in fasi distintive: l'accumulo graduale di token, la fase di "pumping" in cui si diffondono notizie false per aumentare il prezzo, e infine il "dumping", la vendita a prezzi gonfiati.

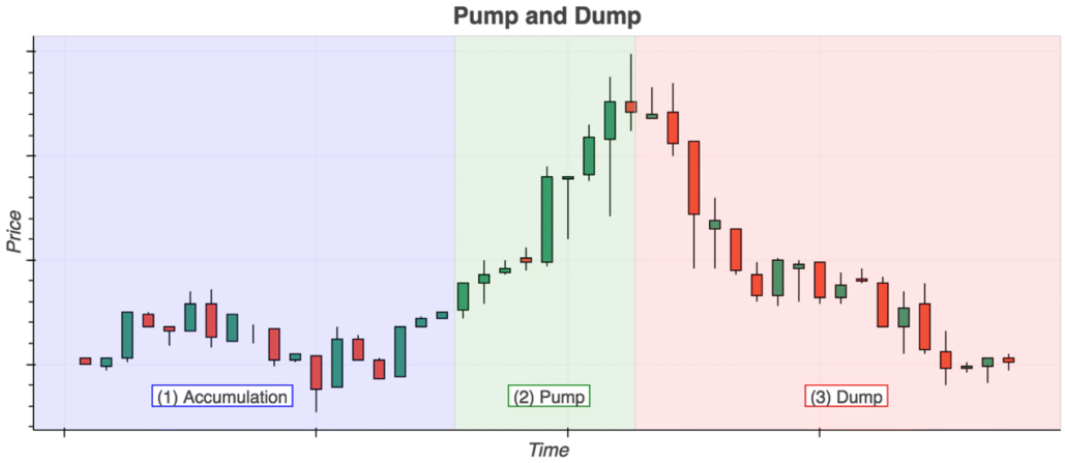


Figura 3.1: Schema Pump and Dump [31]

La creazione di gruppi pubblici specifici è il mezzo attraverso il quale la disinformazione viene disseminata, solitamente attraverso piattaforme di social media come Telegram. Questi gruppi si concentrano spesso su criptovalute meno conosciute con basse capitalizzazioni di mercato, ritenute più suscettibili alla manipolazione. In questo contesto, l'uso di informazioni fuorvianti, come notizie false, false alleanze o sponsorizzazioni fittizie di celebrità, è una pratica comune. Gli utenti vengono spinti a diffondere queste informazioni durante la fase di pumping per influenzare altri investitori [32].

La ricerca suggerisce che i gruppi P&D si dirigono principalmente verso valute meno conosciute, sfruttando la percezione che siano più facili da manipolare per ottenere guadagni considerevoli. In uno schema tipico, i leader del gruppo annunciano la futura manipolazione, inducendo i partecipanti a cercare di acquistare la criptovaluta prima della sua introduzione ufficiale. Questa corsa all'acquisto è spesso accompagnata da false informazioni diffuse per convincere altri a partecipare [31].

Attualmente, la manipolazione di P&D non è sempre considerata illegale, in parte a causa della novità della tecnologia delle criptovalute [33]. Tuttavia, questa situazione presenta rischi significativi per gli investitori. La nostra proposta innovativa si concentra sull'utilizzo di dati liberamente disponibili, sfruttando le API di Telegram per rilevare schemi P&D. Questo approccio, che va oltre i tradizionali dati finanziari, mira a migliorare l'efficacia nella

rilevazione di truffe, contribuendo così a proteggere gli investitori e a preservare l'integrità del mercato delle criptovalute.

L'analisi della letteratura ha guidato l'identificazione e lo studio di pubblicazioni scientifiche dedicate all'individuazione di anomalie nei social media. La ricercatrice Neda Soltani [34] ha svolto un ruolo chiave nel rilevare anomalie e utenti sospetti attraverso l'impiego di tecniche di analisi della rete e lo studio della struttura dei nodi all'interno delle comunità. In un altro interessante contributo [35], è stato proposto il modello di Fama French per misurare e dimostrare l'influenza dei social media come fonte di ispirazione per nuovi investimenti. In questo contesto, è stata analizzata la strategia di investimento suggerita dal gruppo Wallstreetbets per replicarla e generare profitti, con l'analisi del canale sociale che ha suggerito il mercato appropriato in cui investire. Altri ricercatori [36] si sono concentrati sull'identificazione delle attività di manipolazione del prezzo del Bitcoin, utilizzando tecniche di Machine Learning per indagare i periodi di manipolazione attraverso lo studio delle emozioni e dei sentimenti degli utenti nei social media utilizzando algoritmi come SVM e SARIMAX. Diversi studi simili a quelli menzionati sopra evidenziano che i social media sono spesso sfruttati come mezzo per attività fraudolente. Negli ultimi anni, molti di questi studi si sono concentrati sul mercato delle criptovalute e sul fenomeno del Pump and Dump. È interessante notare che tale fenomeno era già in uso nel campo finanziario prima della creazione delle criptovalute. L'esame di lavori analoghi in campo finanziario può fornire preziose intuizioni sulle tecniche utilizzate dai malintenzionati, agevolando l'applicazione di tali metodologie o di nuove nel mercato delle criptovalute.

Uno dei principali mezzi impiegati dai truffatori per manipolare il valore di una valuta è la diffusione di informazioni erronee, spesso con il supporto di bot per automatizzare il processo [37]. Da un punto di vista finanziario, gli schemi di manipolazione del mercato sono generalmente classificati in tre categorie principali: basati sull'informazione, basati sull'azione e basati sul commercio [38]. Un'analisi attenta di questa classificazione permette di

definire gli schemi di Pump & Dump come una combinazione di manipolazione basata sull'informazione e sul commercio. Il lavoro di Kamps [31] rappresenta un primo tentativo di rilevare i Pump and Dump utilizzando un threshold adattativo.

Una delle tecniche di rilevamento delle anomalie utilizzata in questo contesto è una tecnica basata su soglie, la cui ispirazione deriva da ricerche pregresse relative agli attacchi di tipo denial of service su una rete, condotte da Siris e Papagalou nel 2004 [39]. Questa tecnica si basa sulla determinazione di una soglia al di sopra o al di sotto della quale un valore è considerato anomalo. Per implementare questa metodologia, si calcola una media mobile semplice, ovvero una media dei valori precedenti all'interno di una finestra temporale specifica. La lunghezza di questa finestra temporale è comunemente indicata come "lag factor" o fattore di ritardo.

La peculiarità di questa tecnica sta nella possibilità di confrontare un valore con la tendenza generale nel corso di un periodo di tempo, anziché considerarlo in modo isolato. Questo approccio consente di individuare anomalie locali rispetto alla storia recente, identificando deviazioni significative dai normali andamenti.

L'utilizzo di questa tecnica di rilevamento delle anomalie fornisce un punto di partenza funzionale per il monitoraggio e la rilevazione di comportamenti anomali. Tuttavia, rappresenta solo una base su cui ulteriori ricerche e sviluppi possono costruire. L'obiettivo potrebbe essere quello di implementare algoritmi più sofisticati che migliorino la capacità di identificare e interpretare le anomalie in modo più dettagliato e accurato.

Inoltre, considerando che l'analisi riguarda schemi di Pump and Dump nel contesto delle criptovalute, è possibile integrare ulteriori informazioni specifiche del dominio nell'algoritmo di rilevamento delle anomalie. Ad esempio, si potrebbero considerare determinati periodi di tempo, criptovalute specifiche o modelli di trading noti per essere associati a tali schemi fraudolenti. L'obiettivo è aumentare l'accuratezza del rilevamento delle anomalie e adattare il sistema per riconoscere comportamenti sospetti all'interno del contesto delle

criptovalute. In sintesi, questa tecnica basata su soglie rappresenta un punto di partenza solido per il rilevamento delle anomalie, con il potenziale per ulteriori miglioramenti basati sulla conoscenza specifica del dominio delle criptovalute.

3.2 Metodologia

La metodologia proposta si basa sull'utilizzo innovativo e combinato di due diverse tipologie di informazioni inerenti l'analisi delle criptovolaute: dati finanziari e conversazioni social. I dati finanziari per lo studio fanno riferimento alle transazioni dell'exchange Binance e sono stati inizialmente resi pubblici da La Morgia [40]. Per la costruzione di questi data set, sono state etichettate le transazioni di diverse criptovalute che erano state identificate come casi noti di Pump and Dump, partendo dai flussi conversazionali social che facevano riferimento a quelle particolari cripto coinvolte. Più precisamente, gli autori [40], sono partiti analizzando i canali telegram che promuovevano l'acquisto delle criptovalute ed hanno verificato le segnalazioni di pump presenti, forniti dagli stessi amministratori dei gruppi, riuscendo a collezionare 104 occorrenze di dati P&D in due anni. Partendo poi dai timestamp dalle segnalazioni, sono state raccolte le transazioni di bitcoin, moneta utilizzata per acquistare le criptovalute coinvolte, correlate ai pump, attraverso le API di Binance.
Dopo aver ottenuto i dati grezzi dall'API di Binance, sono stati sottoposti a un processo di pre-elaborazione. Le transazioni sono state aggregate in intervalli temporali di cinque, quindici e venticinque secondi, creando tre diversi set di dati che presentano le seguenti caratteristiche:

- **PumpIndex, Symbol**: L'indice di pump e il simbolo della criptovaluta coinvolta nel pump;
- **StdRushOrder, AvgRushOrder**: La variazione percentuale media del numero di ordini urgenti e la deviazione standard mobile;
- **StdTrades**: Il numero di deviazioni standard mobili delle operazioni di acquisto e vendita;

- **StdVolume, AvgVolume**: La variazione percentuale media del volume degli ordini e la deviazione standard mobile;
- **StdPrice, AvgPrice, AvgPriceMax**: La variazione percentuale media, la deviazione standard e la variazione percentuale massima del prezzo.

La manipolazione dei prezzi organizzata attraverso i social media è stata documentata in diversi studi, tra cui quello condotto da Kamps [31]. Questa pratica coinvolge gruppi che variano nell'attività, con quelli meno attivi che effettuano un'operazione di Pump and Dump a settimana, mentre i gruppi più attivi ne effettuano circa una al giorno. In generale, il processo segue le seguenti fasi, come evidenziato da Martineau [41]:

1. **Annuncio Preliminare**: Giorni o ore prima dell'operazione, gli amministratori rendono pubblico l'evento imminente, specificando l'exchange da utilizzare, l'orario preciso dell'inizio e se l'operazione sarà aperta a tutti o sarà riservata a membri VIP e di alto livello.
2. **Consigli Prima dell'Evento**: Poco prima dell'inizio, gli organizzatori forniscono consigli chiari, come verificare la connessione a Internet, acquistare in modo cauto e vendere con decisione, oltre a trattenere la valuta il più a lungo possibile.
3. **Chiusura Temporanea delle Chat Gratuite**: Le chat room gratuite vengono chiuse temporaneamente per evitare "paura, incertezza e dubbio" (FUD), in risposta a tentativi di disturbo dell'operazione e creazione di panico all'interno del gruppo.
4. **Offuscamento del Nome della Criptovaluta**: In base alla gerarchia del gruppo, i membri vengono informati esattamente quando sarà rivelato il nome della criptovaluta bersaglio. Il nome è spesso presentato in un'immagine sfocata, leggibile solo dagli esseri umani per ostacolare l'analisi automatica tramite metodi OCR.

5. **Diffusione delle Informazioni**: Gli amministratori rilasciano notizie poco dopo l'inizio dell'operazione, chiedendo a tutti i membri di diffondere il messaggio che il prezzo della criptovaluta sta aumentando. Questo viene fatto su diverse piattaforme come Twitter, forum e chat speciali, con l'obiettivo di attirare investitori esterni attraverso un'opportunità di investimento apparentemente vantaggiosa.

6. **Analisi Post-Operazione**: Al termine dell'operazione, gli amministratori riaprono le chat gratuite e forniscono agli utenti statistiche dettagliate sull'andamento della Pump and Dump appena conclusa.

Come descritto in precedenza, in aggiunta ai dati delle transazioni, la presente metodologia, basa la sua intuizione ed innovazione sull'utilizzo delle informazioni social come ulteriore caratteristica da considerare nella valutazione della possibilità che avvenga un fenomeno di P&D. Se l'approccio precedente, partiva dai social solo per identificare ed etichettare l'evento di pump, il metodo proposto, ribalta il suo utilizzo facendone diventare parte attiva per la classificazione. Infatti, per ampliare il data set con informazioni provenienti dai social, è stata ricondotta una raccolta di messaggi dal social network Telegram, focalizzandosi sugli orari in cui avvenivano i pump e sui link dei gruppi Telegram interessati. L'accesso diretto a questi gruppi ha permesso di scaricare i messaggi attraverso le API di Telegram, coprendo un periodo di due giorni prima e dopo ciascun pump. Tuttavia, le restrizioni nell'accesso alla cronologia completa dei messaggi hanno limitato la portata del data set raccolto, riducendo il set di dati da 104 a 89 pump.
Successivamente, utilizzando tecniche di Natural Language Processing, è stata effettuata una pre-elaborazione di questi messaggi, che ha coinvolto le seguenti fasi:

- **Rimozione di emoji, immagini, stop-word e link esterni**: processo di riduzione del testo non informativo;

- **Lemmatizzazione**: processo di raggruppamento delle forme inflesse di una parola per poterle analizzare come un unico elemento;

- **Pos Tagging**: processo di marcatura di una parola in un testo come corrispondente a una particolare parte del discorso in base alla sua definizione e al suo contesto, escludendo solo le parole che si riferiscono a sostantivi e nomi propri.

Dopo aver raccolto i messaggi dalla chat pertinente per ciascun pump, si è passati alla fase di estrazione delle caratteristiche.
In particolare, per ciascuno dei tre set di dati creati da La Morgia [40], per ciascuna delle transazioni finanziarie e di conseguenza per ciascun pump definito al loro interno, la nuova caratteristica assume il valore 1 se il simbolo della valuta è presente nei messaggi scambiati 5 minuti prima della transazione, altrimenti 0. Si ottiene così il set di dati finale, composto da 89 pump, aggiungendo solo una caratteristica categorica a quelle finanziarie.
Il classificatore adottato va in continuità con quanto fatto da La Morgia [40], usato proprio per valutare l'efficacia e la rilevanza dell'uso di una nuova caratteristica di Telegram. Si tratta di un classificatore Random Forest, una collezione di alberi decisionali che si basano sui valori di un vettore casuale campionato indipendentemente, dove ciascuno emette un voto, con la previsione rappresentata dalla classe che riceve più voti complessivi [42]. Dato che il data set consiste di 89 P&D, non è stato suddiviso in set di training e test ma è stata applicata una 5-fold cross validation per una valutazione più accurata delle prestazioni.

3.3 Risultati

Nell'analisi di tutti i data set, con l'integrazione della caratteristica sociale, il classificatore è stato in grado di superare l'approccio precedente con un margine statisticamente significativo. Questo dimostra l'efficacia dell'uso di caratteristiche derivanti da dati sociali in quest'area precedentemente

inesplorata. I risultati più promettenti su tutti i data set sono mostrati nella Tabella 3.2 rispetto alla Tabella 3.1.

Chunk-Size	Precision	Recall	F1
5 Sec	89.1%	83.1%	86.04%
15 Sec	97.5%	89.8%	93.5%
25 Sec	97.5%	91.1%	94.1%

Tabella 3.1: Risultati 5-Folds Random Forest considerando solo caratteristiche finanziarie.

Chunk-Size	Precision	Recall	F1
5 Sec	94.2%	91%	92.5%
15 Sec	98.8%	94.3%	96.5%
25 Sec	97.6%	94.3%	95.9%

Tabella 3.2: Risultati 5-Folds Random Forest considerando caratteristiche finanziarie e social

Inoltre, si è scoperto che le previsioni utilizzando il data set suddiviso in frammenti di 5 secondi sono molto meno accurate rispetto a quelle sui data set suddivisi in frammenti di 15 e 25 secondi, il che suggerisce che prevedere anomalie in un lasso di tempo minore corrisponde a un problema più difficile in generale. Questo conferma i risultati di lavori precedenti [40].
Pur avendo ottenuto ottimi risultati, da questi ultimi emerge comunque che è possibile che i gruppi di Pump and Dump riescano ad eludere i sistemi di rilevazione. Nell'analisi condotta, esiste uno scenario in cui il classificatore potrebbe non riuscire a individuare con successo le operazioni di Pump and Dump. Questo scenario si verifica quando gli amministratori di gruppi o alcuni membri VIP iniziano gradualmente ad acquistare una determinata criptovaluta in modo che l'aumento di un parametro chiave, come il prezzo, avvenga in modo graduale e sottile. Nel frattempo, il numero di utenti coinvolti rimane limitato.

In questa situazione, il classificatore potrebbe non riuscire a rilevare il Pump and Dump, in quanto la variazione nei parametri di mercato appare meno drastica e più in linea con le normali fluttuazioni.
Nel data set, questo scenario si è verificato in quattro casi di Pump and Dump, tutti condotti dallo stesso gruppo. In ognuno di questi casi, è stata registrata una fase pre-pump consistente nelle ore che precedono l'inizio dell'operazione. Tuttavia, è importante notare che questa strategia non può essere applicata frequentemente. Un osservatore esperto che conosce l'orario programmato per il prossimo Pump and Dump potrebbe sfruttare questo pattern per individuare in anticipo la criptovaluta oggetto dell'operazione.
Inoltre, va considerato che dopo l'esecuzione di un Pump and Dump di successo, l'operazione diventa evidente per la maggior parte degli utenti del gruppo. Questo potrebbe comportare la perdita di fiducia nei confronti degli amministratori da parte degli utenti e portare all'abbandono del gruppo. Pertanto, anche se questa strategia può occasionalmente sfuggire alla rilevazione, comporta rischi significativi e non può essere sfruttata in modo continuativo senza conseguenze negative a lungo termine.

3.4 Conclusioni e Sviluppi Futuri

In conclusione, questo studio [43] 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.
Tuttavia, si ritiene che l'implementazione di politiche mirate può significativamente ridurre queste manipolazioni di mercato.

Un esempio concreto è fornito dal noto exchange BitTrex, che ha preso una posizione ferma contro la manipolazione di mercato. Dopo aver annunciato misure punitive contro i partecipanti a tali schemi, si è osservato un calo drastico dei casi di Pump and Dump sulla piattaforma.

In futuro, attraverso un mix di monitoraggio avanzato, interventi normativi e politiche attive, è possibile combattere efficacemente le manipolazioni di mercato come i Pump and Dump, proteggendo sia gli investitori che l'integrità dei mercati finanziari.

Dal punto di vista tecnico, per i prossimi sviluppi si intende migliorare ulteriormente il modello utilizzando nuove caratteristiche per tener conto della volatilità tipica delle criptovalute e esplorare le potenzialità delle tecniche di deep learning.

CAPITOLO 4

ANOMALIE COME RICCHEZZA SEMANTICA: TEXT SEGMENTATION NEL DOMINIO LEGALE

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. Si tratta di un

passo in avanti significativo, poiché consente una comprensione più profonda e strutturata del testo, facilitando ulteriormente il lavoro dei professionisti del diritto.

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. In definitiva, l'approccio proposto rappresenta una base solida per la costruzione di soluzioni di segmentazione efficienti e robuste, in grado di affrontare le sfide uniche poste dalla segmentazione del testo in documenti legali di varie lingue.

4.1 Stato dell'Arte

Il campo della segmentazione del testo è stato oggetto di numerosi studi che ne hanno esplorato le tecniche sia in un ambito generale sia specificatamente nel settore dei documenti legali. Tradizionalmente, i metodi

basati su regole sono stati largamente impiegati, grazie alla loro capacità di identificare modelli di scrittura specifici all'interno dei testi. Un esempio emblematico è il lavoro di Bali [44], che si è concentrato sulla segmentazione del testo a livello di frase per la lingua malese, elaborando il testo e costruendo regole adatte a tale scopo. In una direzione simile, Saniati [45] ha sviluppato un approccio di segmentazione basato sulla distribuzione delle entità nominate per identificare i cambiamenti di paragrafi nei testi. Sebbene i metodi basati su regole siano noti per la loro velocità e efficacia, essi richiedono un'attenta manutenzione, specialmente quando si trattano diversi tipi di documenti o quando le parole chiave che indicano l'inizio di un segmento sono imprecise o cambiano.

Di fronte a queste limitazioni, le tecniche statistiche e di deep learning hanno guadagnato terreno nel campo dell'elaborazione del linguaggio naturale. Tra gli approcci più innovativi troviamo l'uso di reti BiLSTM in vari contesti. Per esempio, Koshorek [46] han implementato un'architettura gerarchica che utilizza BiLSTMs sia per generare rappresentazioni delle frasi sia per la classificazione. Un'ulteriore evoluzione è stata proposta da Pinkesh [47], che han integrato le BiLSTMs con reti neurali convoluzionali (CNNs) e meccanismi di attenzione. Una ricerca di particolare rilievo ha coinvolto i Campi Casuali Condizionati [48] alla segmentazione automatica delle decisioni della Corte Ceca, suddividendole in parti multi-paragrafo [49]. Nonostante la scarsità di documenti disponibili e l'uso di un modello probabilistico, questa metodologia ha ottenuto un punteggio F1 medio di 0.703 su 7 sezioni.

Innovazioni ulteriori [50] hanno proposto tre diverse architetture per la segmentazione del testo basate su BERT, sottolineando l'importanza del contesto locale del documento. In una direzione simile, ma con un approccio diverso, Glavaš [51] ha sviluppato un modello consapevole della coerenza basato su due transformer gerarchici, dimostrando grande efficacia nel trasferimento zero-shot tra lingue diverse. Più di recente, Aumiller [52] ha introdotto un nuovo approccio per la segmentazione del testo, considerando il cambio di argomento tra le frasi come indicatore di un nuovo confine di

segmento, attraverso il metodo di Predizione dello Stesso Argomento (STP). Anche in questo caso, sono stati impiegati due modelli basati su BERT per STP e per l'analisi dell'intero documento.
L'impiego di modelli transformer, come BERT, ha segnato un importante progresso nel campo della segmentazione del testo, migliorando notevolmente l'accuratezza grazie alla loro capacità di catturare informazioni contestuali. Il nostro approccio, che si basa su queste innovazioni, introduce il rilevamento di break-point utilizzando un modello basato su transformer, pre-addestrato su testi legali italiani. Il documento procede poi con una dettagliata esposizione della nostra metodologia proposta, dei risultati sperimentali e della valutazione, mettendo in luce l'impatto potenziale del nostro approccio nella segmentazione dei documenti legali, un campo in costante evoluzione e di cruciale importanza nel settore della giurisprudenza.

4.2 Metodologia

In Italia, gli studi legali e i professionisti del settore giuridico possono ottenere significativi vantaggi dall'analisi approfondita dei documenti legali, in particolare quelli provenienti dalla Corte di Cassazione, la più alta corte d'appello del paese, situata all'apice della giurisdizione ordinaria. La Corte di Cassazione riveste un ruolo cruciale nel sistema giudiziario italiano, fungendo da ultimo grado di giudizio per le cause civili e penali e avendo la responsabilità di garantire l'uniformità dell'interpretazione della legge in tutto il territorio nazionale.
La Corte di Cassazione italiana, essendo la più alta corte di appello del paese, rappresenta una fonte inestimabile di dati legali. La ricerca condotta si focalizza sull'analisi e l'estrazione di informazioni da un data set composto da 730 documenti provenienti da questa corte, suddivisi in due gruppi: 520 documenti destinati alla fase di addestramento e 210 per la fase di test.
L'importanza di questi documenti risiede nella loro ricchezza informativa e nella loro rilevanza nel contesto giuridico italiano. La Corte di Cassazione offre

una prospettiva unica sulle interpretazioni e le applicazioni della legge italiana. I documenti selezionati forniscono non solo esempi di decisioni giuridiche, ma anche di ragionamenti e argomentazioni legali, rendendoli una fonte preziosa per l'apprendimento e l'analisi.

Per rendere questi documenti accessibili e utilizzabili, è stato intrapreso un processo dettagliato di organizzazione e preparazione dei dati. Ogni documento è stato scomposto in frasi singole, con l'obiettivo di rendere l'input del modello più gestibile e focalizzato. Questa suddivisione in frasi permette una migliore comprensione e analisi del testo, rendendo più semplice l'identificazione di specifiche informazioni legali e terminologie.

Il data set è notevole non solo per la sua dimensione, ma anche per il suo equilibrio. Il set di addestramento comprende circa 136.000 frasi, mentre il set di test è costituito da 60.000 frasi. Questa vasta raccolta di dati è stata bilanciata in modo da garantire che una varietà di argomenti e casi sia rappresentata equamente, fornendo così una visione completa e diversificata del panorama legale italiano.

Le fonti utilizzate per la raccolta dei documenti includono *ItalgiureWeb*[1], l'archivio del Centro di Documentazione Elettronica della Corte Suprema, che costituisce una risorsa fondamentale per l'accesso a documenti giuridici ufficiali. Vi è poi *Ricerca Giuridica*[2], un catalogo esaustivo delle sentenze della Corte Suprema, che fornisce un'ampia panoramica delle decisioni giuridiche prese nel corso degli anni. Infine, è stata utilizzata la fonte *La Legge per tutti*[3], che raccoglie migliaia di sentenze di cassazione dal 2012 al 2022, offrendo una visione attuale e dettagliata dell'evoluzione della giurisprudenza italiana.

Per quanto riguarda l'etichettatura dei testi, questa è stata eseguita manualmente con l'ausilio di un motore a regole ad hoc, creato specificatamente per questo compito. Questo strumento utilizza un insieme di regole basate su schemi, definite attraverso condizioni if-then, dove specifiche parole chiave o espressioni innescano la classificazione. Il motore a regole, sfruttando la libreria

[1] https://www.italgiure.giustizia.it/

[2] https://www.ricercagiuridica.com/

[3] https://www.laleggepertutti.it/

regex, è in grado di identificare pattern testuali ricorrenti all'interno dei testi legali, manipolando array e stringhe in modo efficiente e accurato.

I documenti sono stati organizzati in una struttura complessa, divisa in tre sezioni principali: la sezione "testo", che contiene il testo grezzo del documento legale; la sezione "linee", che include un dizionario in cui ogni elemento rappresenta il testo della i-esima riga del documento, il relativo allineamento (sinistra, destra, centro, giustificato) e l'etichetta corrispondente (1 se la riga rappresenta un break-point, 0 in caso contrario); e infine la sezione "sezioni", un dizionario che dettaglia l'indice di inizio e fine della j-esima sezione (o segmento) del documento e l'etichetta di sezione corrispondente.

L'analisi dettagliata dei testi delle sentenze ha permesso di identificare diverse sezioni fondamentali, ognuna con una funzione specifica all'interno del documento legale. Queste includono: l'intestazione, che fornisce dettagli sulla corte di riferimento per la sentenza (questa sezione non è considerata nella formazione e nel test del modello); l'elenco degli attori, che comprende i nomi dei ricorrenti nel processo; l'elenco dei convenuti, ovvero i soggetti contro cui è stata intentata la causa; la sezione dei fatti, che riassume le circostanze e i contesti in cui si è sviluppato il caso; la descrizione dei procedimenti, che dettaglia l'andamento effettivo del processo; le motivazioni, dove sono esposti i motivi che hanno portato all'esito della sentenza; le informazioni aggiuntive, che includono dati supplementari ritenuti rilevanti per la sentenza; l'esito del processo, che può comprendere verdetti, ordinanze, ingiunzioni o altre determinazioni legali fatte dalla corte; e infine la firma del giudice o del presidente del consiglio, che convalida ufficialmente il documento.

L'approccio adottato in questo studio, basato su un'analisi meticolosa e sistematica dei documenti legali, offre una panoramica completa e dettagliata delle diverse componenti che costituiscono una sentenza della Corte di Cassazione. Questa metodologia non solo migliora la comprensione dei documenti legali ma apre anche la strada a sviluppi futuri nell'ambito dell'elaborazione del linguaggio naturale applicata al settore giuridico, fornendo agli studiosi e ai professionisti del diritto strumenti sempre più

sofisticati per l'analisi e l'interpretazione dei testi legali.

Nel nostro studio, il data set presentava uno squilibrio significativo: solamente il 5.7% delle linee totali rappresentava i break-point, mentre la maggior parte delle linee costituiva i segmenti senza interruzioni. Questa disparità era particolarmente evidente considerando l'ampiezza del data set, che comprendeva più di 98.000 linee. Nello specifico, del set di addestramento di 68.524 linee, soltanto 4.015 erano classificate come break-point. Tale distribuzione sbilanciata poteva potenzialmente alterare il processo di apprendimento del modello, enfatizzando eccessivamente le linee senza interruzioni e compromettendo la capacità di rilevare efficacemente i break-point.

Nonostante questo squilibrio pronunciato, il modello di segmentazione del testo ha mostrato prestazioni eccezionali, raggiungendo risultati notevoli anche con una minima percentuale di break-point. Abbiamo tentato di contrastare lo squilibrio del data set impiegando diverse tecniche di ri-campionamento. Sorprendentemente, però, questi sforzi non hanno portato a miglioramenti significativi. Invece, è emersa la capacità intrinseca del modello di adattarsi allo squilibrio esistente, senza ricorrere a correzioni artificiali. Questo ha evidenziato l'efficacia dell'architettura del modello, la sua abilità nell'estrazione delle caratteristiche e la qualità dei dati utilizzati per l'addestramento.

L'approccio proposto per la segmentazione di documenti legali si avvale di un modello basato su Transformer, specificamente ottimizzato per adattarsi al nostro data set di documenti legali. Tale approccio rappresenta un salto qualitativo nel campo dell'elaborazione automatica dei testi nel settore giuridico, ponendosi come un punto di riferimento nello sviluppo di tecnologie avanzate per la comprensione del linguaggio legale.

Il punto di partenza della ricerca è stato il lavoro pionieristico di [53], che hanno sviluppato il primo modello basato su transformer per il dominio legale inglese. Questo modello innovativo, chiamato BERT [29], è stato addestrato da zero su corpora specifici del settore legale. In maniera analoga, il modello ITALIAN-LEGAL-BERT [54] è nato dall'ulteriore

addestramento del modello ITALIAN-XXL-BERT su corpora di diritto civile italiano. Questo addestramento supplementare ha incluso quattro epoche aggiuntive su un corpus di 3.7 GB di testi provenienti dall'Archivio Nazionale della Giurisprudenza, un vasto repository pubblico che raccoglie milioni di documenti legali, come decreti, ordinanze e sentenze civili, emessi dai tribunali italiani e dalla corte d'appello. È stato utilizzato ITALIAN-LEGAL-BERT in quanto offre vantaggi specifici e una rilevanza cruciale per il compito di segmentazione di documenti legali, grazie alla sua capacità di catturare le sfumature e le complessità del linguaggio giuridico.

La metodologia che si propone parte dal presupposto che all'interno dei testi legali è possibile identificare modelli e strutture ricorrenti, sui quali basare la segmentazione del testo. Questo approccio consente di sviluppare un sistema in grado di identificare innanzitutto paragrafi o sezioni caratteristici, delimitati da intestazioni specifiche o formule introduttive, per poi procedere alla segmentazione del testo. In questo contesto, si indica un "break-point" in un testo legale come una linea che corrisponde all'inizio di un nuovo segmento all'interno del documento. Poiché queste break-lines rappresentano le intestazioni o le formule introduttive delle sezioni, vengono incluse come prima linea del segmento. Invece, la linea che precede il break-point costituisce l'ultima linea del segmento precedente e viene inclusa come ultima linea della sezione precedente.

Consideriamo un data set $D = \{T_1, T_2, \ldots, T_K\}$ costituito da K documenti di testo, e definiamo ogni documento $T_i = \{s_1^i, s_1^i, \ldots, s_N^i\}$ come una sequenza di N sezioni. Ciascuna sezione $s_j = \{l_1^{i,j}, l_1^{i,j}, \ldots, l_M^{i,j}\}$ è composta da M linee (o frasi). Ogni riga è associata a un'etichetta di classificazione $y \in (0, 8)$, dove 0 indica che la frase non è un break-point e tutti gli altri numeri indicano che la riga è un break-point che determina l'inizio di una sezione (ogni numero rappresenta una sezione diversa).

Il metodo proposto si concentra, quindi, sull'identificazione delle frasi che definiscono i confini tra diverse sezioni all'interno di un documento legale, che chiamiamo "break-point". Quando un break-point viene rilevato,

l'etichetta associata fornisce informazioni sulla classe della sezione a cui appartiene. Rilevando efficacemente questi break-point, possiamo suddividere con precisione il documento in unità significative, migliorando sia il recupero sia l'analisi delle informazioni.

Questo approccio si distingue per la sua capacità di affrontare una delle principali sfide nell'elaborazione dei testi legali: la varietà e la complessità delle strutture testuali. I documenti legali, infatti, sono notoriamente complessi e articolati, presentando un'ampia gamma di stili, formati e terminologie. La nostra metodologia, grazie all'impiego di un modello transformer specificamente addestrato e ottimizzato, è in grado di catturare e analizzare queste caratteristiche distintive, fornendo così un'analisi accurata e dettagliata del contenuto dei documenti.

In sintesi, il metodo sviluppato propone un approccio innovativo e altamente specializzato per la segmentazione di documenti legali, che apre nuove prospettive nell'ambito dell'intelligenza artificiale applicata al settore giuridico. Questa metodologia non solo migliora la comprensione e l'analisi dei testi legali ma rappresenta anche un passo avanti significativo nella ricerca e nello sviluppo di soluzioni tecnologiche avanzate per il trattamento automatico dei documenti legali. Con questo approccio, si intende contribuire a definire nuovi standard nel campo dell'elaborazione del linguaggio naturale, offrendo strumenti sempre più sofisticati e accurati per la gestione e l'analisi dei testi nel dominio legale.

4.3 Risultati

In questa sezione, presentiamo i risultati del metodo proposto per la segmentazione dei documenti legali utilizzando il modello ITALIAN-LEGAL-BERT. Per valutarne l'efficacia, abbiamo utilizzato diverse metriche di valutazione, tra cui lo score F1 bilanciato, l'accuratezza e la finestra categorica di WindowDiff. Queste metriche ci consentono di ottenere una visione completa della qualità della segmentazione, tenendo conto delle sfide legate allo sbilanciamento delle etichette nei documenti legali.

Inoltre, sono stati condotti due confronti significativi. Il primo confronto è stato effettuato con il modello ITALIAN-XXL-BERT[4]. Questo confronto ha permesso di comprendere le differenze tra i due modelli e sottolineare l'importanza dell'addestramento specifico per il contesto legale. I risultati di questo confronto evidenziano l'efficacia del modello ITALIAN-LEGAL-BERT nelle sfide uniche presentate dai documenti legali.

Il secondo confronto è stato realizzato con un motore basato su regole. Questo motore si basa su pattern predefiniti, regole linguistiche e euristiche per identificare i punti di separazione nei documenti legali. Questo confronto ha permesso di mettere in luce sia i vantaggi sia i limiti di un approccio basato su regole rispetto a un modello di apprendimento automatico come quello utilizzato. Mentre il motore basato su regole può avere prestazioni decenti in alcune situazioni, il modello ITALIAN-LEGAL-BERT dimostra una maggiore flessibilità e adattabilità nei confronti di diverse tipologie di documenti legali e contesti.

In sintesi, i risultati dimostrano che ITALIAN-LEGAL-BERT offre una solida soluzione per la segmentazione di documenti legali in italiano, superando sia un modello generico come ITALIAN-XXL-BERT sia un motore basato su regole. Questo rappresenta un importante passo avanti nell'automatizzazione delle attività legate all'analisi dei documenti legali, migliorando l'efficienza e la precisione nelle operazioni giuridiche.

Nei primi anni dello sviluppo della Segmentazione del Testo, la metrica di valutazione P_k, introdotta nel paper intitolato *Statistical Models for Text Segmentation* [55], si affermò come uno standard per la valutazione degli algoritmi di segmentazione del testo. Tuttavia, una successiva analisi teorica [56] ha sollevato alcune preoccupazioni riguardo all'adeguatezza di questa metrica.

Le principali criticità individuate nella metrica P_k riguardano il suo squilibrio nella penalizzazione tra falsi negativi e falsi positivi. In particolare, questa metrica tende a punire in modo eccessivo le mancate identificazioni di confini

[4]https://huggingface.co/dbmdz/bert-base-italian-xxl-uncased

tra segmenti del testo, trascurando gli errori di sovrapposizione o vicinanza tra segmenti. Inoltre, è sensibile alle variazioni nella distribuzione delle dimensioni dei segmenti, il che potrebbe distorcere la valutazione.

Per affrontare queste criticità, è stata introdotta la metrica WindowDiff [56], che adotta un approccio differente. Questa metrica utilizza una finestra mobile di dimensioni fisse lungo il testo e penalizza l'algoritmo di segmentazione ogni volta che il numero di confini identificati all'interno della finestra non corrisponde al numero effettivo di confini presenti in quella porzione di testo. WindowDiff offre un modo più equilibrato per valutare le prestazioni dell'algoritmo, affrontando le preoccupazioni sollevate dalla metrica P_k.

Tuttavia, va notato che WindowDiff tiene conto solo della sequenza di segmentazione, senza considerare la classe o la tipologia di segmento identificato. In risposta a questa limitazione, nell'articolo viene proposta la metrica Categorical WindowDiff. Questa nuova metrica calcola i valori di WindowDiff in modo individuale per ciascun tipo di segmento o categoria, e successivamente aggrega tali valori senza richiedere operazioni aggiuntive complesse.

In sintesi, questo studio sottolinea le sfide legate alla valutazione degli algoritmi di segmentazione del testo, presentando una serie di metriche, tra cui P_k, WindowDiff e Categorical WindowDiff, che cercano di affrontare diverse criticità e offrire una valutazione più completa delle prestazioni degli algoritmi di segmentazione del testo nei documenti legali.

Al fine di preparare i dati testuali per l'analisi, è stata attuata una fase di pre-elaborazione del testo. L'obiettivo principale di questa pre-elaborazione è la pulizia delle frasi di input al fine di renderle adatte alla fase successiva di apprendimento. Durante questa fase, sono stati individuati e rimossi vari schemi presenti nel testo, utilizzando espressioni regolari. Applicando questo metodo personalizzato di pre-elaborazione a ciascuna frase, si ottiene una raffinazione dei dati testuali, il che contribuisce in maniera significativa al miglioramento dell'accuratezza delle previsioni.

È importante notare che i documenti legali seguono una struttura specifica,

influenzata dalla giurisdizione del paese a cui fanno riferimento. Questo modello di scrittura è caratterizzato da macro-sezioni, all'interno delle quali le frasi, gli acronimi o le parole iniziali sono allineati in modo distintivo, distinguendoli dal resto del testo presente nella stessa sezione. L'obiettivo di questa formattazione è attirare l'attenzione del lettore sul passaggio tra diverse aree o concetti all'interno del documento.

Per arricchire la comprensione del testo da parte del modello, è stata introdotta una caratteristica aggiuntiva relativa all'allineamento del testo all'interno del documento. Questa caratteristica è rappresentata da una stringa di testo di allineamento aggiunta all'inizio del testo di ciascuna linea, come ad esempio: "centro Questa è una riga del documento". È emerso dagli studi effettuati che l'inclusione di questa informazione sull'allineamento sia fondamentale per migliorare la capacità del modello di comprendere le informazioni spaziali presenti nel testo.

La Figura 4.1 fornisce ulteriori prove a sostegno delle ipotesi che sono state formulate, mostrando la distribuzione degli allineamenti sulle breaklines. Tale distribuzione è fortemente sbilanciata, con una predominanza dell'allineamento "centro". Tuttavia, le implicazioni e la validità di questo concetto verranno ulteriormente sviluppate e consolidate nelle sezioni successive dell'articolo.

In conclusione, questa fase di pre-elaborazione dei dati e l'aggiunta dell'informazione sull'allineamento migliorano notevolmente la capacità del modello di analizzare e interpretare i documenti legali, contribuendo a ottenere previsioni più accurate e informative.

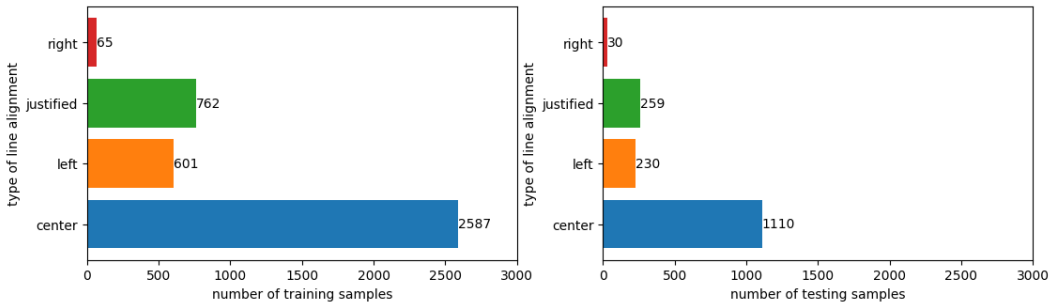


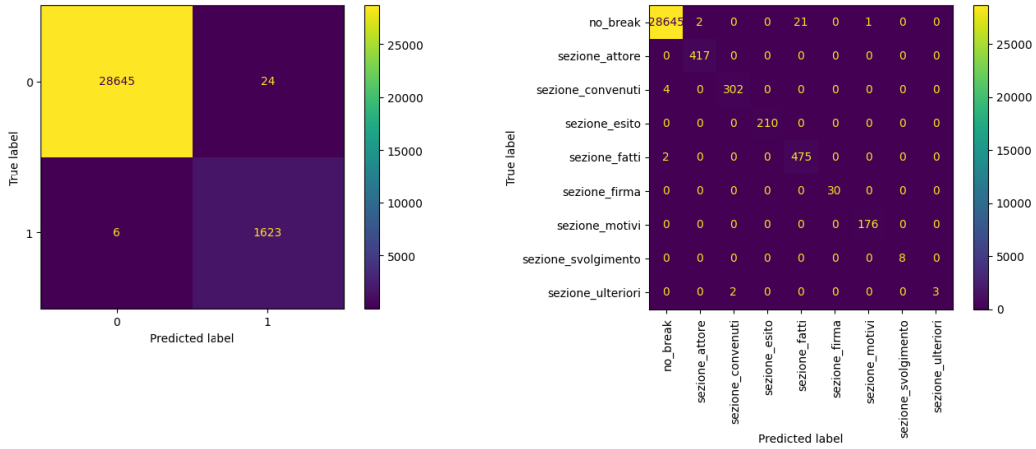
Figura 4.1: Distribuzione dell'allineamento del testo

Prima di esaminare i risultati circa la segmentazione del testo all'interno dei documenti, è stata inizialmente valutata l'accuratezza della stima dei break-point sulle linee dell'insieme di test. Per farlo, sono state utilizzate due matrici di confusione presentate nella Figura 4.2. La prima matrice riguarda la stima dei break-point, senza considerare la loro classe (Immagine a), mentre la seconda matrice tiene conto delle diverse classi nella stima dei break-point (Immagine b).

I risultati ottenuti da queste matrici di confusione indicano prestazioni notevoli del modello. Nel caso della classificazione binaria dei break-point (ovvero prevedere se una data linea è un break-point o meno), il modello ha raggiunto un eccezionale punteggio F1 di 0,98. Questo risultato dimostra l'efficacia del modello nella previsione dei break-point all'interno dei documenti legali.

È importante sottolineare che sono stati condotti test supplementari utilizzando una fase di pre-elaborazione dei dati che escludeva l'allineamento del testo. In queste circostanze, il modello non ha ottenuto risultati soddisfacenti, con un punteggio F1 di circa 0,86. Ciò evidenzia l'importanza dell'inclusione dell'allineamento del testo nella fase di pre-elaborazione, che contribuisce in modo significativo all'accuratezza complessiva del modello nelle previsioni di segmenti di testo.

In conclusione, prima di analizzare i risultati della previsione dei segmenti di testo, è stato valutato con successo la capacità del modello di prevedere i break-point sulle linee dei documenti legali, ottenendo prestazioni notevoli. Questo risultato sottolinea l'efficacia dell'approccio proposto e l'importanza di considerare l'allineamento del testo nella fase di pre-elaborazione dei dati.



(a) Confusion Matrix per break-point prediction.

(b) Confusion Matrix per break-point class prediction.

Figura 4.2: Confusion Matrix per Breakpoint Prediction.

In seguito, sono stati sottoposti i modelli al vero task di segmentazione del testo, che tiene conto sia della stima delle break-lines sia della stima delle classi (o sezioni) a cui appartengono tali break-point. I risultati, presentati nella Tabella 4.1, includono i punteggi F1 per la classificazione binaria e multi classe, insieme alle metriche Categorical WindowDiff.

Model	**F1 Binary**	**F1 Multi - Macro avg**	**Categorical WindowDiff**
ITALIAN-XXL-BERT	0.90	0.78	6.833
ITALIAN-LEGAL-BERT	0.98	0.96	0.324
Rule-Engine	0.99	0.99	0.002

Tabella 4.1: Risultati Text Segmentation

La metrica Categorical WindowDiff fornisce una misura della differenza media tra le classificazioni previste e quelle effettive delle sezioni nei documenti, offrendo ulteriori dettagli sulle prestazioni dei modelli. Un valore più basso di Categorical WindowDiff indica una migliore corrispondenza tra le previsioni del modello e le classificazioni reali delle sezioni.

I risultati ottenuti mettono in luce una notevole differenza nelle prestazioni tra i modelli ITALIAN-XXL-BERT e ITALIAN-LEGAL-BERT per il compito di segmentazione del testo. Ciò può essere spiegato da diverse ragioni. In primo

luogo, il modello ITALIAN-LEGAL-BERT è stato appositamente addestrato su testi legali italiani, il che lo rende più adatto a catturare i modelli linguistici unici, i termini tecnici e le caratteristiche strutturali specifiche dei documenti legali. Questo addestramento specializzato consente al modello di comprendere meglio e interpretare le complessità dei testi legali, migliorando la sua capacità di prevedere con precisione i break-point e classificare le sezioni in modo accurato.

D'altra parte, il modello ITALIAN-XXL-BERT, sebbene sia un potente modello linguistico, non dispone della stessa conoscenza specifica di dominio e mostra una capacità limitata nel catturare le sfumature e le complessità dei testi legali.

Per quanto riguarda il motore basato su regole, è importante sottolineare che esso si basa su schemi specifici e regole linguistiche definite per il data set in esame, il che spiega la sua superiorità nei risultati ottenuti. Tuttavia, è importante notare che i sistemi basati su regole possono essere meno adattabili e flessibili rispetto ai modelli di apprendimento automatico, specialmente quando si tratta di dati complessi e variegati.

In definitiva, i risultati riflettono l'importanza dell'addestramento specifico per il dominio nei modelli di elaborazione del linguaggio naturale e sottolineano le sfide e le limitazioni dei sistemi basati su regole nei contesti reali.

4.4 Conclusioni e Sviluppi Futuri

In conclusione, in questo studio [57], è stato presentato un approccio altamente efficace per affrontare una sfida critica nel campo dei documenti legali: la segmentazione del testo. Questa sfida è fondamentale perché la corretta segmentazione del testo in documenti giuridici può avere un impatto significativo sull'interpretazione e sull'analisi dei contenuti legali. L'approccio si basa su tecnologie avanzate, dimostrando prestazioni eccezionali e apportando un contributo innovativo alla comunità accademica e professionale. La metodologia sviluppata si basa su un modello basato su Transformer, noto

per la sua capacità di catturare relazioni complesse nel testo. Utilizzando la rilevazione dei break-point, il nostro modello è in grado di identificare in modo preciso e coerente i confini tra le diverse sezioni di un documento legale. Inoltre, viene enfatizzata l'importanza della pre-elaborazione del testo, introducendo una caratteristica "spaziale" che tiene conto dell'allineamento delle righe all'interno del documento. Questo miglioramento nella segmentazione può aumentare notevolmente l'efficienza delle operazioni di ricerca e analisi giuridica.

Per ulteriori sviluppi, si intende esplorare l'utilizzo di modelli multilingua per testare la robustezza del metodo su diverse lingue. Inoltre, considerando la rapida evoluzione dei Large Language Models, sarà sicuramente valutata l'opportunità di utilizzare modelli più sofisticati per migliorare ulteriormente la segmentazione del testo nei documenti legali.

Questo lavoro rappresenta un passo significativo nell'avanzamento della comprensione e dell'analisi dei documenti legali, con possibili applicazioni importanti nel campo giuridico, nella ricerca e nell'automazione dei processi legali.

CAPITOLO 5

ANOMALIE COME ERRORI: I 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.
La tecnologia OCR si avvale di diversi campi dell'informatica:

- **Elaborazione delle Immagini**: Per migliorare la qualità delle immagini e rendere i caratteri più distinguibili.
- **Riconoscimento dei Pattern e Machine Learning**: Per identificare e classificare i caratteri in base a grandi database di esempi.
- **Reti Neurali e Intelligenza Artificiale**: Per migliorare l'accuratezza del riconoscimento in situazioni complesse, come la scrittura a mano o font insoliti.

Nonostante i notevoli progressi, l'OCR deve ancora affrontare sfide significative, soprattutto nel dominio legale, come:

- **Riconoscimento di testi su sfondi complessi**: I documenti legali possono contenere sfondi complessi o essere stampati su carta con filigrane di sicurezza. Questo può rendere difficile per l'OCR riconoscere correttamente il testo.
- **Gestione di formati di testo non standard**: I documenti giuridici spesso utilizzano un formato di testo che può variare notevolmente, con diverse dimensioni e stili di carattere, colonne, riquadri, note a piè di pagina, ecc. Ciò richiede agli algoritmi OCR di essere estremamente flessibili e sofisticati.
- **Interpretazione della scrittura manuale**: La scrittura manuale è ancora comune in molti documenti legali, come note a margine, firme e annotazioni. L'interpretazione accurata della scrittura manuale è una delle sfide più difficili per l'OCR, a causa della varietà e dell'irregolarità dei caratteri.

- **Precisione e affidabilità**: Nel campo giuridico, è fondamentale che il riconoscimento del testo sia estremamente preciso, poiché anche un piccolo errore può avere conseguenze significative. Questo pone l'accento sulla necessità di sistemi OCR altamente affidabili e precisi.

- **Privacy e sicurezza dei dati**: La manipolazione di documenti legali richiede un alto livello di sicurezza e privacy, poiché spesso contengono informazioni sensibili. Gli algoritmi OCR e i sistemi di gestione documentale devono quindi essere progettati con rigide misure di sicurezza.

- **Integrazione con sistemi legali esistenti**: Integrare la tecnologia OCR in sistemi legali esistenti può essere una sfida, in quanto richiede non solo la compatibilità tecnologica, ma anche l'adattamento ai processi e alle normative del settore giuridico.

Nonostante queste sfide, l'OCR sta diventando sempre più sofisticato e continua a offrire grandi benefici nel campo giuridico, migliorando l'efficienza e l'accessibilità dei documenti legali.

5.1 Stato dell'Arte

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.

- **Rilevamento degli Errori**: Questo processo 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.

- **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.

Nel corso degli anni, sono stati sviluppati vari approcci per il post-processing OCR. Tuttavia, molti di questi si concentrano principalmente sulla correzione degli errori, spesso trascurando l'importanza della rilevazione degli errori. Un'efficace rilevazione degli errori è fondamentale; senza la precisa identificazione di dove e quali sono gli errori, la loro correzione diventa un compito arduo.
Gli errori in un testo OCR possono essere generalmente classificati in due categorie principali [58]:

- **Non-Word Error Detection**: Questo tipo identifica errori che risultano in stringhe di caratteri che non formano parole riconoscibili. Esempi includono errori di battitura o caratteri mal interpretati dall'OCR.

- **Real-Word Error Detection**: A differenza del primo tipo, qui gli errori si verificano in parole che esistono ma sono usate in modo inappropriato nel contesto. Questi errori sono più difficili da rilevare poiché richiedono una comprensione del contesto del testo.

Gli approcci al Non-Word Error Detection nel campo dell'OCR sono principalmente metodi basati su dizionari. Questi approcci si sono rivelati efficaci in diversi compiti del NLP, come sottolineato da studi quali quelli di Hull [59] e Sindhu [60].
L'utilizzo di dizionari e ampie liste di parole, compilati da corpora e altre fonti, permette di confrontare ogni parola del testo con quelle presenti nel dizionario, identificando così possibili errori.

Tale metodo è stato applicato anche nel rilevamento di errori di Optical Character Recognition, come dimostrato da [61] e Nguyen [62].
Questi metodi includono il Character n-gram analysis e le Lexicon look-up techniques, ognuno con i propri vantaggi e limitazioni:

1. **Character n-gram analysis**: Questa tecnica impiega tabelle di n-grammi di caratteri per identificare errori di ortografia.
   Gli n-grammi sono sequenze di caratteri (ad esempio, trigrammi sono sequenze di tre caratteri) utilizzati per analizzare la probabilità di sequenze di lettere in una lingua.
   Questo metodo è particolarmente utile per rilevare errori in parole che non seguono modelli comuni della lingua.
   Esistono diverse varianti di tabelle di n-grammi, da quelle non-posizionali a quelle posizionali, e da tabelle binarie a quelle di frequenza.
   Tali varianti offrono diversi livelli di efficacia nel rilevamento degli errori.
   Zamora [63] evidenziato che le tabelle trigrammi di frequenza possono non essere efficaci nel distinguere tra parole valide e non valide.
   Morris [64] ha osservato che gli errori di ortografia spesso contengono trigrammi unici per il documento specifico, suggerendo l'uso di una formula per calcolare l'indice di peculiarità di ciascun token.

2. **Lexicon look-up techniques**: Queste tecniche confrontano ogni token del testo di input con le voci di un lessico. Se un token non corrisponde a nessuna voce, viene considerato un errore.
   Questo approccio è diretto e può essere efficace per identificare parole completamente fuori luogo. La scelta del lessico è cruciale per l'efficacia di questo metodo. Un lessico troppo limitato può portare al rifiuto di parole valide (falsi negativi), mentre un lessico eccessivamente ampio può portare all'accettazione di errori (falsi positivi).

Entrambe le tecniche di Non-Word Error Detection hanno i loro punti di forza e di debolezza. L'analisi degli n-grammi di caratteri è utile per identificare errori ortografici basati sulla frequenza e sulla struttura delle parole, ma può

non essere efficace nel contesto di parole peculiari a un documento.
D'altra parte, la ricerca nel lessico è efficace per identificare parole che non esistono affatto, ma richiede un lessico ben calibrato e può lottare con errori di contesto.
Il Real-Word Error Detection supera i problemi del Non-Word Error Detection considerando il contesto delle parole, ma la rilevazione di errori derivanti dai sistemi OCR rimane una sfida complessa, con vari approcci che offrono diversi vantaggi e limitazioni.
L'ingegnerizzazione delle caratteristiche testuali rimane un collo di bottiglia sia per l'apprendimento automatico che per altri approcci statistici.
La continua esplorazione e sviluppo di nuove tecniche e miglioramenti degli approcci esistenti sono cruciali per il progresso in questo campo.
Inoltre, lo studio di metodi più avanzati e l'integrazione di tecniche innovative potrebbero portare a soluzioni più efficaci e efficienti nel futuro della rilevazione di errori OCR e nell'elaborazione del linguaggio naturale in generale.

5.2 Metodologia

L'elaborazione di un data set strutturato per affrontare le sfide poste dai sistemi di OCR Anomaly Detection richiede un approccio meticoloso e innovativo. Il punto di partenza di questo processo è la creazione di una replica esatta del data set prima presentato nello studio di Legal Text Segmentation. Questa fase è fondamentale per garantire che tutte le frasi siano oggetto di una perturbazione accurata, che riflette le anomalie tipicamente riscontrate durante l'analisi dei testi tramite OCR.
Questo studio vuole intercettare gli errori che possono scaturire da un processo di OCR, sviluppando un modello di OCR Anomaly Detection.
Per fare ciò è stato necessario costruire un data set che contenesse errori tipici di un sistema OCR.
Per fare ciò è stata utilizzata la tecnica di "Data Augmentation", utilizzata nell'ambito dell'apprendimento automatico, specialmente nel deep learning,

per aumentare la quantità e la diversità dei dati di addestramento.
Questa pratica è particolarmente importante quando si dispone di un data set limitato, poiché l'aumento dei dati può aiutare a prevenire l'overfitting e migliorare la capacità del modello di generalizzare su nuovi dati.
Nel contesto della visione artificiale, la data augmentation può includere una varietà di trasformazioni applicate alle immagini, come rotazione, ridimensionamento, ritaglio, capovolgimento (orizzontale o verticale), variazioni di colore, aggiunta di rumore, e così via.
Queste trasformazioni creano versioni leggermente differenti della stessa immagine, il che aiuta il modello ad apprendere caratteristiche rilevanti indipendentemente da piccole variazioni nell'input.
Nel trattamento del linguaggio naturale, la data augmentation può includere tecniche come la parafrasi, l'alterazione della struttura grammaticale, l'inserimento o la rimozione di parole, e l'uso di sinonimi.
Tale approccio aiuta i modelli linguistici a essere meno sensibili alle specifiche formulazioni di una frase e a concentrarsi di più sul significato generale.
In sintesi, la data augmentation è un approccio fondamentale per migliorare la robustezza e l'efficacia dei modelli di machine learning e deep learning, specialmente quando i dati disponibili per l'addestramento sono limitati o poco vari. La data augmentation nel contesto testuale si riferisce a tecniche utilizzate per aumentare artificialmente la quantità di dati testuali disponibili, generalmente con lo scopo di migliorare la performance di modelli di machine learning e intelligenza artificiale.
Tutto ciò è particolarmente utile in scenari dove i dati disponibili sono limitati.
Alcune delle tecniche comuni di data augmentation per i dati testuali sono:

1. **Paraphrasing**: Riformulare lo stesso contenuto in modi diversi. Ad esempio, utilizzando strumenti come modelli di linguaggio per generare diverse versioni dello stesso testo.

2. **Back Translation**: Tradurre un testo in una lingua e poi ritradurlo indietro nella lingua originale. Questo può portare a piccole variazioni

nel testo che sono utili per la data augmentation.

3. **Perturbation**: Introdurre errori ortografici o grammaticali intenzionali per rendere il modello più robusto a tali variazioni.

4. **Synonym Replacement**: Sostituire le parole nel testo con i loro sinonimi per creare una nuova versione del testo.

5. **Random Insertion**: Inserire parole casuali nel testo per aumentare la diversità dei dati.

6. **Random Deletion**: Eliminare casualmente parole dal testo per simulare dati incompleti.

7. **Sentence Shuffling**: Cambiare l'ordine delle frasi in un paragrafo per creare nuove varianti del testo mantenendo lo stesso contesto generale.

Queste tecniche possono essere utilizzate singolarmente o in combinazione per generare un set di dati più grande e variegato, che può aiutare a migliorare la generalizzazione e la robustezza dei modelli di apprendimento automatico.
Nel caso specifico di questo studio è stata utilizzata la tecnica della perturbazione. La perturbazione delle frasi è un processo critico che mira a simulare le condizioni reali nelle quali i professionisti del settore legale lavorano con documenti testuali digitalizzati.
In molti casi, questi documenti vengono elaborati utilizzando sistemi OCR. Pertanto, è essenziale che il data set rifletta accuratamente le sfide associate a questa tecnologia. La perturbazione introduce variazioni significative, quali anomalie nel riconoscimento dei caratteri, discrepanze di formattazione e interpretazioni occasionali errate. Tali variazioni sono cruciali per mimare le difficoltà incontrate nelle applicazioni OCR.
Un aspetto chiave di questo processo è la limitazione del grado di perturbazione applicato a ciascuna frase. È stato deciso di perturbare solo il 10% dei caratteri per ogni frase, seguendo le indicazioni della ricerca condotta da Ma [65].
Tale scelta è dettata dalla necessità di creare un modello in grado di rilevare gradi lievi di perturbazione, con l'obiettivo finale di rendere il modello capace

di individuare tassi più elevati di disturbo.
Questa strategia mira a ottimizzare l'efficacia del modello nell'ambiente di lavoro reale, dove i professionisti legali si trovano frequentemente a dover gestire documenti con vari livelli di qualità testuale.
Il dataset risultante è una collezione di linee di testo, ciascuna delle quali è etichettata in modo distinto in base alla presenza o assenza di perturbazione. Le linee soggette a perturbazione ricevono l'etichetta "1", mentre quelle non perturbate sono contrassegnate con "0".
La classificazione binaria è essenziale per l'addestramento del modello, poiché consente di distinguere chiaramente tra testi originali e quelli alterati.
Inoltre, facilita l'analisi delle performance del modello in scenari diversi, valutando la sua capacità di riconoscere e gestire testi con varie tipologie di anomalie. Una volta ottenuto il data set finale è stato realizzato un classificatore 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.
Anche in questo caso, come per la Legal Text Segmentation, è stato utilizzato ITALIAN-LEGAL-BERT in quanto offre vantaggi specifici grazie alla sua capacità di catturare le sfumature e le complessità del linguaggio giuridico.

5.3 Risultati

Consideriamo $D = \{l_1, l_2, \ldots, l_K\}$ come l'insieme dei dati che comprende K frasi. A ciascuna frase è assegnata un'etichetta di classificazione y, che può essere 0 o 1, indicando rispettivamente frasi non perturbate e frasi perturbate. Su tale data set è stato addestrato il modello OCR Anomaly Detection per determinare se una frase è stata perturbata o meno. In particolare, dopo otto epoche di fine-tuning, il modello ha raggiunto uno score F1 del 97.5% sui dati

di test. Questo risultato, come mostrato nella Tabella 5.1, è significativo poiché evidenzia non solo l'alta precisione del modello nella classificazione delle frasi, ma anche la sua robustezza.

Modello	Precision	Recall	F1
ITALIAN-LEGAL-BERT	0.985	0.965	0.975

Tabella 5.1: Risultati OCR Anomaly Detection

Inoltre, dalla Figura 5.1 emerge la capacità del modello di ridurre al minimo sia i falsi positivi che i falsi negativi rendendolo particolarmente affidabile per l'identificazione di anomalie in contesti di OCR Anomaly Detection. La sua efficacia nel rilevare irregolarità nelle frasi è quindi un punto di forza chiave nella nostra ricerca.

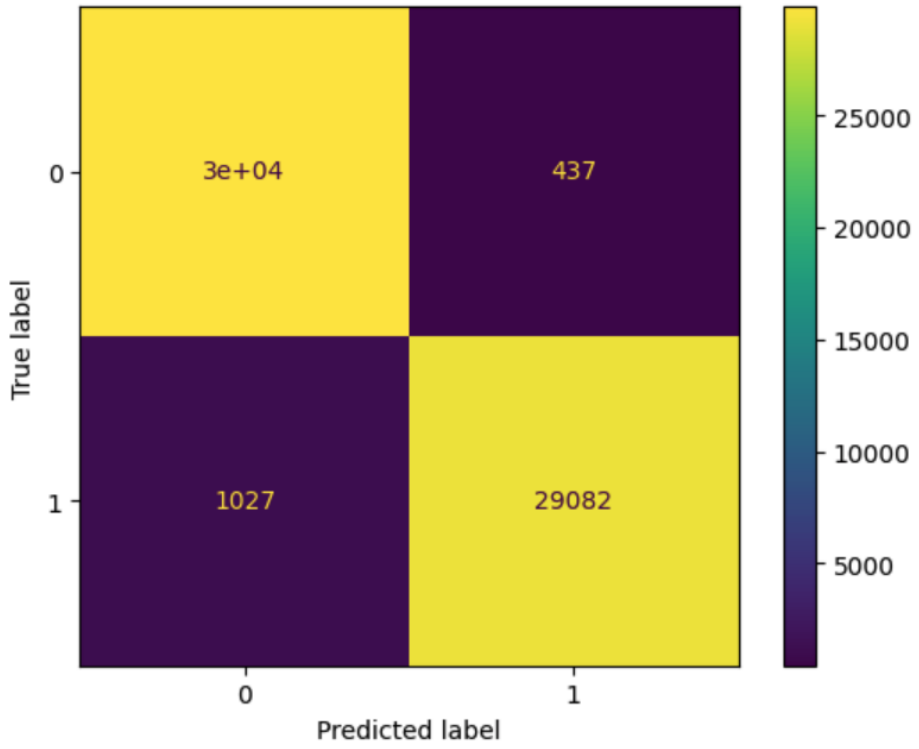


Figura 5.1: Confusion Matrix per OCR Anomaly Detection

5.4 Conclusioni e Sviluppi Futuri

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) [66], 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.

Il metodo REMOAC è basato su un processo complesso composto da più fasi. Innanzitutto, si verifica se il testo contiene errori attraverso il modello di OCR Anomaly Detection sopra descritto; se ci sono, vengono identificati i Most Important Error Token più significativi (MIET), che sono gli errori che hanno il maggiore impatto sulla qualità del testo riconosciuto. Questi token vengono poi confrontati con un insieme di parole chiave, denominato Bag of Most Important Words (BOMIW), che viene derivato dai dati di addestramento del classificatore di testi legali. Questo processo consente di individuare con precisione gli errori più critici e di apportare le correzioni necessarie.

La ricerca ha anche esplorato la possibilità di estendere l'applicazione di REMOAC oltre il dominio legale. Considerando l'ampia varietà di domini che presentano terminologie, gergo e strutture documentali uniche, come il settore medico, tecnico o scientifico, l'adattabilità di REMOAC a questi contesti risulta particolarmente promettente. Ciò implicherebbe non solo l'adattamento del metodo alle specificità linguistiche e terminologiche di ogni ambito, ma anche la sua validazione in contesti diversi.

Un'altra area di indagine futura riguarda la valutazione del metodo in diversi contesti linguistici. Attualmente, la ricerca si è concentrata sulla lingua italiana; tuttavia, per comprendere appieno il potenziale del metodo, è essenziale estendere la sua applicazione e valutazione ad altre lingue. Questo approccio non solo confermerebbe la versatilità del metodo come strumento

indipendente dalla lingua, ma fornirebbe anche dati preziosi sulla sua efficacia in diversi contesti linguistici e culturali.

L'aspetto innovativo di REMOAC risiede nella sua capacità di combinare tecniche avanzate di apprendimento automatico e analisi linguistica per affrontare un problema specifico e complesso come gli errori di stistemi OCR nel testo legale. La metodologia proposta offre un bilanciamento tra accuratezza e spiegabilità, due aspetti fondamentali nel contesto legale, dove le decisioni e le correzioni devono essere trasparenti e giustificabili.

In conclusione, la ricerca effettuata rappresenta un passo importante verso la comprensione e il superamento delle sfide poste dagli errori di sistemi OCR nel dominio legale. Attraverso l'implementazione di REMOAC, si mira a migliorare l'accuratezza e l'affidabilità dei sistemi di riconoscimento testuale in questo settore, fornendo così un contributo significativo non solo alla comunità legale, ma anche a un'ampia gamma di altri settori che si affidano alla tecnologia OCR. Con ulteriori ricerche e sviluppi, si prevede che il metodo proposto possa essere adattato e applicato con successo a una varietà di contesti linguistici e professionali, migliorando così l'efficacia complessiva di questi sistemi critici.

CAPITOLO 6

CONCLUSIONI

Nel mio percorso di dottorato, ho esplorato in profondità il concetto di Anomaly Detection, applicato in tre distinti ambiti: il rilevamento di schemi Pump and Dump nel mercato delle criptovalute, la segmentazione del testo in documenti legali, e l'ottimizzazione dei sistemi di Optical Character Recognition nel contesto legale attraverso il rilevamento degli errori che ne scaturiscono. Queste ricerche hanno offerto una visione unica delle diverse forme che le anomalie possono assumere e delle metodologie per rilevarle e affrontarle efficacemente.

Nel primo studio, è stato indagato il fenomeno del Pump and Dump, emergente nel mercato delle criptovalute. Questa frode, esemplificativa delle irregolarità del mercato, consiste nella manipolazione artificiale del valore di una criptovaluta per trarne profitto a discapito di investitori ignari. Abbiamo identificato il Natural Language Processing (NLP) e l'analisi dei social media come strumenti fondamentali per rilevare questi schemi fraudolenti. L'analisi dettagliata del linguaggio utilizzato nei canali online, come chat e social media, ha permesso di individuare segnali di manipolazione, come l'uso di espressioni iperboliche o la diffusione di informazioni contraddittorie. La nostra ricerca ha dimostrato come il NLP possa efficacemente identificare pattern di

linguaggio e comportamenti sospetti, fondamentali per la rilevazione precoce e la prevenzione delle truffe Pump and Dump.

Nel secondo ambito di ricerca, è stato esplorata la segmentazione del testo nei documenti legali. Questi documenti, caratterizzati da una ricchezza semantica e da una complessità strutturale, presentano sfide uniche in termini di navigabilità e analisi. Abbiamo sviluppato un metodo basato sul rilevamento di "break-points" che va oltre la semplice segmentazione del testo. Questo approccio innovativo non solo identifica i segmenti all'interno di un documento legale, ma classifica anche le sezioni in base al loro contenuto. Il modello, basato su una rete neurale di tipo transformer pre-addestrata su testi legali, ha mostrato notevole efficacia nel migliorare l'accessibilità e la gestione dei documenti legali, offrendo ai professionisti del diritto uno strumento potente per una più rapida identificazione e analisi delle sezioni rilevanti.

Nell'ultima parte della mia ricerca, sono state affrontate le sfide poste dall'implementazione di modelli per il rilevamento degli errori provenienti dai sistemi OCR in ambito legale. Nonostante i significativi progressi tecnologici, l'OCR deve confrontarsi con difficoltà come il riconoscimento di testi su sfondi complessi, la gestione di formati di testo non standard, e l'interpretazione della scrittura manuale, soprattutto nei documenti legali. La nostra ricerca ha evidenziato l'importanza di sviluppare algoritmi più sofisticati e flessibili, in grado di affrontare queste sfide e garantire un'alta precisione nel riconoscimento degli errori prodotti. Inoltre, abbiamo sottolineato l'importanza di considerare la privacy e la sicurezza dei dati, essenziali nella manipolazione di documenti legali contenenti informazioni sensibili.

Le ricerche condotte offrono contributi significativi in ciascuno dei tre ambiti studiati. Nel campo delle criptovalute, l'approccio basato sul NLP per il rilevamento dei Pump and Dump rappresenta un passo avanti nella protezione degli investitori e nella preservazione dell'integrità del mercato delle criptovalute. Nel contesto legale, la segmentazione avanzata del testo e il rilevamento degli errori degli OCR non solo migliorano l'efficienza e l'accuratezza del lavoro dei professionisti del diritto, ma rendono anche i

documenti legali più accessibili e gestibili. Questi progressi sottolineano l'importanza dell'integrazione delle tecnologie avanzate come l'intelligenza artificiale e il machine learning nelle pratiche quotidiane.

Inoltre, le metodologie sviluppate hanno il potenziale di essere adattate ed estese a una varietà di contesti e lingue, aprendo la strada a future ricerche e applicazioni. L'adattabilità e la versatilità di queste tecniche potrebbero rivoluzionare il modo in cui affrontiamo e gestiamo le anomalie in diversi settori.

In conclusione, questo percorso di ricerca ha dimostrato che l'impiego di tecniche avanzate di Machine Learning e soprattutto di Natural Language Processing offrono nuove prospettive nell'anomaly detection evidenziando il vasto potenziale e la necessità di ulteriori sviluppi e ricerche andando a comparare i risultati con pipeline più complesse che utilizzano anche Large Language Models.

BIBLIOGRAFIA

[1] V. Kumar. «Parallel and distributed computing for cybersecurity». In: *IEEE Distributed Systems Online* 6.10 (2005). DOI: 10.1109/MDSO.2005.53.

[2] Clay Spence, Lucas Parra e Paul Sajda. «Detection, Synthesis and Compression in Mammographic Image Analysis with a Hierarchical Image Probability Model». In: *Proceedings of the Workshop on Mathematical Methods in Biomedical Image Analysis* (ott. 2001). DOI: 10.1109/MMBIA.2001.991693.

[3] E. Aleskerov, B. Freisleben e B. Rao. «CARDWATCH: a neural network based database mining system for credit card fraud detection». In: *Proceedings of the IEEE/IAFE 1997 Computational Intelligence for Financial Engineering (CIFEr)*. 1997, pp. 220–226. DOI: 10.1109/CIFER.1997.618940.

[4] Ryohei Fujimaki, Takehisa Yairi e Kazuo Machida. «An approach to spacecraft anomaly detection problem using Kernel Feature Space». In: ago. 2005, pp. 401–410. DOI: 10.1145/1081870.1081917.

[5] Francis Ysidro Edgeworth. «XLI. On discordant observations». In: *Philosophical Magazine Series 1* 23 (1887), pp. 364–375. URL: https://api.semanticscholar.org/CorpusID:120568135.

[6] Hwee San. Teng, K. Chen e S. C. Lu. «Adaptive real-time anomaly detection using inductively generated sequential patterns». In: *Proceedings. 1990 IEEE Computer Society Symposium on Research in Security and Privacy* (1990), pp. 278–284. URL: https://api.semanticscholar.org/CorpusID:35632142.

[7] Allan Seheult et al. «Robust Regression and Outlier Detection.» In: *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 152 (gen. 1989), p. 133. DOI: 10.2307/2982847.

[8] Peter J. Huber. «Robust Statistics». In: *International Encyclopedia of Statistical Science*. A cura di Miodrag Lovric. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1248–1251. ISBN: 978-3-642-04898-2. DOI: 10.1007/978-3-642-04898-2_594. URL: https://doi.org/10.1007/978-3-642-04898-2_594.

[9] Markos Markou e Manjeet Singh. «Novelty detection: A review—Part 1: Statistical Approaches». In: *Signal Processing* 83 (dic. 2003), pp. 2481–2497. DOI: 10.1016/j.sigpro.2003.07.018.

[10] Markos Markou e Sameer Singh. «Novelty detection: A review - Part 2:: Neural network based approaches». In: *Signal Processing* 83 (dic. 2003), pp. 2499–2521. DOI: 10.1016/j.sigpro.2003.07.019.

[11] Rob Saunders e John Gero. «The Importance Of Being Emergent». In: (mar. 2000).

[12] Pang-Ning Tan, Michael Steinbach e Vipin Kumar. *Introduction to Data Mining, (First Edition)*. USA: Addison-Wesley Longman Publishing Co., Inc., 2005. ISBN: 0321321367.

[13] Xiuyao Song et al. «Conditional Anomaly Detection». In: *IEEE Transactions on Knowledge and Data Engineering* 19.5 (2007), pp. 631–645. DOI: 10.1109/TKDE.2007.1009.

[14] Mahesh Joshi, Ramesh Agarwal e Vipin Kumar. «Predicting rare classes: can boosting make any weak learner strong?» In: *Proc. 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ACM* (gen. 2002), pp. 297–306. DOI: 10.1145/775047.775092.

[15] Nitesh Chawla, Nathalie Japkowicz e Aleksander Kołcz. «Editorial: Special Issue on Learning from Imbalanced Data Sets». In: *SIGKDD Explorations* 6 (giu. 2004), pp. 1–6. DOI: 10.1145/1007730.1007733.

[16] Clifton Phua, Damminda Alahakoon e Vincent Lee. «Minority Report in Fraud Detection: Classification of Skewed Data». In: *SIGKDD Explorations* 6 (gen. 2004), pp. 50–59.

[17] Gary Weiss e Haym Hirsh. «Learning to Predict Rare Events in Event Sequences». In: (gen. 1998).

[18] Ricardo Vilalta e Sheng Ma. «Predicting rare events in temporal domain». In: feb. 2002, pp. 474–481. ISBN: 0-7695-1754-4. DOI: 10.1109/ICDM.2002.1183991.

[19] James Theiler e Michael Cai. «Resampling Approach for Anomaly Detection in Multispectral Images». In: *Proc SPIE* 5093 (apr. 2003). DOI: 10.1117/12.487069.

[20] Naoki Abe, Bianca Zadrozny e John Langford. «Outlier detection by active learning». In: vol. 2006. Ago. 2006, pp. 504–509. DOI: 10.1145/1150402.1150459.

[21] Ingo Steinwart, Don R. Hush e Clint Scovel. «A Classification Framework for Anomaly Detection». In: *J. Mach. Learn. Res.* 6 (2005), pp. 211–232. URL: https://api.semanticscholar.org/CorpusID:17427420.

[22] Dipankar Dasgupta e Luis Fernando Nino. «A comparison of negative and positive selection algorithms in novel pattern detection». In: vol. 1. Feb. 2000, 125–130 vol.1. ISBN: 0-7803-6583-6. DOI: 10.1109/ICSMC.2000.884976.

[23] Dipankar Dasgupta e Nivedita Majumdar. «Anomaly detection in multidimensional data using negative selection algorithm». In: vol. 2. Feb. 2002, pp. 1039–1044. ISBN: 0-7803-7282-4. DOI: 10.1109/CEC.2002.1004386.

[24] Patrik D'haeseleer, Stephanie Forrest e Paul Helman. «An immunological approach to change detection: algorithms, analysis and implications». In: mag. 1996, pp. 110–119.

[25] Claudio De Stefano, Carlo Sansone e Mario Vento. «To reject or not to reject: That is the question - An answer in case of neural classifiers». In: *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* 30 (mar. 2000), pp. 84–94. DOI: 10.1109/5326.827457.

[26] Daniel Barbara, Ningning Wu e Sushil Jajodia. «Detecting novel network intrusions using bayes estimators». In: *Proceedings of the 2001 SIAM International Conference on Data Mining*. SIAM. 2001, pp. 1–17.

[27] Bernhard Schölkopf et al. «Estimating Support of a High-Dimensional Distribution». In: *Neural Computation* 13 (lug. 2001), pp. 1443–1471. DOI: 10.1162/089976601750264965.

[28] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer: New York, 2000.

[29] Jacob Devlin et al. «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding». In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies)*. 2019.

[30] Ashish Vaswani et al. «Attention is All you Need». In: *Neural Information Processing Systems*. 2017. URL: https://api.semanticscholar.org/CorpusID:13756489.

[31] Josh Kamps e Bennett Kleinberg. *To the moon: defining and detecting cryptocurrency pump-and-dumps*. Crime Science, 2018.

[32] A. S. Kyle e S. Viswanathan. *How to define illegal price manipulation*. American Economic Review, 2008.

[33] D. B. Kramer. *The way it is and the way it should be: Liability under § 10 (b) of the exchange act and rule 10b-5 thereunder for making false and misleading statements as part of a scheme to pump and dump a stock*. University of Miami Business Law Review, 2005.

[34] S. Alireza Hashemi Golpayegani Neda Soltani Elham Hormizi. *Anomaly Detection in Q&A Based Social Networks*. Advances in Intelligent Systems e Computing, 2018.

[35] Ruixiang Wang Ryan G. Chacon Thibaut G. Morillon. *Will the reddit rebellion take you to the moon? Evidence from WallStreetBets*. Financial Markets e Portfolio Management, 2022.

[36] Mehmet Serdar Guzel Firat Akba Ihsan Tolga Medeni e Iman Askerzade. *Manipulator Detection in Cryptocurrency Markets Based on Forecasting Anomalies*. IEEE International Conference on Artificial Intelligence in Engineering e Technology, 2022.

[37] Fred Morstatter Mehrnoosh Mirtaheri Sami Abu-El-Haija, Greg Ver Steeg e Aram Galstyan. *Identifying and Analyzing Cryptocurrency Manipulations in Social Media*. IEEE Transactions on Computational Social Systems, 2021.

[38] F. Allen e D. Gale. *Stock-price manipulation*. The Review of Financial Studies, 1992.

[39] V.A. Siris e F. Papagalou. «Application of anomaly detection algorithms for detecting SYN flooding attacks». In: *IEEE Global Telecommunications Conference, 2004. GLOBECOM '04.* Vol. 4. 2004, 2050–2054 Vol.4. DOI: 10.1109/GLOCOM.2004.1378372.

[40] F. Sassi La Morgia A. Mei. *Pump and Dumps in the Bitcoin Era: Real Time Detection of Cryptocurrency Market Manipulations*. IEEE, 2020.

[41] P. Martineau. *Inside the group chats where people pump and dump cryptocurrency*. The Outline, 2018.

[42] L. Breiman. *Random forests*. Machine learning, vol. 45, no. 1, pp. 5–32, 2001.

[43] Domenico Alfano, Roberto Abbruzzese e Domenico Parente. «Pump and Dump Cryptocurrency Detection Using Social Media». In: *Proceedings of the 12th International Conference on Data Science, Technology and Applications, DATA 2023, Rome, Italy, July 11-13, 2023*. A cura di Oleg Gusikhin, Slimane Hammoudi e Alfredo Cuzzocrea. SCITEPRESS, 2023, pp. 235–240. ISBN: 978-989-758-664-4. DOI: 10.5220/0012059300003541. URL: https://doi.org/10.5220/0012059300003541.

[44] Bali Ranaivo-Malancon. «Building a Rule-Based Malay Text Segmentation Tool». In: *2011 International Conference on Asian Language Processing*. 2011.

[45] Saniati e Ayu Purwarianti. «Rule based approach for text segmentation on Indonesian news article using named entity distribution». In: *2014 International Conference on Data and Software Engineering (ICODSE)*. 2014.

[46] Omri Koshorek et al. «Text Segmentation as a Supervised Learning Task». In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2018.

[47] Pinkesh Badjatiya et al. «Attention-Based Neural Text Segmentation». In: 2018.

[48] John D. Lafferty, Andrew McCallum e Fernando Pereira. «Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data». In: *International Conference on Machine Learning*. 2001.

[49] Jakub Harasta et al. «Automatic Segmentation of Czech Court Decisions into Multi-Paragraph Parts». In: 2019.

[50] Michal Lukasik et al. «Text Segmentation by Cross Segment Attention». In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020.

[51] Goran Glavaš e Swapna Somasundaran. «Two-Level Transformer and Auxiliary Coherence Modeling for Improved Text Segmentation». In: *Proceedings of the AAAI Conference on Artificial Intelligence* (2020).

[52] Dennis Aumiller et al. «Structural Text Segmentation of Legal Documents». In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*. 2021.

[53] Ilias Chalkidis et al. «LEGAL-BERT: The Muppets straight out of Law School». In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. 2020.

[54] Daniele Licari e Giovanni Comandè. «ITALIAN-LEGAL-BERT: A Pre-trained Transformer Language Model for Italian Law». In: *Companion Proceedings of the 23rd International Conference on Knowledge Engineering and Knowledge Management*. 2022.

[55] Doug Beeferman, Adam L. Berger e John D. Lafferty. «Statistical Models for Text Segmentation». In: *Machine Learning* (1999).

[56] Lev Pevzner e Marti A. Hearst. «A Critique and Improvement of an Evaluation Metric for Text Segmentation». In: *Computational Linguistics* (2002).

[57] Andrea Lombardi, Domenico Alfano e Roberto Abbruzzese. «Legal Text Segmentation Through Breakpoint Detection». In: dic. 2023. ISBN: 9781643684727. DOI: 10.3233/FAIA230968.

[58] Karen Kukich. «Techniques for Automatically Correcting Words in Text». In: *ACM Comput. Surv.* 24.4 (1992), pp. 377–439. ISSN:

0360-0300. DOI: 10.1145/146370.146380. URL: https://doi.org/10.1145/146370.146380.

[59] David Hull e Gregory Grefenstette. «Querying Across Languages: A Dictionary-Based Approach to Multilingual Information Retrieval.» In: ago. 1996, pp. 49–57. DOI: 10.1145/243199.243212.

[60] Sindhu D V e B Sagar. «Dictionary Based Machine Translation from Kannada to Telugu». In: *IOP Conference Series: Materials Science and Engineering* 225 (ago. 2017), p. 012182. DOI: 10.1088/1757-899X/225/1/012182.

[61] Sarah Schulz e Jonas Kuhn. «Multi-modular domain-tailored OCR post-correction». In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. A cura di Martha Palmer, Rebecca Hwa e Sebastian Riedel. Copenhagen, Denmark: Association for Computational Linguistics, set. 2017, pp. 2716–2726. DOI: 10.18653/v1/D17-1288. URL: https://aclanthology.org/D17-1288.

[62] Thi-Tuyet-Hai Nguyen et al. «Adaptive Edit-Distance and Regression Approach for Post-OCR Text Correction: 20th International Conference on Asia-Pacific Digital Libraries, ICADL 2018, Hamilton, New Zealand, November 19-22, 2018, Proceedings». In: nov. 2018, pp. 278–289. ISBN: 978-3-030-04256-1. DOI: 10.1007/978-3-030-04257-8_29.

[63] E.M. Zamora, J.J. Pollock e Antonio Zamora. «The use of trigram analysis for spelling error detection». In: *Information Processing & Management* 17.6 (1981), pp. 305–316. ISSN: 0306-4573. DOI: https://doi.org/10.1016/0306-4573(81)90044-3. URL: https://www.sciencedirect.com/science/article/pii/0306457381900443.

[64] Robert Morris, Lorinda e Cherry. «Computer detection of typographical errors». In: *IEEE Transactions on Professional Communication* PC-18 (1975), pp. 54–56. URL: https://api.semanticscholar.org/CorpusID:8303019.

[65] Edward Ma. *NLP Augmentation*. https://github.com/makcedward/nlpaug. 2019.

[66] Roberto Abbruzzese, Domenico Alfano e Andrea Lombardi. «REMOAC: A Retroactive Explainable Method for OCR Anomalies Correction in Legal Domain». In: dic. 2023. ISBN: 9781643684727. DOI: 10.3233/FAIA230985.

ELENCO DELLE FIGURE

ELENCO DELLE TABELLE