UNIVERSITÀ POLITECNICA DELLE MARCHE

FACOLTÀ DI ECONOMIA "GIORGIO FUÁ"

Corso di Laurea Magistrale in Data Science per l'Economia e le Imprese

Analisi degli Incidenti Causati dall'Intelligenza Artificiale e delle Regolamentazioni dell'EU AI Act

Analysis of Incidents Caused by Artificial Intelligence and the EU AI Act Regulations

Relatore

Prof. Luca Virgili

Candidato

Leonardo Galassi

Anno Accademico 2023-2024

Indice

Elenco delle figure

Abstract

Negli ultimi anni, l'*Intelligenza Artificiale* (IA) ha conosciuto una rapida espansione in diversi settori, portando a significativi benefici ma anche a nuovi rischi e incidenti. Questi incidenti hanno sollevato preoccupazioni riguardanti la sicurezza, l'affidabilità e l'impatto etico delle tecnologie basate su *IA*. In risposta a tali sfide, l'Unione Europea ha proposto l'*EU AI Act*, un quadro normativo mirato a regolamentare l'utilizzo dell'*IA* in modo da garantire un uso sicuro e responsabile di queste tecnologie. Tuttavia, nonostante l'introduzione di questo quadro normativo, è essenziale continuare ad aggiornare la regolamentazione per rimanere al passo con l'evoluzione delle tecnologie emergenti, che presentano nuove sfide e potenziali rischi che devono essere affrontati tempestivamente.

L'obiettivo di questa tesi è analizzare gli incidenti legati all'*Intelligenza Artificiale* (*IA*) e valutare l'efficacia dell'*EU AI Act*, individuando inoltre possibili miglioramenti

da apportare alla regolamentazione in specifici settori dell'*IA*.

Le analisi realizzate hanno evidenziato la necessità di una regolamentazione aggiornata e flessibile. L'*EU AI Act* rappresenta un progresso significativo per la sicurezza e la trasparenza delle tecnologie *IA*, ma l'evoluzione rapida delle tecnologie emergenti richiede un continuo miglioramento delle normative. Le analisi testuali di *Natural Language Processing* (*NLP*) hanno rivelato una percezione negativa degli incidenti, sottolineando l'importanza di misure di sicurezza più rigorose, mentre settori come *AI* e *chatbot* e *deepfake* necessitano di una regolamentazione più mirata per garantire un utilizzo sicuro ed etico dell'*intelligenza artificiale*.

CAPITOLO 1

Introduzione

Negli ultimi anni, l'intelligenza artificiale (*IA*) ha trasformato profondamente numerosi settori migliorando l'efficienza operativa, automatizzando processi complessi e aprendo nuove possibilità per l'innovazione. Tuttavia, con l'espansione delle applicazioni dell'IA, sono emerse anche nuove sfide e rischi. Gli incidenti legati all'utilizzo dell'IA hanno sollevato preoccupazioni significative riguardo alla *sicurezza*, all'*affidabilità* e all'*etica* di questi sistemi. Da un lato, l'IA offre enormi vantaggi in termini di automazione, predizione e *decision-making*; dall'altro, la sua applicazione incontrollata può portare a malfunzionamenti, discriminazioni e persino danni fisici o psicologici.

La crescente complessità e autonomia delle tecnologie *IA* hanno quindi reso evidente la necessità di un quadro normativo che ne regoli l'utilizzo, mitigando i

rischi senza soffocare l'innovazione. Con l'aumento delle applicazioni in settori critici come la sanità, i trasporti e la finanza, diventa fondamentale garantire che tali tecnologie vengano utilizzate in modo sicuro ed etico. A tal proposito, l'Unione Europea ha sviluppato l'*EU AI ACT*, una proposta di legge mirata a creare un ambiente normativo chiaro e sicuro per l'adozione dell'*IA*. Questo atto legislativo intende non solo proteggere i cittadini europei dai potenziali rischi, ma anche promuovere l'innovazione responsabile, assicurando che l'Europa rimanga all'avanguardia nello sviluppo e nell'implementazione delle tecnologie *IA*.

Nell'ambito della ricerca di soluzioni efficaci per la gestione dell'intelligenza artificiale, abbiamo avviato un progetto focalizzato sull'analisi degli incidenti associati all'uso di queste tecnologie. L'obiettivo principale è esaminare un dataset di tali incidenti per rispondere a tre domande chiave: la necessità di regolamentare l'intelligenza artificiale, l'efficacia del regolamento EU AI Act, e le potenziali migliorie da apportare alla regolamentazione in settori specifici dell'IA. Per affrontare queste domande, abbiamo condotto diverse analisi, tra cui una sentiment analysis per valutare le percezioni degli incidenti, e tecniche di NLP per suddividere il dataset in topic e categorie, associando ogni incidente alle categorie di rischio previste dall'EU AI Act. I risultati mostrano che gli incidenti sono percepiti quasi esclusivamente in modo negativo e che molti di essi rientrano nelle categorie di rischio inaccettabile, suggerendo la necessità di miglioramenti nel regolamento, in particolare per quanto

riguarda le tecnologie emergenti.

Questa tesi contribuisce in modo significativo alla letteratura esistente sulla governance dell'intelligenza artificiale, offrendo un'analisi dettagliata e rigorosa degli incidenti associati all'IA tramite l'impiego di tecniche avanzate di NLP e analisi testuale. L'approfondimento condotto permette di delineare un quadro chiaro delle percezioni prevalentemente negative legate agli incidenti, evidenziando in modo puntuale le principali lacune e criticità dell'attuale quadro normativo. Particolare attenzione è riservata alla necessità di rafforzare il regolamento EU AI Act, con raccomandazioni specifiche per affrontare le sfide poste dalle tecnologie emergenti e dalle categorie di rischio inaccettabile.

Il lavoro non solo mette in luce la necessità di una regolamentazione più robusta e adattiva, ma propone anche soluzioni concrete per colmare le attuali carenze. Inoltre, la metodologia sviluppata rappresenta un importante strumento che può essere applicato in ricerche future per monitorare, valutare e migliorare continuamente l'efficacia delle regolamentazioni sull'IA in una varietà di contesti, contribuendo così a una governance più sicura e responsabile dell'intelligenza artificiale.

Nel Capitolo 2 viene fornita una panoramica generale sull'intelligenza artificiale, includendo un'introduzione alla tecnologia, una disamina del suo impatto sulla società e un'analisi del contesto normativo, con particolare riferimento al regolamento EU AI Act. Questo capitolo evidenzia come l'IA stia trasformando i processi tradizionali

e creando nuove opportunità, sottolineando al contempo i principali vantaggi e rischi associati al suo utilizzo.

Nel Capitolo 3, l'analisi si sposta su una trattazione più specifica dei diversi settori in cui l'IA viene applicata, con un focus particolare sulla computer vision e, soprattutto, sul Natural Language Processing (NLP). Viene approfondita la rilevanza del NLP, una disciplina dell'IA che ha registrato una crescita esponenziale negli ultimi anni, e che è stata utilizzata come strumento chiave nelle analisi condotte in questa tesi. Il capitolo esplora in dettaglio le potenziali implicazioni etiche e regolatorie dell'IA, dimostrando come il NLP sia cruciale per affrontare le domande di ricerca e per comprendere meglio le sfide e le opportunità offerte da questa tecnologia.

Nel capitolo 4, vengono presentate le analisi svolte per rispondere alle domande di ricerca sopra menzionate. La prima fase del progetto ha riguardato un'analisi esplorativa del *dataset*, seguita da una serie di analisi approfondite che hanno permesso di comprendere meglio le dinamiche degli incidenti di IA e la necessità di regolamentazione. In particolare, sono stati esaminati i rischi associati a tecnologie emergenti come *chatbot*, veicoli autonomi e *deepfake*, e si è valutata l'efficacia dell'EU AI ACT nel trattare queste problematiche.

Nel Capitolo 5 vengono illustrati i risultati ottenuti dall'analisi, insieme ai benefici e alle limitazioni riscontrate nel corso dello studio. Viene inoltre discusso come le conoscenze esistenti siano state utilizzate e integrate con nuove informazioni

acquisite durante il processo di ricerca.

Infine, il capitolo 6 della tesi è dedicato alla conclusione. Questo capitolo offre una riflessione approfondita sulle aree in cui l'attuale regolamentazione potrebbe essere perfezionata. Vengono formulate raccomandazioni per futuri interventi normativi, con l'obiettivo di ottimizzare l'uso dell'IA, assicurando che sia gestita in modo *sicuro*, *etico* e *vantaggioso* per la società.

CAPITOLO 2

Panoramica sull'Intelligenza Artificiale

2.1 Introduzione all'Intelligenza Artificiale

L'*Intelligenza Artificiale* (IA) rappresenta uno dei campi più *dinamici* e *rivoluzionari* dell'informatica moderna. Questa disciplina, che si propone di creare *sistemi* capaci di emulare o superare le capacità cognitive umane, ha radici profonde nella storia della tecnologia e continua a evolversi a un ritmo sorprendente.

2.1.1 Definizione e Origini

L'IA può essere definita come la branca dell'informatica dedicata alla creazione di *sistemi* in grado di eseguire compiti che tradizionalmente richiedono l'intelligenza umana. Questi compiti includono, ma non si limitano a:

- *Riconoscimento* di pattern visivi e acustici
- *Apprendimento* come la capacità di acquisire nuove conoscenze
- Ragionamento logico e problem-solving
- Pianificazione e presa di decisioni strategiche
- Elaborazione e comprensione del linguaggio naturale

Le origini dell'IA risalgono agli anni '50 del XX secolo, un periodo di fermento intellettuale e tecnologico. Figure pionieristiche come *Alan Turing*, *John McCarthy*, *Marvin Minsky* e *Claude Shannon* gettarono le basi teoriche e pratiche di questo campo. In particolare:

- Alan Turing propose nel 1950 il famoso "*Test di Turing*", un criterio per valutare la capacità di una macchina di esibire un comportamento *intelligente* indistinguibile da quello umano. Questo test, sebbene controverso, rimane un punto di riferimento concettuale nel campo dell'IA.
- John McCarthy, considerato il padre dell'IA, coniò il termine "*Intelligenza Artificiale*" nel 1956 durante la conferenza di Dartmouth, un evento seminale che segnò l'inizio formale della ricerca sull'IA.
- Marvin Minsky, co-fondatore del laboratorio di Intelligenza Artificiale del MIT, contribuì in modo significativo allo sviluppo della teoria e delle applicazioni

pratiche dell'IA, esplorando concetti come le reti neurali e la rappresentazione della conoscenza.

- Claude Shannon, noto come il padre della teoria dell'informazione, applicò le sue idee innovative alla crittografia e alla costruzione di macchine intelligenti, ponendo le basi per lo sviluppo dell'IA attraverso la manipolazione e l'interpretazione dei dati.

2.1.2 Evoluzione Storica e Paradigmi

L'evoluzione dell'IA può essere suddivisa in diverse fasi, ciascuna caratterizzata da approcci e paradigmi distinti:

- *IA Simbolica (1950-1980)*: Anche nota come IA classica, questa fase si concentrava sulla *manipolazione* di simboli e sull'uso della *logica formale* per risolvere problemi. L'idea centrale era che l'intelligenza potesse essere riprodotta attraverso la manipolazione di simboli secondo regole precise. Esempi significativi di questo approccio includono:
  - Sistemi Esperti: programmi che utilizzano una base di *conoscenze* e un *motore inferenziale* per prendere decisioni in domini specifici, come la diagnosi medica o l'analisi geologica.

  - Programmi di Dimostrazione Automatica di Teoremi: sistemi in grado di provare *teoremi* matematici attraverso la logica formale.
  - Pianificatori Automatici: sistemi capaci di generare sequenze di *azioni* per raggiungere obiettivi specifici.

  Nonostante i successi iniziali, l'IA simbolica mostrò limiti significativi nell'affrontare problemi del mondo reale, caratterizzati da *incertezza* e *ambiguità*.

- *Connessionismo e Reti Neurali (1980-1990)*: In risposta ai limiti dell'IA simbolica, emerse un nuovo paradigma ispirato al funzionamento del *cervello umano*. Il *connessionismo* si basa sull'idea che l'intelligenza emerga dall'interazione di semplici unità di elaborazione interconnesse, analoghe ai *neuroni biologici*. Questo approccio portò allo sviluppo di:
  - Reti Neurali Artificiali: modelli computazionali composti da "neuroni" artificiali organizzati in *strati* e connessi tra loro.
  - Algoritmi di Backpropagation: metodi per addestrare le reti neurali attraverso la *correzione degli errori*.

  Sebbene promettente, il connessionismo incontrò difficoltà *tecniche* e *computazionali* che ne limitarono l'applicazione pratica in questa fase.

- *Apprendimento Automatico e Big Data (1990-2010)*: Con l'aumento della *potenza di calcolo* e la disponibilità di grandi quantità di *dati*, l'attenzione si spostò verso approcci basati sull'*apprendimento automatico*. Questa fase vide lo sviluppo di:
  - Algoritmi di Machine Learning: come *Support Vector Machine (SVM)*, *Random Forest* e *Gradient Boosting*.
  - Data Mining: per estrarre *conoscenza* da grandi volumi di dati.
  - Approcci probabilistici: come le *Reti Bayesiane*, per gestire l'*incertezza*.
- *Deep Learning e IA Moderna (2010-presente)*: L'avvento del *deep learning* ha segnato una svolta *rivoluzionaria* nel campo dell'IA. Basato su *reti neurali profonde* con molti *strati*, il deep learning ha portato a progressi straordinari in vari settori:
  - Reti Neurali Convoluzionali (CNN): hanno rivoluzionato il *riconoscimento di immagini* e la *computer vision*.
  - Reti Neurali Ricorrenti (RNN) e Long Short-Term Memory (LSTM): hanno migliorato significativamente l'*elaborazione di sequenze* e il *riconoscimento del linguaggio*.

  - Transformer e modelli di linguaggio di grandi dimensioni: come *GPT* e *BERT*, hanno portato a progressi senza precedenti nell'*elaborazione del linguaggio naturale*.
  - Reti Generative Avversarie (GAN): hanno aperto nuove frontiere nella *generazione di contenuti multimediali*.

2.1.3 Tipologie di IA

L'IA può essere categorizzata in due principali tipologie:

IA Debole (o Ristretta): Si riferisce a *sistemi* progettati per svolgere compiti specifici. Questi sistemi, pur potendo essere estremamente efficienti nel loro dominio, non possiedono una vera *comprensione* o *coscienza*. Esempi includono:

- Assistenti virtuali come *Siri*, *Alexa* o *Google Assistant*, che sono in grado di rispondere a domande, eseguire comandi vocali, gestire calendari e controllare dispositivi domestici intelligenti.
- Sistemi di raccomandazione utilizzati da piattaforme come *Netflix* o *Amazon*, che analizzano i dati delle preferenze degli utenti per suggerire film, serie TV o prodotti che potrebbero interessarli.
- Software di riconoscimento facciale o vocale, impiegati per autenticazione biometrica in smartphone, sicurezza in aeroporti, e applicazioni di sorveglianza.

- Sistemi di guida autonoma, sviluppati da aziende come *Tesla*, *Waymo* e *Uber*, che utilizzano sensori e algoritmi complessi per navigare e operare veicoli senza intervento umano.

- Algoritmi di trading automatico nel settore finanziario, che eseguono operazioni di compravendita di azioni in millisecondi basandosi su modelli predittivi.

- Chatbot utilizzati nel servizio clienti, capaci di rispondere a domande frequenti e fornire assistenza iniziale agli utenti.

IA Forte (o Generale): Si riferisce a un'ipotetica forma di IA che possiede capacità cognitive generali *paragonabili* o superiori a quelle umane. Un'IA forte sarebbe in grado di:

- Comprendere e apprendere qualsiasi compito intellettuale che un essere umano può svolgere, come ad esempio risolvere problemi complessi in vari campi del sapere, dall'ingegneria alla medicina.

- Trasferire *conoscenze* e *abilità* tra domini diversi, per esempio applicando tecniche apprese nella fisica quantistica alla biologia molecolare.

- Possedere *autocoscienza* e *comprensione* profonda, il che implicherebbe una capacità di riflettere su se stessa, avere emozioni e sviluppare una forma di moralità o etica.

- Creare arte originale, come scrivere romanzi, comporre musica o dipingere quadri, con una qualità paragonabile o superiore a quella umana.
- Interagire in modo naturale e significativo con gli esseri umani, comprendendo e rispondendo non solo alle parole ma anche alle emozioni e alle intenzioni dietro di esse.
- Innovare e proporre soluzioni nuove e creative a problemi globali come il cambiamento climatico, la fame nel mondo o le pandemie.

2.2 Impatto dell'IA nella Società

L'Intelligenza Artificiale (IA) sta emergendo come una forza trasformativa in quasi ogni aspetto della società contemporanea, con un impatto che si estende ben oltre i confini della tecnologia. Questa rivoluzione tecnologica offre opportunità senza precedenti per il progresso umano, ma solleva anche questioni etiche, sociali ed economiche di vasta portata che richiedono un'attenta considerazione e regolamentazione.

2.2.1 Applicazioni dell'IA

L'Intelligenza Artificiale (IA) sta trasformando diversi ambiti della società con un impatto significativo sul mondo del lavoro e sull'economia, la sanità, l'educazione, la

sicurezza, la sorveglianza e altri molteplici settori. Nell'economia, *l'automazione* e l'aumento della *produttività* sono evidenti in settori come la manifattura e la logistica, con robot industriali intelligenti e sistemi di gestione del magazzino basati su IA. Tuttavia, l'automazione solleva preoccupazioni sulla *disoccupazione tecnologica*, pur creando nuove opportunità lavorative in campi come lo sviluppo di sistemi IA e l'analisi dei dati. Per affrontare queste sfide, è fondamentale promuovere la *riqualificazione professionale* e l'apprendimento continuo. L'IA può anche esacerbare le *disuguaglianze economiche* se non gestita correttamente, rendendo necessarie politiche per garantire una distribuzione equa dei benefici. Nella sanità, l'IA sta rivoluzionando la *diagnostica avanzata*, la *medicina personalizzata*, la ricerca medica e la telemedicina, migliorando la precisione delle diagnosi e il monitoraggio dei pazienti. Nel campo dell'educazione, l'IA facilita l'*apprendimento personalizzato*, supporta gli educatori e migliora l'accessibilità per studenti con disabilità. Infine, nella sicurezza e sorveglianza, l'IA è utilizzata per *prevenire il crimine*, migliorare la *sicurezza informatica* e solleva importanti questioni etiche e di privacy, richiedendo un bilanciamento tra sicurezza e diritti civili. Questi esempi dimostrano come l'IA stia trasformando profondamente la società, offrendo sia opportunità che sfide.

2.2.2 Sfide Etiche e Regolatorie

Le sfide etiche e regolatorie legate all'intelligenza artificiale sono vastamente complesse e comprendono una serie di problematiche che vanno ben oltre la semplice implementazione tecnologica. Queste sfide si intrecciano con questioni di *giustizia sociale*, *diritti umani* e *sostenibilità ambientale*, richiedendo un'attenta considerazione e un approccio integrato per garantire che l'IA possa servire al meglio l'interesse pubblico.

Uno dei principali problemi riguarda la *trasparenza* e la *spiegabilità* dei sistemi di IA. I modelli avanzati di deep learning, per esempio, sono spesso definiti come vere e proprie "*black box*" perché il processo attraverso il quale giungono alle loro conclusioni è complesso e non facilmente decifrabile. Questa opacità può diventare particolarmente problematica in settori dove le decisioni hanno un impatto diretto e significativo sulla vita delle persone, come nella sanità e nel sistema giudiziario. In ambito sanitario, gli algoritmi di IA possono influenzare le diagnosi e i piani di trattamento, ma senza una comprensione chiara del processo decisionale dell'IA, è difficile per i medici e i pazienti fidarsi completamente delle raccomandazioni generate. La capacità di comprendere e spiegare come un'IA giunge a una decisione è fondamentale non solo per validare le sue raccomandazioni, ma anche per garantire che tali decisioni siano *giuste* e basate su *evidenze*. In ambito giudiziario, la mancanza di trasparenza può sollevare interrogativi sulla *correttezza* e l'*imparzialità* delle

decisioni automatizzate, rischiando di compromettere i principi fondamentali di giustizia ed equità. Per affrontare queste problematiche, è necessario sviluppare tecniche di interpretabilità più sofisticate e strumenti di visualizzazione che rendano il processo decisionale delle IA più chiaro e comprensibile.

Il *bias* e la *discriminazione* rappresentano un'altra sfida cruciale. Gli algoritmi di IA, quando addestrati su dati storici, possono riprodurre e persino amplificare i *pregiudizi* esistenti nei dati. Questo è particolarmente pericoloso in contesti come il reclutamento, la concessione di prestiti e il sistema di giustizia penale. Ad esempio, se un algoritmo di assunzione è addestrato su dati che riflettono pregiudizi di *genere* o *razza*, potrebbe perpetuare tali pregiudizi, discriminando candidati qualificati. Analogamente, negli ambiti finanziari e legali, i modelli di IA possono esacerbare le disuguaglianze esistenti, portando a decisioni che sono *inique* e dannose per determinati gruppi di persone. Per mitigare questo rischio, è essenziale sviluppare e implementare metodologie di rilevamento e correzione dei bias. Ciò include l'uso di tecniche di *auditing* dei dati e degli algoritmi, la *diversificazione* dei dataset utilizzati per l'addestramento e la creazione di meccanismi per monitorare e correggere gli effetti discriminatori. Inoltre, le organizzazioni devono impegnarsi a formare i propri team su questioni di bias e giustizia, promuovendo una *cultura di inclusione* e equità.

La *privacy* e la *protezione dei dati* sono altre aree di grande preoccupazione. Gli algoritmi di IA spesso richiedono enormi quantità di dati personali per funzionare

correttamente, sollevando interrogativi su come questi dati vengono raccolti, conservati e utilizzati. Le preoccupazioni relative alla privacy includono il rischio di *violazioni dei dati*, la raccolta non autorizzata di informazioni e l'uso improprio dei dati personali. È quindi imperativo implementare misure di protezione dei dati rigorose, come la *crittografia*, l'*anonimizzazione* e il *controllo degli accessi*, per garantire che le informazioni sensibili siano al sicuro. Inoltre, è necessario adottare politiche chiare e trasparenti riguardo alla raccolta e all'uso dei dati, assicurandosi che gli utenti diano un *consenso informato* e che abbiano il controllo sui propri dati. Le normative come il *GDPR* (Regolamento Generale sulla Protezione dei Dati) in Europa offrono un modello per garantire che la privacy degli individui sia rispettata, ma è necessario un continuo aggiornamento e rafforzamento di tali normative per affrontare le nuove sfide poste dall'IA.

La questione della *responsabilità* e dell'*accountability* è complessa e cruciale. Quando un sistema di IA causa un errore o un danno, stabilire chi è responsabile è spesso complicato. Questo può includere sviluppatori, utilizzatori del sistema o le stesse organizzazioni che implementano tali tecnologie. È fondamentale sviluppare quadri giuridici che attribuiscano chiaramente la responsabilità per le decisioni e le azioni dei sistemi di IA. Ciò richiede non solo normative specifiche, ma anche l'adozione di pratiche di *audit* e monitoraggio regolari per garantire che i sistemi di IA funzionino in modo sicuro ed etico. Le organizzazioni devono essere pronte a

rispondere in modo adeguato in caso di malfunzionamenti o conseguenze negative, e devono implementare politiche di gestione dei rischi che prevedano procedure di risposta e di risoluzione dei problemi.

L'*impatto ambientale* dell'IA è una preoccupazione crescente. L'addestramento di modelli di IA di grandi dimensioni richiede notevoli risorse computazionali e di energia, con conseguente elevato consumo di energia e emissioni di carbonio. Questo solleva interrogativi riguardo all'*impronta ecologica* delle tecnologie di IA e alla sostenibilità a lungo termine. Per affrontare questo problema, è essenziale sviluppare algoritmi più *efficienti* dal punto di vista energetico e considerare l'impatto ambientale in tutte le fasi dello sviluppo e dell'implementazione dei sistemi di IA. Le organizzazioni devono adottare pratiche di *sostenibilità*, come l'uso di *energie rinnovabili* e l'ottimizzazione dell'efficienza energetica dei data center, per ridurre l'impronta ecologica. È importante che le aziende e i ricercatori lavorino per sviluppare tecnologie di IA che minimizzino il consumo di risorse e promuovano la sostenibilità ambientale.

2.2.3 Prospettive Future

L'intelligenza artificiale sta rapidamente diventando una tecnologia pervasiva con il potenziale di trasformare radicalmente la società. Mentre offre straordinarie opportunità per migliorare la qualità della vita, aumentare l'efficienza e risolvere

problemi complessi, solleva anche importanti sfide etiche, sociali ed economiche.

Per massimizzare i benefici dell'IA e mitigare i rischi, è essenziale:

- Sviluppare *quadri normativi flessibili* e adattivi che possano tenere il passo con i rapidi progressi tecnologici.
- Promuovere un *approccio interdisciplinare* allo sviluppo dell'IA, integrando competenze tecniche con prospettive etiche, sociali e legali.
- Investire nell'*educazione e nella formazione* per preparare la società alle trasformazioni indotte dall'IA.
- Incoraggiare la *ricerca su IA etica, trasparente* e centrata sull'uomo.
- Incentivare il *dialogo tra esperti, decisori politici, aziende e cittadini* per affrontare in modo sinergico le sfide poste dall'IA.

2.3 Regolamentazione Europea dell'IA

L'Unione Europea sta assumendo un ruolo di leadership nella regolamentazione dell'intelligenza artificiale, con l'obiettivo di promuovere un uso etico e responsabile di questa tecnologia rivoluzionaria. La regolamentazione europea dell'*IA* mira a garantire che l'adozione dell'intelligenza artificiale avvenga in modo trasparente, sicuro e rispettoso dei diritti fondamentali degli individui.

Nel 2021, la Commissione Europea ha presentato una proposta di regolamento sull'*IA*, nota come *Artificial Intelligence Act* (AIA). Questo regolamento rappresenta un tentativo pionieristico di creare un quadro normativo omnicomprensivo per l'*IA*, basato su un approccio di gestione del rischio. L'AIA classifica i sistemi di *IA* in base al livello di rischio che rappresentano per i diritti e le libertà fondamentali degli individui, adottando misure regolatorie proporzionali a tali rischi.

2.3.1 Classificazione dei Sistemi di IA

I sistemi di *IA* vengono suddivisi in quattro categorie principali:

1. **Rischio inaccettabile**: Questa categoria include applicazioni di *IA* che rappresentano una minaccia chiara per la sicurezza, i diritti e i valori dell'Unione Europea. Esempi di tali applicazioni sono i sistemi di *IA* che utilizzano il *social scoring* da parte di governi e le tecniche di manipolazione subliminale. Questi sistemi sono vietati all'interno dell'UE. Il divieto di tali applicazioni riflette un forte impegno verso la protezione dei diritti umani e la prevenzione di abusi di potere e discriminazioni sistematiche.

2. **Rischio alto**: I sistemi di *IA* che rientrano in questa categoria richiedono una regolamentazione stringente e il rispetto di requisiti rigorosi prima della loro immissione sul mercato. Esempi di sistemi ad alto rischio includono

quelli utilizzati in infrastrutture critiche (come i trasporti), nell'istruzione, nell'occupazione e nei servizi pubblici essenziali. Questi sistemi devono essere soggetti a controlli e certificazioni per garantire la loro sicurezza e affidabilità. Le autorità competenti devono effettuare valutazioni dettagliate per assicurarsi che tali sistemi non compromettano la sicurezza e la privacy degli individui.

3. **Rischio limitato**: Questa categoria comprende i sistemi di *IA* che presentano un rischio moderato. Per tali sistemi, la normativa richiede principalmente misure di trasparenza, come l'obbligo di informare gli utenti che stanno interagendo con un sistema di *IA*. Un esempio comune di sistemi a rischio limitato sono i chatbot che forniscono assistenza ai clienti. Sebbene questi sistemi non presentino rischi significativi, la trasparenza è fondamentale per mantenere la fiducia degli utenti e garantire un uso consapevole della tecnologia.

4. **Rischio minimo**: I sistemi di *IA* a rischio minimo sono considerati sicuri e non richiedono interventi normativi specifici. In questa categoria rientrano, ad esempio, i filtri antispam e le raccomandazioni di acquisto sui siti di e-commerce. Nonostante la loro semplicità, questi sistemi contribuiscono significativamente all'efficienza delle operazioni quotidiane e migliorano l'esperienza dell'utente.

2.3.2 Requisiti Orizzontali per i Sistemi di IA

L'AIA introduce anche requisiti orizzontali per tutti i sistemi di *IA*, indipendentemente dal loro livello di rischio. Questi includono:

- **Trasparenza**: Gli sviluppatori e gli utilizzatori di sistemi di *IA* devono garantire che i processi decisionali siano trasparenti. Gli utenti devono essere informati quando stanno interagendo con un sistema di *IA*, e devono avere accesso a spiegazioni comprensibili sulle decisioni automatizzate che li riguardano.

- **Tracciabilità**: I sistemi di *IA* devono essere progettati in modo da permettere la tracciabilità delle decisioni. Questo include la possibilità di esaminare i dati di input e le operazioni del sistema per comprendere come sono state prese determinate decisioni.

- **Qualità dei dati**: I dati utilizzati per addestrare i sistemi di *IA* devono essere di alta qualità, accurati e rappresentativi. Questo aiuta a prevenire bias e discriminazioni nei risultati del sistema.

- **Sicurezza**: I sistemi di *IA* devono essere sicuri e resilienti contro attacchi e malfunzionamenti. Devono essere implementate misure di sicurezza per proteggere i dati e le operazioni del sistema.

- **Supervisione umana**: Deve essere garantita la possibilità di intervento umano in tutte le fasi di operazione del sistema di *IA*. Questo principio di *human-in-the-loop* è fondamentale per mantenere il controllo umano sulle decisioni critiche e per prevenire l'abuso di sistemi automatizzati.

2.3.3 Protezione dei Diritti Fondamentali

Un aspetto cruciale della regolamentazione europea è l'enfasi sulla protezione dei diritti fondamentali. L'AIA prevede disposizioni specifiche per prevenire discriminazioni e bias algoritmici, garantendo che i sistemi di *IA* siano progettati e utilizzati in modo equo. Questo include l'obbligo di effettuare valutazioni d'impatto sui diritti fondamentali per i sistemi di *IA* ad alto rischio, nonché misure per garantire la trasparenza e l'esplicabilità degli algoritmi.

La proposta di regolamento sottolinea anche l'importanza della supervisione umana e della responsabilità. Gli operatori di sistemi di *IA* devono garantire che i loro sistemi possano essere disattivati o modificati in caso di malfunzionamenti o comportamenti indesiderati. Questo principio di *human-in-the-loop* è fondamentale per mantenere il controllo umano sulle decisioni critiche e per prevenire l'abuso di sistemi automatizzati.

2.3.4 Strategia Complessiva per l'Innovazione

Oltre alla regolamentazione specifica, l'UE sta promuovendo una strategia complessiva per sostenere l'innovazione e lo sviluppo dell'*IA* in Europa. Questa strategia include:

- **Investimenti in Ricerca e Sviluppo**: L'UE sta aumentando gli investimenti in progetti di ricerca e sviluppo nel campo dell'*IA*, con l'obiettivo di mantenere l'Europa all'avanguardia delle innovazioni tecnologiche. Questo include finanziamenti per progetti di ricerca accademica, collaborazioni pubblico-privato e iniziative di ricerca industriale.

- **Supporto alle Piccole e Medie Imprese (PMI)**: Le PMI sono fondamentali per l'ecosistema dell'innovazione. L'UE sta implementando programmi di supporto per aiutare le PMI a sviluppare e adottare soluzioni di *IA*, fornendo accesso a risorse finanziarie, tecniche e di mentoring.

- **Iniziative per Migliorare le Competenze Digitali**: La formazione continua e l'aggiornamento delle competenze digitali sono essenziali per garantire che la forza lavoro europea possa sfruttare appieno le opportunità offerte dall'*IA*. L'UE sta promuovendo programmi di formazione e riqualificazione per preparare i lavoratori alle sfide della trasformazione digitale.

2.3.5 Dialogo Aperto e Collaborazione

Un elemento fondamentale nella regolamentazione dell'Intelligenza Artificiale a livello europeo è la promozione di un dialogo aperto e collaborativo tra tutte le parti interessate: scienziati, legislatori, industrie e cittadini. Questo approccio è cruciale per affrontare le complesse sfide etiche, legali e tecniche che emergono con l'avanzamento della tecnologia IA.

Il dialogo aperto implica la creazione di canali di comunicazione trasparenti e accessibili, che permettano a tutti i soggetti coinvolti di esprimere le proprie opinioni, preoccupazioni e suggerimenti. La trasparenza è essenziale per costruire fiducia e garantire che le normative siano comprese e accettate dalla società. I cittadini devono avere l'opportunità di comprendere come l'IA influisca sulle loro vite e partecipare al processo decisionale che riguardano le tecnologie emergenti.

La collaborazione tra scienziati e legislatori è particolarmente importante per garantire che le normative siano basate su una solida comprensione scientifica e tecnologica. Gli scienziati possono fornire un'informazione dettagliata sulle potenzialità e i rischi associati all'IA, mentre i legislatori possono tradurre queste informazioni in regolamenti pratici e giuridici. Questa sinergia aiuta a evitare la creazione di normative troppo rigide o poco adeguate, che potrebbero ostacolare l'innovazione o non proteggere adeguatamente i diritti dei cittadini.

L'industria, d'altra parte, gioca un ruolo cruciale nella sperimentazione e imple-

mentazione delle tecnologie IA. Le aziende devono essere coinvolte nel processo di regolamentazione per garantire che le normative siano realistiche e non compromettano l'efficacia delle soluzioni tecnologiche. Inoltre, le imprese possono contribuire a identificare le migliori pratiche e le soluzioni per mitigare i rischi associati all'IA.

Le decisioni riguardanti l'adozione e la regolamentazione dell'IA devono essere basate su una comprensione approfondita delle implicazioni tecnologiche, sociali ed etiche. Per questo motivo, è fondamentale che il dialogo tra tutte le parti sia continuo e strutturato. Questo permette di anticipare e risolvere le problematiche emergenti in modo proattivo e di aggiornare le normative in risposta ai cambiamenti tecnologici e sociali.

In conclusione, un dialogo aperto e collaborativo è la chiave per una regolamentazione efficace dell'IA che equilibri l'innovazione con la protezione dei diritti umani e dei valori fondamentali. Solo attraverso un impegno condiviso e una comunicazione chiara è possibile costruire un quadro normativo che favorisca lo sviluppo responsabile dell'IA e affronti le sfide future in modo equo e informato.

CAPITOLO 3

Settori dell'Intelligenza Artificiale

L'*Intelligenza Artificiale* rappresenta uno dei campi più dinamici e affascinanti della moderna *data science*. Le sue radici affondano nella storia dell'*informatica*, della *matematica* e della statistica, ma è solo negli ultimi decenni che ha visto un'esplosione di interesse e applicazioni pratiche. Le discipline dell'IA sono variegate e interconnesse, ciascuna contribuendo in modo unico al progresso del campo. In questo capitolo, esploreremo le principali discipline che costituiscono l'IA, fornendo una panoramica dei loro fondamenti teorici, delle tecniche più utilizzate e delle applicazioni pratiche.

L'IA si suddivide in numerose sotto-discipline [1], ciascuna con obiettivi e metodologie distinti. Tra le più rilevanti troviamo l'*apprendimento automatico* (*machine*

[1] https://doi.org/10.1080/08874417.2023.2261010

learning), la *computer vision*, il *Natural Language Process* (*NLP*), i *sistemi esperti*, la *robotica* e l'*ottimizzazione*. Ognuna di queste aree contribuisce a rendere le macchine capaci di eseguire compiti che, fino a pochi anni fa, erano esclusivamente dominio dell'intelligenza umana.

3.1 Computer Vision

In questo paragrafo andremo a trattare la computer vision, una disciplina dell'IA che si occupa di come i computer possono acquisire, interpretare e comprendere informazioni visive dal mondo circostante. La computer vision sfrutta algoritmi e tecniche avanzate di intelligenza artificiale e machine learning per permettere alle macchine di "vedere" e analizzare immagini e video, replicando in qualche modo la capacità visiva umana.

3.1.1 Concetti Fondamentali

La *computer vision* è una disciplina complessa che integra concetti fondamentali per permettere ai computer di interpretare e comprendere le informazioni visive. Si tratta di un campo dell'informatica e dell'intelligenza artificiale che si occupa dell'estrazione automatica, analisi e comprensione di informazioni utili da una singola immagine o una sequenza di immagini. L'obiettivo principale della *computer vision*

è quello di replicare e superare le capacità visive umane, consentendo ai computer di identificare, riconoscere e interpretare oggetti e scene nel mondo reale.

Questa disciplina utilizza una varietà di tecniche, che spaziano dall'elaborazione delle immagini alla geometria computazionale, fino all'apprendimento automatico e alle reti neurali. La *computer vision* trova applicazione in numerosi campi e aree dove è fondamentale analizzare e comprendere il contenuto visivo.

Di seguito approfondiamo i concetti principali:

- *Acquisizione dell'immagine*: La prima fase della *computer vision* consiste nella cattura delle immagini. Questo può essere realizzato attraverso vari dispositivi, tra cui fotocamere digitali, scanner e sensori specifici. La qualità dell'immagine, determinata dalla risoluzione e dalla precisione del dispositivo di acquisizione, è cruciale per l'efficacia delle successive fasi di elaborazione. La scelta del dispositivo dipende spesso dall'applicazione specifica, ad esempio, le fotocamere ad alta risoluzione sono essenziali per il riconoscimento facciale, mentre i sensori LIDAR[2] (acronimo dall'inglese Light Detection and Ranging o Laser Imaging Detection and Ranging) sono utilizzati nei veicoli autonomi per mappare l'ambiente circostante in 3D.

- *Pre-elaborazione*: Prima che un'immagine possa essere analizzata, deve essere sottoposta a pre-elaborazione per migliorare la qualità visiva e rimuovere il

[2]https://it.wikipedia.org/wiki/Lidar

rumore. Questo passo può includere una serie di tecniche come:

  - *Filtraggio*: Utilizzato per ridurre il rumore e migliorare i dettagli dell'immagine. I filtri spaziali, ad esempio, aiutano a preservare i bordi mentre rimuovono il rumore.
  - *Normalizzazione*: Questo processo adegua le intensità dei pixel in modo che l'immagine risultante sia più uniforme e priva di distorsioni causate da variazioni di illuminazione.
  - *Trasformazione*: Include operazioni come la trasformazione di Fourier, utilizzata per analizzare le frequenze presenti nell'immagine, e la trasformazione logaritmica, che può migliorare i dettagli nelle immagini con alto contrasto.

- *Segmentazione*: La segmentazione è il processo di suddivisione di un'immagine in regioni significative, solitamente basato su caratteristiche comuni come colore, intensità o texture. Questa fase è critica perché permette di isolare le aree di interesse per ulteriori analisi. Le tecniche di segmentazione più comuni includono:

  - *Thresholding*: Un metodo semplice che separa i pixel in base a un valore di soglia, utile per immagini con contrasti ben definiti.

  - *Crescita delle regioni*: Inizia da un insieme di pixel "seme" e si espande per includere i pixel adiacenti che hanno proprietà simili.
  - *Segmentazione basata su contorno*: Identifica i contorni degli oggetti tramite l'uso di operatori di rilevamento dei bordi come l'operatore di Canny o Sobel.

- *Riconoscimento e descrizione*: Dopo la segmentazione, il sistema deve essere in grado di riconoscere e descrivere le forme e i modelli all'interno delle regioni segmentate. Questo implica l'uso di tecniche avanzate per l'estrazione delle caratteristiche e il matching dei modelli:
  - *Feature extraction*: Estrazione di caratteristiche distintive come angoli, bordi e texture. Tecniche come SIFT (Scale-Invariant Feature Transform) e SURF (Speeded Up Robust Features) sono ampiamente utilizzate per identificare punti di interesse nelle immagini.
  - *Pattern recognition*: Applicazione di algoritmi di *machine learning* e *deep learning* per riconoscere oggetti specifici. Le reti neurali convoluzionali (CNN) sono particolarmente efficaci per il riconoscimento di immagini.

- *Interpretazione e comprensione*: L'obiettivo finale della *computer vision* è interpretare e comprendere il contenuto delle immagini. Questo può includere diversi livelli di analisi:

- *Riconoscimento degli oggetti*: Identificazione e classificazione degli oggetti presenti nell'immagine. Algoritmi come YOLO (You Only Look Once) e R-CNN (Region-Based Convolutional Neural Networks) sono tra i più utilizzati per la rilevazione e classificazione in tempo reale.
- *Stima della posizione e dell'orientamento*: Determinare la posizione spaziale e l'orientamento degli oggetti, che è fondamentale per applicazioni come la robotica e la navigazione autonoma.
- *Comprensione delle relazioni spaziali*: Analizzare le relazioni spaziali tra gli oggetti per comprendere meglio la scena. Questo è essenziale per compiti come la navigazione autonoma, dove è importante sapere non solo cosa sono gli oggetti, ma anche come sono posizionati e relazionati tra loro.

3.1.2 Image Classification

L'*image classification* (Figura 3.1) è uno dei compiti fondamentali della *computer vision*, che consiste nell'assegnare un'etichetta o una classe a un'immagine in base al suo contenuto visivo. Questo processo implica l'analisi e la comprensione delle caratteristiche presenti nell'immagine per determinare a quale categoria essa appartiene. L'*image classification* ha molte applicazioni pratiche, tra cui il riconoscimento

di oggetti, la diagnosi medica assistita da computer, la sorveglianza e la gestione delle risorse naturali.

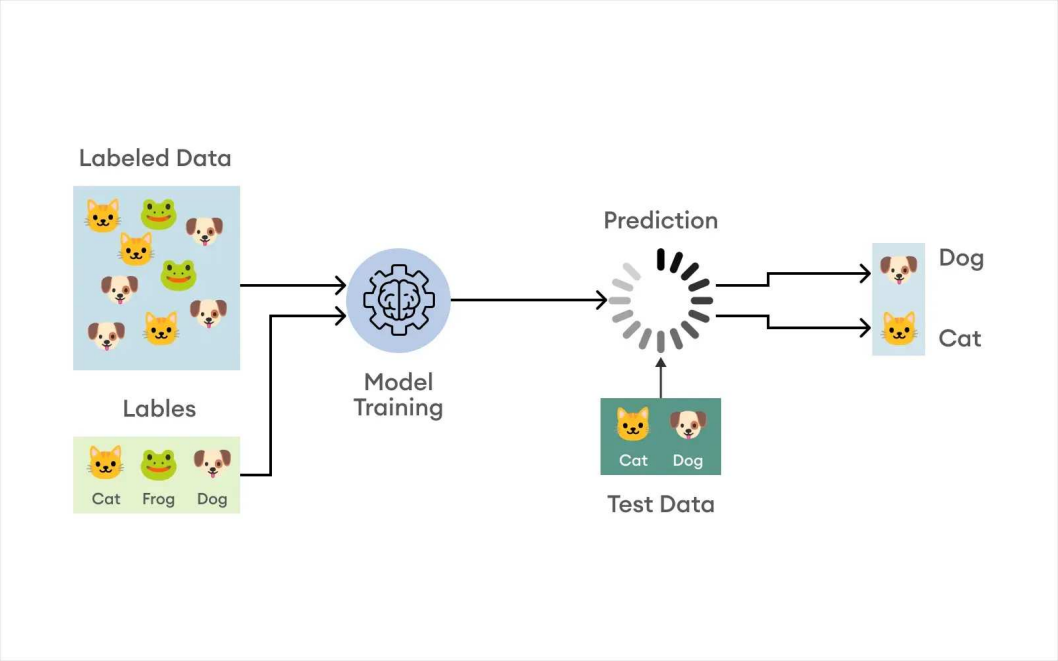

Figura 3.1: Diagramma di classificazione delle immagini

Il processo di *image classification* inizia con la pre-elaborazione delle immagini, fase nella quale la qualità delle immagini stesse viene migliorata e il rumore viene rimosso. Le immagini grezze spesso devono essere filtrate per ridurre il rumore e normalizzate per uniformare le intensità dei pixel. Inoltre, possono essere applicate trasformazioni geometriche per standardizzare le dimensioni e l'orientamento delle immagini, rendendo così più semplice l'analisi successiva.

Una volta completata la pre-elaborazione, si passa all'estrazione delle caratteristiche. Questa fase è cruciale poiché implica l'estrazione di *feature* significative dalle

immagini, come bordi, angoli, texture e forme. Le caratteristiche estratte devono essere rappresentative del contenuto visivo dell'immagine e discriminanti rispetto alle diverse classi. Metodi tradizionali per l'estrazione delle caratteristiche includono SIFT (Scale-Invariant Feature Transform) e HOG (Histogram of Oriented Gradients), mentre le tecniche più recenti fanno uso di reti neurali convoluzionali (CNN) che possono apprendere automaticamente le caratteristiche rilevanti dai dati grezzi.

La selezione delle caratteristiche più rilevanti e la riduzione della dimensionalità sono passi successivi che aiutano a migliorare l'accuratezza e l'efficienza del modello di classificazione. Tecniche come PCA (Principal Component Analysis) e LDA (Linear Discriminant Analysis) vengono utilizzate per ridurre il numero di feature mantenendo quelle più informative.

L'addestramento del modello rappresenta un'altra fase critica del processo. Utilizzando un insieme di immagini etichettate, il modello di classificazione viene addestrato mediante tecniche di apprendimento automatico. Negli ultimi anni, le CNN hanno rivoluzionato il campo dell'*image classification* grazie alla loro capacità di apprendere caratteristiche complesse direttamente dai dati grezzi.

Per garantire che il modello di classificazione sia accurato e generalizzabile, è essenziale eseguire una fase di validazione utilizzando dati di test separati dai dati di addestramento. Metriche come l'accuratezza, la precisione, il richiamo e la curva ROC (Receiver Operating Characteristic) vengono utilizzate per valutare le

prestazioni del modello. Una corretta validazione è fondamentale per assicurarsi che il modello possa operare efficacemente su dati non visti precedentemente.

Una volta che il modello è stato validato, può essere implementato in applicazioni reali. L'*image classification* trova applicazione in numerosi settori [3]. Nel riconoscimento facciale, viene utilizzato in sistemi di sicurezza e per lo sblocco dei dispositivi. Nella diagnostica medica, aiuta a identificare patologie in immagini mediche come radiografie e risonanze magnetiche. Nell'automazione industriale, consente il riconoscimento di difetti nei prodotti e il monitoraggio della qualità. Nelle automobili a guida autonoma, è impiegato per riconoscere segnali stradali, pedoni e altri veicoli, migliorando la sicurezza e l'efficienza della guida.

L'*image classification* è un campo in continua evoluzione, con nuove tecniche e modelli che vengono sviluppati costantemente. Le reti neurali profonde, in particolare, hanno portato a miglioramenti significativi nella precisione e nell'efficienza dei sistemi di classificazione delle immagini. L'integrazione di grandi dataset e l'uso di tecnologie avanzate di elaborazione del calcolo, come le GPU, continuano a spingere i limiti di ciò che è possibile con l'*image classification*, aprendo nuove frontiere per la *computer vision*.

[3] https://www.kaspersky.com/resource-center/definitions/what-is-facial-recognition

3.1.3 Generative Adversarial Networks

Le *Generative Adversarial Networks* (GAN) sono una classe di modelli di *machine learning* introdotti da Ian Goodfellow nel 2014. Le GAN rappresentano una delle innovazioni più significative nel campo dell'apprendimento automatico e della *computer vision*, poiché permettono la generazione di nuovi dati che sono indistinguibili da quelli reali. Questi modelli hanno aperto nuove possibilità in settori come la sintesi di immagini, la creazione di arte digitale, il miglioramento della risoluzione delle immagini e molti altri.

Una GAN è composta da due reti neurali principali, chiamate *generatore* e *discriminatore*, che sono addestrate in modo competitivo. Il *generatore* crea dati falsi a partire da un input casuale, cercando di ingannare il *discriminatore*, il quale invece ha il compito di distinguere tra dati reali e dati generati. Durante l'addestramento, entrambe le reti migliorano continuamente: il *generatore* impara a creare dati sempre più realistici, mentre il *discriminatore* diventa sempre più abile nel rilevare i dati falsi.

Il processo di addestramento delle GAN può essere suddiviso in diverse fasi chiave. Inizialmente, il *generatore* produce un set di dati falsi basandosi su un input di rumore casuale. Questi dati generati vengono quindi combinati con un set di dati reali e presentati al *discriminatore*. Il *discriminatore* valuta se ogni esempio è reale o generato, fornendo feedback al *generatore*. Questo ciclo si ripete molte volte, con

entrambe le reti che migliorano progressivamente le loro capacità.

Una delle sfide principali nell'addestramento delle GAN è la stabilità. Le GAN possono essere difficili da addestrare a causa della natura competitiva del processo, che può portare a instabilità e oscillazioni nelle prestazioni. Varianti delle GAN, come le WGAN (Wasserstein GAN) e le DCGAN (Deep Convolutional GAN), sono state proposte per migliorare la stabilità e l'efficacia dell'addestramento.

Le applicazioni delle GAN sono molteplici e innovative. Nella sintesi di immagini, le GAN possono generare volti umani realistici, paesaggi e oggetti che non esistono nella realtà. Questa capacità è utilizzata in settori come l'arte digitale, la pubblicità e l'intrattenimento. Nella medicina, le GAN possono essere utilizzate per generare immagini mediche sintetiche per migliorare l'addestramento dei modelli di *machine learning* o per aumentare dataset limitati. Inoltre, le GAN sono impiegate nel miglioramento della risoluzione delle immagini, noto come *super-resolution*, dove immagini a bassa risoluzione vengono convertite in versioni ad alta risoluzione con dettagli migliorati.

Per esempio, nei deepfake, il *generatore* della GAN produce immagini sintetiche, mentre il *discriminatore* valuta se le immagini sono reali o generate. Con il tempo, il *generatore* diventa sempre più abile nel creare immagini realistiche che ingannano il *discriminatore*.

Il processo di addestramento delle GAN per la creazione di *deepfake* richiede diverse

fasi. In primo luogo, è necessario raccogliere un vasto dataset di immagini o video della persona target. Queste immagini devono essere di alta qualità e catturare il soggetto da vari angoli e in diverse condizioni di illuminazione. Il *generatore* utilizza queste immagini per imparare a riprodurre le caratteristiche distintive del volto, come la struttura ossea, la texture della pelle e le espressioni facciali.
Durante l'addestramento, il *generatore* crea immagini sintetiche che vengono valutate dal *discriminatore*. Inizialmente, il *discriminatore* è molto bravo a identificare le immagini sintetiche, ma con il tempo, il *generatore* migliora le sue capacità, producendo immagini sempre più realistiche che ingannano il *discriminatore*. Questo ciclo di feedback continuo è ciò che rende le GAN così potenti e in grado di generare *deepfake* estremamente convincenti.

La ricerca sulle GAN continua ad evolversi rapidamente, con nuovi modelli e tecniche che vengono sviluppati per superare le limitazioni esistenti e ampliare le applicazioni. L'integrazione di tecnologie avanzate di calcolo, come le GPU, sta rendendo possibile l'addestramento di GAN sempre più complesse e potenti. Inoltre, l'uso delle GAN in combinazione con altre tecniche di *machine learning* sta aprendo nuove frontiere nella *computer vision* e in molti altri campi, portando a innovazioni significative e nuove possibilità creative.

3.1.4 Applicazioni della Computer Vision

La *computer vision* ha trovato numerose applicazioni in diversi settori, trasformando radicalmente il modo in cui interagiamo con la tecnologia e migliorando molteplici aspetti della vita quotidiana. Tra le applicazioni più rilevanti, possiamo evidenziare l' *autonomous driving*, i *deepfake* e il *face recognition*.

- **Autonomous Driving**: La guida autonoma è una delle applicazioni più ambiziose e innovative della *computer vision*. I veicoli autonomi utilizzano una combinazione di sensori, radar, LIDAR e telecamere per percepire e interpretare l'ambiente circostante. La *computer vision* permette ai veicoli di riconoscere segnali stradali, rilevare ostacoli, identificare corsie e prevedere il comportamento di pedoni e altri veicoli.

  Il processo di guida autonoma inizia con l'acquisizione di immagini e dati dai sensori del veicolo. Questi dati vengono elaborati in tempo reale per costruire una rappresentazione tridimensionale dell'ambiente. Algoritmi avanzati di *computer vision* segmentano le immagini per distinguere tra strade, marciapiedi, veicoli e pedoni. Il riconoscimento degli oggetti è essenziale per identificare segnali stradali, semafori e ostacoli. Inoltre, la stima della posizione e dell'orientamento degli oggetti è cruciale per la navigazione sicura.

  I sistemi di guida autonoma utilizzano CNN per il riconoscimento degli oggetti

e la segmentazione delle immagini. Algoritmi di *deep learning* addestrati su enormi dataset permettono di migliorare continuamente l'accuratezza del sistema. Un esempio noto è il sistema Autopilot di Tesla, che utilizza una combinazione di radar, LIDAR e telecamere per fornire funzionalità di guida autonoma e assistenza alla guida.

Il flusso di dati dai sensori viene elaborato in tempo reale da potenti unità di calcolo a bordo del veicolo. Algoritmi di fusione dei sensori combinano le informazioni provenienti da LIDAR, radar e telecamere per creare una rappresentazione coerente e accurata dell'ambiente. Questo processo di fusione è fondamentale per garantire che il veicolo possa prendere decisioni sicure e informate durante la guida.

Uno degli aspetti cruciali della guida autonoma è la capacità di prevedere il comportamento degli altri utenti della strada. Algoritmi di previsione analizzano i movimenti dei pedoni, dei ciclisti e degli altri veicoli per anticipare le loro azioni e pianificare manovre sicure. Ad esempio, il sistema può prevedere se un pedone è probabile che attraversi la strada o se un veicolo cambierà corsia improvvisamente.

La guida autonoma richiede anche sofisticati algoritmi di controllo e pianificazione del percorso. Questi algoritmi determinano la traiettoria ottimale per il veicolo, tenendo conto delle condizioni del traffico, delle leggi stradali e

delle preferenze di guida. Il controllo del veicolo viene gestito da sistemi che regolano la velocità, la sterzata e la frenata per seguire la traiettoria pianificata in modo sicuro ed efficiente.

La sicurezza è una priorità assoluta nella guida autonoma. I veicoli autonomi sono dotati di sistemi di ridondanza per garantire che un guasto in un componente non comprometta la sicurezza complessiva del sistema. Ad esempio, se un sensore LIDAR smette di funzionare, il sistema può fare affidamento su radar e telecamere per continuare a operare in sicurezza. Inoltre, i veicoli autonomi sono programmati per seguire protocolli di emergenza in caso di situazioni critiche, come guasti ai componenti o condizioni stradali impreviste.

Nonostante i notevoli progressi, la guida autonoma presenta ancora diverse sfide tecniche e regolatorie [4]. Una delle principali sfide è l'addestramento dei sistemi di *machine learning* su dataset sufficientemente vari e completi da coprire tutte le possibili situazioni stradali. Inoltre, la guida autonoma richiede un'infrastruttura di comunicazione avanzata per consentire la condivisione di informazioni tra veicoli e infrastrutture stradali.

Le questioni etiche e legali legate alla guida autonoma sono altrettanto importanti. La responsabilità in caso di incidenti, la privacy dei dati raccolti dai veicoli e l'accettazione sociale della tecnologia sono aspetti che devono essere

[4] https://www.technologyreview.com/2015/07/29/166941/how-to-help-self-driving-cars-make-ethical-decisions/

affrontati con attenzione. Le normative devono evolversi per garantire che i veicoli autonomi possano operare in modo sicuro e conforme alle leggi stradali.

- **Deepfake**: I *deepfake* sono una delle applicazioni più controverse e impressionanti della *computer vision*. Utilizzando tecniche di *deep learning*, in particolare le GAN, è possibile creare video e immagini che sembrano reali ma che in realtà sono completamente sintetici. I *deepfake* possono essere utilizzati per creare video di persone che dicono o fanno cose che in realtà non hanno mai detto o fatto.

  La creazione di *deepfake* inizia con l'addestramento di una GAN su un ampio dataset di immagini o video della persona che si desidera simulare. Il *generatore* della GAN produce immagini sintetiche, mentre il *discriminatore* valuta se le immagini sono reali o generate. Con il tempo, il *generatore* diventa sempre più abile nel creare immagini realistiche che ingannano il *discriminatore*.

  Il processo di addestramento delle GAN per la creazione di *deepfake* richiede diverse fasi. In primo luogo, è necessario raccogliere un vasto dataset di immagini o video della persona target. Queste immagini devono essere di alta qualità e catturare il soggetto da vari angoli e in diverse condizioni di illuminazione. Il *generatore* utilizza queste immagini per imparare a riprodurre le caratteristiche distintive del volto, come la struttura ossea, la texture della pelle e le espressioni facciali.

Durante l'addestramento, il *generatore* crea immagini sintetiche che vengono valutate dal *discriminatore*. Inizialmente, il *discriminatore* è molto bravo a identificare le immagini sintetiche, ma con il tempo, il *generatore* migliora le sue capacità, producendo immagini sempre più realistiche che ingannano il *discriminatore*. Questo ciclo di feedback continuo è ciò che rende le GAN così potenti e in grado di generare *deepfake* estremamente convincenti.

Una volta che il *generatore* è sufficientemente addestrato, può creare video di persone che parlano o eseguono azioni che non hanno mai compiuto. Questo è possibile combinando le immagini generate con tecniche di animazione che sincronizzano i movimenti delle labbra e le espressioni facciali con un audio pre-registrato. Il risultato finale è un video che appare completamente autentico, ma che è in realtà una creazione artificiale.

I *deepfake* hanno applicazioni sia positive che negative. Positivamente, possono essere utilizzati in produzioni cinematografiche per effetti speciali, permettendo agli attori di apparire in scene senza essere fisicamente presenti. Possono anche essere impiegati nel restauro di film, per riportare in vita attori scomparsi o per migliorare la qualità delle immagini in vecchi filmati. Inoltre, i *deepfake* trovano applicazione nell'educazione e nella ricerca, dove possono essere utilizzati per creare simulazioni e modelli didattici.

Tuttavia, i *deepfake* possono anche essere utilizzati per scopi malevoli, come

la diffusione di disinformazione, la creazione di notizie false e la violazione della privacy [5]. Video falsi di politici, celebrità o persone comuni possono essere creati e diffusi per manipolare l'opinione pubblica, estorcere denaro o compromettere la reputazione di individui. Questo ha sollevato preoccupazioni etiche e legali significative, portando allo sviluppo di tecnologie per rilevare e contrastarli.

La rilevazione dei *deepfake* è un campo di ricerca attivo. Gli algoritmi di rilevamento analizzano le immagini e i video alla ricerca di imperfezioni che possono rivelare la natura sintetica del contenuto. Queste imperfezioni includono inconsistenze nell'illuminazione, artefatti nei bordi del volto, anomalie nei movimenti delle labbra e degli occhi, e discrepanze nei pattern di pixel.

Nonostante le sfide etiche e legali, la tecnologia dei *deepfake* continua a evolversi e migliorare. La comprensione delle loro potenzialità e dei loro rischi è cruciale per sfruttare al meglio questa potente tecnologia, garantendo al contempo la protezione dei diritti individuali e la veridicità delle informazioni.

- **Face Recognition**: Il riconoscimento facciale è una delle applicazioni più diffuse della *computer vision*. Questa tecnologia consente di identificare e verificare l'identità di una persona analizzando le caratteristiche del viso. Utilizzato in una varietà di settori, il riconoscimento facciale ha applicazioni

[5] https://doi.org/10.1016/j.clsr.2022.105716

in sicurezza, sorveglianza, marketing e interazione con i dispositivi.

Il processo di riconoscimento facciale inizia con l'acquisizione di un'immagine del volto attraverso una fotocamera. L'immagine viene quindi elaborata per rilevare il volto e estrarre le caratteristiche rilevanti, come la distanza tra gli occhi, la forma del naso, la lunghezza della mascella e altri tratti distintivi. Queste caratteristiche vengono convertite in un vettore numerico, noto come *embedding*, che rappresenta il volto in uno spazio n-dimensionale.

Per identificare una persona, l'*embedding* del volto acquisito viene confrontato con quelli presenti in un database di volti noti utilizzando tecniche di *machine learning* e *deep learning*. Algoritmi come le reti neurali convoluzionali sono particolarmente efficaci per l'estrazione delle caratteristiche e la comparazione dei volti. Il riconoscimento facciale è utilizzato in applicazioni di sicurezza, come il controllo degli accessi e la sorveglianza, oltre che in applicazioni commerciali, come il pagamento senza contatto e il marketing personalizzato.

La tecnologia di riconoscimento facciale si basa su una serie di fasi critiche. In primo luogo, il sistema deve rilevare la presenza di un volto all'interno di un'immagine o di un video. Questo viene solitamente effettuato utilizzando algoritmi di rilevamento dei volti che analizzano i pixel dell'immagine per individuare caratteristiche tipiche del volto umano, come la simmetria e le proporzioni relative delle varie parti del viso.

Una volta rilevato il volto, il sistema procede con l'estrazione delle caratteristiche. Questo passaggio implica la misura e l'analisi delle diverse parti del volto, come la distanza tra gli occhi, la larghezza del naso, la forma della bocca e la struttura della mascella. Queste informazioni vengono utilizzate per creare un *embedding*, una rappresentazione numerica unica del volto che può essere facilmente confrontata con altre rappresentazioni.

Il riconoscimento facciale ha numerose applicazioni pratiche. In ambito di sicurezza, è utilizzato per il controllo degli accessi in edifici sicuri, aeroporti e altri luoghi sensibili. Ad esempio, molti smartphone moderni utilizzano il riconoscimento facciale per sbloccare il dispositivo, fornendo un metodo di autenticazione sicuro e conveniente. Inoltre, la tecnologia è impiegata in sistemi di sorveglianza per identificare individui sospetti o ricercati in tempo reale.

In ambito commerciale, il riconoscimento facciale può migliorare l'esperienza del cliente. Ad esempio, nei negozi al dettaglio, può essere utilizzato per riconoscere i clienti abituali e offrire loro un servizio personalizzato. Inoltre, alcune aziende stanno sperimentando il pagamento tramite riconoscimento facciale, eliminando la necessità di contanti o carte di credito. Questo non solo velocizza le transazioni ma riduce anche il rischio di frodi.

Anche in questo caso, l'uso diffuso del riconoscimento facciale solleva anche

importanti questioni di privacy ed etica. La raccolta e l'archiviazione di dati biometrici sensibili possono portare a preoccupazioni riguardo la sorveglianza di massa e l'uso improprio delle informazioni personali. È essenziale che le tecnologie di riconoscimento facciale siano implementate con rigorose misure di protezione dei dati e regolamenti chiari per prevenire abusi.

In conclusione, le applicazioni della *computer vision* continuano a espandersi con l'evoluzione delle tecnologie e l'aumento della potenza di calcolo. La capacità di interpretare e comprendere le informazioni visive sta trasformando molteplici settori, portando innovazioni significative e migliorando la qualità della vita.

3.2 Natural Language Processing

Il *Natural Language Processing* (NLP) si distingue per la sua capacità di permettere alle macchine di comprendere, interpretare e generare *linguaggio umano* in modo naturale. Questa disciplina si colloca all'incrocio tra *linguistica*, *informatica* e *intelligenza artificiale*, e si propone di colmare il divario tra il linguaggio umano, ricco di sfumature e complessità, e il linguaggio formale delle macchine.

Il NLP si basa su una serie di tecniche e algoritmi che analizzano il testo e il parlato per estrarre significato, identificare relazioni e persino generare contenuti. Le applicazioni di questa tecnologia sono molteplici e spaziano dai *chatbot* e assistenti

virtuali, come *Siri* e *Alexa*, ai sistemi di traduzione automatica, come *Google Translate*, fino all'analisi del sentiment nei social media. Queste applicazioni non solo migliorano l'interazione tra uomo e macchina, ma offrono anche strumenti preziosi per l'analisi dei dati e la comprensione delle tendenze sociali.

Tuttavia, il percorso verso una comprensione completa del linguaggio naturale è costellato di sfide. La varietà di lingue, dialetti, espressioni idiomatiche e contesti culturali rende il compito di un sistema NLP estremamente complesso. Inoltre, la necessità di gestire ambiguità, ironia e sottintesi richiede modelli sempre più sofisticati, spesso basati su tecniche di *apprendimento profondo*.

Esploreremo i fondamenti del Natural Language Processing, analizzando le sue tecniche principali, le sfide attuali e le prospettive future. Attraverso una revisione della letteratura e casi studio, cercheremo di comprendere come il NLP stia plasmando il nostro rapporto con la tecnologia e quali opportunità offre per il futuro.

3.2.1 Concetti Fondamentali

Come anticipato in precedenza, *il Natural Language Processing* è un campo interdisciplinare che integra linguistica, informatica e intelligenza artificiale per permettere alle macchine di comprendere, interpretare e generare il linguaggio umano in modo significativo. I concetti fondamentali del NLP si possono suddividere in diverse nozioni chiave:

1. **Tokenizzazione**. La *tokenizzazione* è il processo di suddivisione di un testo in unità più piccole, chiamate *token*. Questi token possono essere parole, frasi o anche caratteri singoli. La tokenizzazione è il primo passo nel NLP, poiché fornisce la base per ulteriori analisi testuali.

2. **Lemmatizzazione e Stemming**. La *lemmatizzazione* riduce le parole alla loro forma base o *lemma*, considerando il contesto lessicale. Ad esempio, "mangiando" e "mangiato" vengono ricondotti a "mangiare". Lo *stemming*, invece, rimuove i suffissi dalle parole per ottenere una radice comune, ad esempio "correndo" e "corse" diventano "corr". La lemmatizzazione è più accurata rispetto allo stemming, ma anche più complessa.

3. **Parsing Sintattico**. Il *parsing sintattico* implica l'analisi della struttura grammaticale di una frase. Esso identifica le relazioni tra le parole e crea un albero sintattico che rappresenta la struttura gerarchica della frase. Questo processo è essenziale per comprendere il significato e il contesto delle frasi.

4. **Rappresentazione Semantica**. La *rappresentazione semantica* cerca di catturare il significato del testo. Tecniche come i *word embeddings* (ad esempio BERT) rappresentano le parole come vettori in uno spazio multidimensionale, dove la distanza tra i vettori riflette la somiglianza semantica. Questi modelli permettono di gestire compiti complessi come l'analisi del sentiment e la traduzione automatica.

5. **Named Entity Recognition** (NER). Il *NER* è il processo di identificazione e

classificazione delle entità menzionate nel testo in categorie predefinite come nomi di persone, organizzazioni, luoghi, date e altre entità rilevanti. Questa tecnica è cruciale per l'estrazione di informazioni e per la comprensione del contesto in cui le entità sono menzionate.

6. **Trasformer e Modelli Pre-addestrati**. I modelli di *trasformer*, come BERT e GPT (che saranno approfonditi nei prossimi paragrafi), hanno rivoluzionato il campo del NLP. Questi modelli utilizzano meccanismi di *attenzione* per catturare le dipendenze a lungo raggio nel testo e sono pre-addestrati su grandi corpora di dati, consentendo loro di apprendere rappresentazioni linguistiche profonde. Successivamente, possono essere *fine-tuned* per specifici compiti NLP.

7. **Machine Translation**. La *traduzione automatica* è uno dei compiti più complessi del NLP, che coinvolge la conversione di testo da una lingua a un'altra. I modelli di *traduzione neurale* (NMT) hanno superato i metodi tradizionali basati su regole e statistici, utilizzando reti neurali profonde per ottenere traduzioni più fluide e accurate.

8. **Sentiment Analysis**. La *sentiment analysis* implica la classificazione delle opinioni espresse nel testo in categorie come positivo, negativo o neutro. Questo compito è ampiamente utilizzato nel monitoraggio dei social media, nelle recensioni dei prodotti e nel *customer feedback*.

3.2.2 Transformer

I transformer sono una classe di modelli di deep learning introdotti da Vaswani et al. nel 2017, che hanno rivoluzionato il campo dell'elaborazione del linguaggio naturale (NLP). La loro architettura si basa principalmente sul meccanismo di self-attention, che consente al modello di attribuire diversi pesi a diverse parti della sequenza di input, migliorando così la capacità di catturare le dipendenze a lungo termine nel testo. Questa caratteristica li rende particolarmente efficaci nel trattare sequenze di dati, come testi, dove la comprensione del contesto e delle relazioni tra parole è cruciale.

Una delle principali innovazioni dei transformer è l'abilità di processare simultaneamente tutte le parole di una frase, piuttosto che una alla volta come avviene nei modelli sequenziali tradizionali come le RNN (Recurrent Neural Networks). Questo è reso possibile dal meccanismo di self-attention, che permette al modello di mettere in relazione ogni parola con ogni altra parola nella stessa frase, indipendentemente dalla loro distanza reciproca. Questo approccio consente ai transformer di catturare relazioni a lungo termine molto più efficacemente rispetto ai modelli sequenziali, che tendono a dimenticare le informazioni man mano che la sequenza si allunga.

L'architettura del transformer[6] è composta da due parti principali: l'encoder e il decoder. L'encoder è responsabile dell'elaborazione della sequenza di input e della

[6]https://huggingface.co/learn/nlp-course/it/chapter1/4

costruzione di un embedding della stessa. Questo processo avviene attraverso una serie di strati, ognuno dei quali applica meccanismi di self-attention e reti neurali feed-forward. Il decoder, invece, utilizza questa rappresentazione per generare una sequenza di output, anch'esso attraverso una serie di strati simili. Questo schema encoder-decoder rende i transformer particolarmente adatti per compiti di traduzione automatica, dove una sequenza di testo in una lingua deve essere convertita in una sequenza di testo in un'altra lingua.

Un'altra caratteristica chiave dei transformer è l'uso del meccanismo di "positional encoding", che consente al modello di mantenere informazioni sull'ordine delle parole nella sequenza di input. Poiché il self-attention tratta tutte le parole simultaneamente e indipendentemente dalla loro posizione, senza un modo per codificare l'ordine delle parole il modello non potrebbe distinguere tra diverse permutazioni delle stesse parole. Il positional encoding risolve questo problema aggiungendo una componente di posizione alle rappresentazioni delle parole, permettendo al modello di apprendere non solo quali parole sono presenti, ma anche in che ordine appaiono.

I transformer hanno dimostrato una straordinaria capacità di generalizzare su una vasta gamma di compiti NLP, non solo nella traduzione automatica, ma anche nella comprensione del linguaggio, nella generazione di testo e in molte altre applicazioni [7]. Questo ha portato a un'adozione diffusa dei transformer in molte aree della ricerca

[7] https://www.analyticsvidhya.com/blog/2019/06/understanding-transformers-nlp-state-of-the-art-models/

e dell'industria, con modelli come BERT e GPT che diventano standard de facto per numerosi compiti di NLP.

BERT[8] (Bidirectional Encoder Representations from Transformers) è un modello di language representation basato sull'architettura del transformer, sviluppato da Google. La sua innovazione principale è l'uso di una pre-training bidirezionale che permette al modello di avere una comprensione più profonda del contesto. BERT viene pre-addestrato su enormi corpus di testi utilizzando due compiti: il "Masked Language Model" (MLM), dove alcune parole nella frase sono mascherate e il modello deve predirle, e il "Next Sentence Prediction" (NSP), dove il modello predice se una frase segue un'altra. Questo approccio bidirezionale consente a BERT di catturare relazioni più complesse tra le parole rispetto ai modelli tradizionali unidirezionali.

GPT[9] (Generative Pre-trained Transformer) è un'altra famiglia di modelli basati sui transformer, sviluppata da OpenAI. A differenza di BERT, GPT è un modello autoregressivo, il che significa che predice il testo parola per parola, utilizzando il contesto delle parole precedenti per generare la parola successiva. Questo lo rende particolarmente efficace nei compiti di generazione del testo, come la scrittura automatica e la risposta a domande. GPT viene pre-addestrato su grandi quantità di testo non etichettato e poi fine-tuned su specifici compiti. La sua architettura

[8]https://it.wikipedia.org/wiki/BERT
[9]https://aws.amazon.com/it/what-is/gpt/

unidirezionale consente una generazione di testo più fluida e coerente, risultando in applicazioni pratiche notevoli, tra cui i chatbot avanzati e gli assistenti virtuali.

3.2.3 Possibili Applicazioni

I modelli basati su transformer hanno trovato applicazione in una vasta gamma di compiti nel campo degli NLP, grazie alla loro capacità di comprendere e generare testo con un alto grado di precisione e coerenza. Le loro applicazioni, come detto in precedenza, spaziano dalla traduzione automatica alla sintesi del testo, passando per l'analisi del sentiment e molto altro. Una delle aree in cui i transformer hanno avuto un impatto particolarmente significativo è quella dei chatbot.

I *chatbot* sono applicazioni progettate per interagire con gli utenti tramite il linguaggio naturale, simulando una conversazione umana. I modelli transformer, grazie alla loro capacità di comprendere il contesto e generare risposte coerenti, sono diventati fondamentali nello sviluppo di chatbot avanzati.

L'uso dei transformer nei chatbot consente una gestione più efficace delle conversazioni complesse rispetto ai modelli tradizionali. Ad esempio, i modelli come GPT possono generare risposte in modo fluido e naturale, grazie alla loro capacità di prevedere e produrre testo basato sul contesto delle conversazioni precedenti. Questo permette ai chatbot di mantenere conversazioni più coerenti e di adattarsi meglio alle diverse esigenze degli utenti.

Inoltre, i transformer possono migliorare la comprensione delle intenzioni degli utenti attraverso il meccanismo di self-attention, che consente al modello di considerare l'intero contesto della conversazione piuttosto che solo le ultime interazioni. Questo approccio aiuta i chatbot a fornire risposte più pertinenti e contestualmente appropriate.

I *chatbot* basati su transformer sono anche in grado di gestire una varietà di compiti, dalla semplice risposta a domande frequenti alla conduzione di conversazioni più complesse che richiedono una comprensione profonda del testo. Essi possono essere utilizzati in diversi contesti, come assistenti virtuali per il servizio clienti, strumenti di supporto per l'e-commerce, e anche in ambito educativo per offrire tutoraggio e supporto agli studenti.

L'integrazione dei transformer nei chatbot non solo migliora l'efficacia delle interazioni, ma contribuisce anche a un'esperienza utente più soddisfacente e coinvolgente. Questo ha portato a una crescente adozione di chatbot basati su transformer in molte industrie, dove sono utilizzati per automatizzare e ottimizzare le comunicazioni con i clienti e per fornire un servizio più personalizzato e reattivo.

CAPITOLO 4

Analisi dell'EU AI ACT

In questo capitolo, ci concentreremo sull'analisi di un dataset riguardante incidenti attribuiti all'uso dell'intelligenza artificiale. L'obiettivo è esplorare le questioni cruciali emerse da questi eventi per individuare possibili aree di miglioramento del regolamento europeo sull'intelligenza artificiale (EU AI Act) e altre normative pertinenti. Inoltre, valuteremo se l'EU AI Act copre adeguatamente questi incidenti, considerando se vi siano lacune da colmare per prevenire tali eventi in futuro. Attraverso un esame approfondito dei dati, ci porremo domande fondamentali su come prevenire tali incidenti in futuro, garantendo che l'adozione dell'intelligenza artificiale avvenga in modo sicuro ed etico. Questa analisi mira a offrire spunti e suggerimenti utili per rafforzare la gestione dell'IA in diversi contesti.

4.1 Introduzione al Dataset

In questa sezione, forniremo una descrizione dettagliata del dataset utilizzato per l'analisi. Esamineremo attentamente le variabili che lo compongono, analizzandone la struttura e le caratteristiche principali. Questo ci permetterà di comprendere meglio i dati a disposizione e di impostare l'analisi in modo efficace, ponendo le basi per le valutazioni successive.

4.1.1 AI Incidents

Il dataset "AI Incidents" è una risorsa fondamentale per chi studia o lavora nel campo dell'intelligenza artificiale, specialmente in relazione agli aspetti di sicurezza, etica e governance. Raccolto dalla piattaforma "incidentsdatabase.ai"[1], questo dataset offre una documentazione dettagliata di una serie di incidenti in cui l'utilizzo di sistemi di intelligenza artificiale ha portato a conseguenze negative, evidenziando le sfide e i rischi associati all'adozione di queste tecnologie in diversi contesti applicativi.

Il sito, da cui è stato preso il database, è una piattaforma dedicata alla raccolta e alla diffusione di informazioni riguardanti incidenti legati all'intelligenza artificiale. La piattaforma nasce con l'intento di aumentare la trasparenza e la consapevolezza riguardo agli errori, ai fallimenti e ai rischi associati all'uso di sistemi IA. Esso offre

[1] https://incidentdatabase.ai/

un archivio accessibile al pubblico, che permette di esplorare e comprendere meglio gli effetti negativi che possono derivare dall'uso improprio di queste tecnologie avanzate.

Il database si arricchisce grazie a un duplice approccio: le segnalazioni degli utenti e un'attività di ricerca meticolosa. Include casi documentati in pubblicazioni accademiche, fonti giornalistiche attendibili e altre risorse verificate. Questo metodo assicura che la piattaforma rimanga una fonte aggiornata e autorevole per chi desidera esplorare gli incidenti correlati all'intelligenza artificiale. Il dataset viene costantemente aggiornato in tempo reale, incorporando nuovi incidenti non appena vengono riportati e verificati.

4.1.2 Descrizione delle Variabili del Dataset

Il dataset contiene informazioni riguardanti incidenti legati all'utilizzo di sistemi di intelligenza artificiale. Ogni record rappresenta un incidente specifico e include una serie di variabili che descrivono vari aspetti dell'evento. Di seguito viene fornita una descrizione dettagliata delle variabili presenti nel dataset, seguita da un'analisi critica delle stesse. Il dataset è composto da 720 record e 9 colonne, di cui 8 sono di tipo testuale (`object`) e una è di tipo numerico intero (`incident_id`).

- **_id**: rappresenta un identificatore univoco per ciascun record all'interno del dataset. È *essenziale* per mantenere l'integrità e la tracciabilità dei dati.

- **incident_id**: è un identificatore numerico univoco che distingue ciascun incidente all'interno del dataset. Questa variabile è *fondamentale* per tracciare e riferirsi a singoli incidenti in modo coerente, specialmente quando si effettuano analisi comparative o si esaminano pattern specifici all'interno del dataset.

- **date**: indica la data in cui è stata riportata l'informazione ad esso relativa. Questa variabile è di tipo *data*, il che consente di effettuare analisi temporali come la valutazione delle tendenze nel tempo e l'identificazione di periodi con una maggiore concentrazione di incidenti.

- **reports**: contiene una lista di identificatori numerici, ciascuno dei quali corrisponde a un report specifico associato all'incidente documentato. Ogni identificatore rappresenta un documento o una fonte esterna che fornisce informazioni aggiuntive e approfondite relative all'incidente. Questi report possono includere dettagli come analisi tecniche, valutazioni delle cause dell'incidente, testimonianze di testimoni, o resoconti ufficiali provenienti da enti regolatori o investigativi. La presenza di più report collegati a un singolo incidente suggerisce una maggiore complessità o rilevanza dell'evento, offrendo così un contesto più ampio e articolato per l'analisi.

- **Alleged deployer of AI system**: raccoglie una lista di entità o organizzazioni *implementatrici* del sistema di IA coinvolto nell'incidente. I "deployer" sono

coloro che hanno distribuito o utilizzato il sistema di IA nel contesto in cui si è verificato l'incidente. L'analisi di questa variabile può fornire informazioni su quali entità sono più frequentemente coinvolte in incidenti, permettendo di identificare potenziali problemi sistematici o responsabilità ricorrenti.

- **Alleged developer of AI system**: similmente, elenca gli sviluppatori del sistema di IA implicato nell'incidente. Gli sviluppatori sono le entità che hanno progettato e creato il sistema di IA. Questa variabile è *fondamentale* per identificare quali aziende o team di sviluppo sono più frequentemente associati a incidenti, offrendo spunti per un'analisi più approfondita delle pratiche di sviluppo, dei test di qualità e delle responsabilità etiche nel processo di creazione dei sistemi di intelligenza artificiale.

- **Alleged harmed or nearly harmed parties**: indica le persone o i gruppi che sono stati danneggiati o che sono stati quasi danneggiati dall'incidente. La comprensione di questa variabile è *cruciale* per valutare l'impatto sociale e umano degli incidenti legati ai sistemi di IA. Analizzare chi sono le parti lese può aiutare a identificare gruppi vulnerabili o a rischio, evidenziando la necessità di regolamentazioni o protezioni aggiuntive per determinate categorie di utenti.

- **description**: fornisce una descrizione testuale dell'incidente. Questo campo è *fondamentale* per comprendere nel dettaglio cosa è accaduto, come si è svolto l'incidente e quali sono stati gli effetti. L'analisi di queste descrizioni può contribuire a identificare temi comuni tra diversi incidenti, fornendo una base per la categorizzazione degli incidenti e per l'analisi del sentiment (attraverso strumenti di NLP).

- **title**: rappresenta una sintesi dell'incidente, descrivendolo in modo conciso. Il titolo è utile per una rapida identificazione e categorizzazione dell'incidente.

Le variabili descritte offrono una visione dettagliata degli incidenti legati ai sistemi di intelligenza artificiale, consentendo una vasta gamma di analisi, dalla valutazione dei trend temporali all'identificazione di responsabilità specifiche e all'impatto sociale. L'uso di questo dataset in un contesto di ricerca può contribuire significativamente alla comprensione dei rischi associati all'implementazione e allo sviluppo di sistemi di IA, offrendo spunti per migliorare la sicurezza, l'affidabilità e l'etica di tali sistemi.

4.1.3 Analisi Esplorativa

Nel contesto della crescente diffusione dei sistemi di IA, è essenziale comprendere e analizzare gli incidenti legati a queste tecnologie. Attraverso un'analisi esplorativa,

cercheremo di individuare i principali problemi e di ottenere una visione complessiva delle informazioni fornite dal dataset. In questo lavoro, ci proponiamo di esplorare tre aree principali:

La **distribuzione temporale** degli incidenti rappresenta un aspetto fondamentale per comprendere le dinamiche e le tendenze nel tempo. Attraverso una dettagliata analisi temporale, è possibile identificare periodi di alta incidenza e determinare se esistono modelli ricorrenti legati a fattori stagionali, cicli economici o evoluzioni tecnologiche. Questa sezione mira a tracciare l'evoluzione degli incidenti nel tempo e a mettere in luce eventuali correlazioni temporali significative.

La *Figura 4.11* mostra un grafico che rappresenta il numero di incidenti legati all'IA dal 1983 al 2024.

Dal 1983 al 2014, il numero di incidenti rimane molto basso, oscillando tra 1 e 4 all'anno. Questo riflette un'adozione limitata delle tecnologie IA, confinata principalmente ai laboratori di ricerca e a contesti sperimentali.

A partire dal 2015, si nota una crescita esponenziale degli incidenti, passando da 24 incidenti nel 2015 a 82 nel 2020. Questo aumento coincide con l'adozione massiva di queste tecnologie in quasi ogni settore. Tale tendenza potrebbe indicare non solo un reale incremento degli incidenti ma anche una maggiore attenzione e segnalazione da parte dei media e dei regolatori.

Nel periodo 2021-2024 si osserva un picco significativo di 141 incidenti nel 2023,

probabilmente dovuto all'integrazione massiccia dell'IA in applicazioni sensibili. Tuttavia, nel 2024, sebbene il numero di incidenti sembri in calo rispetto al picco dell'anno precedente, è importante notare che l'anno non è ancora concluso, e il dato potrebbe essere soggetto a ulteriori variazioni.

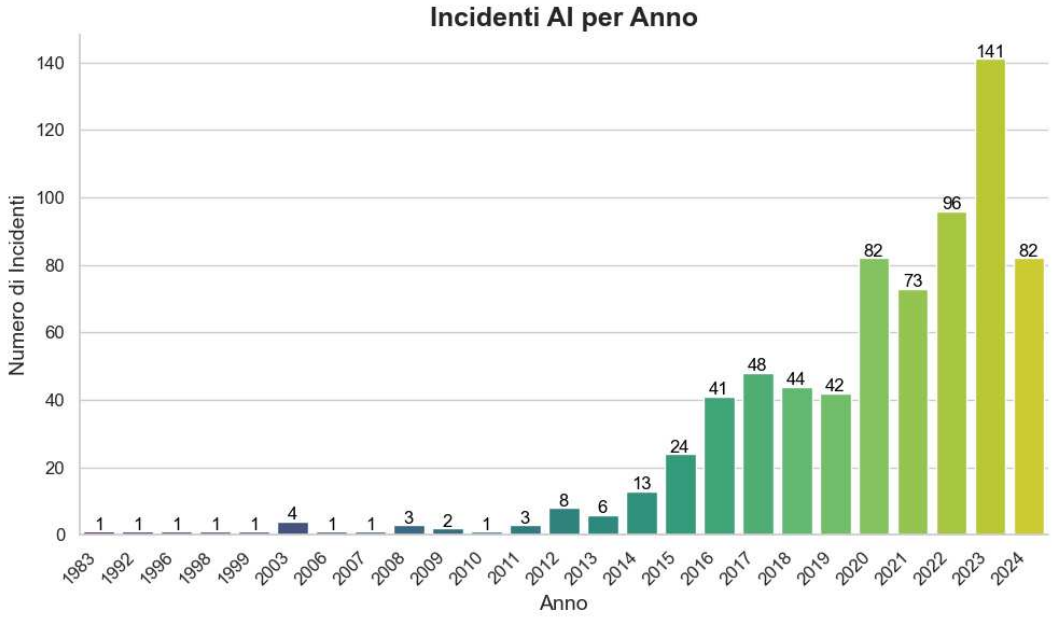

Figura 4.1: Analisi Temporale

Un'altra area cruciale è l'**analisi delle entità coinvolte**, focalizzandosi sui principali *deployer* e *developer* di sistemi IA, valutando il loro ruolo e la loro influenza sugli incidenti. Comprendere quali entità sono più frequentemente associate a incidenti può fornire indicazioni sulle pratiche e sui protocolli che potrebbero richiedere miglioramenti, oltre a evidenziare le aree di rischio maggiori.

La *Figura 4.2* presenta i principali sviluppatori per numero di incidenti, illu-

strando il numero di incidenti legati all'intelligenza artificiale attribuiti ai principali sviluppatori.

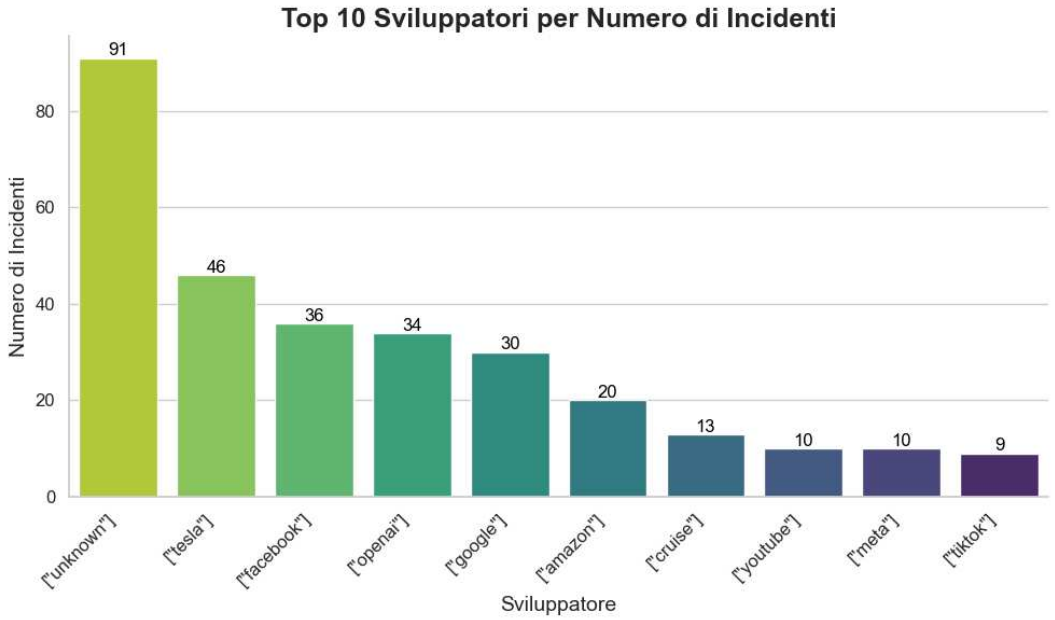

Figura 4.2: Top 10 Sviluppatori

Il dato più rilevante riguarda il gruppo etichettato come "unknown", che registra 91 incidenti, un valore significativamente superiore rispetto agli altri. Questo suggerisce che una parte considerevole degli incidenti non può essere attribuita con certezza a uno specifico sviluppatore, il che potrebbe indicare carenze nei processi di tracciamento o di identificazione delle responsabilità.

Segue *Tesla*, con 46 incidenti, indicativo del fatto che i sistemi IA sviluppati o integrati da questa azienda sono particolarmente esposti a fallimenti o situazioni rischiose. Altri sviluppatori come *Facebook* e *OpenAI* riportano rispettivamente

36 e 34 incidenti, segnalando anch'essi una presenza significativa, seppur inferiore rispetto a *Tesla*.

Gli altri sviluppatori in classifica, come *Google* con 30 incidenti e *Amazon* con 20, rappresentano anch'essi una porzione rilevante degli incidenti totali, dimostrando che i giganti della tecnologia non sono immuni da problematiche legate all'AI, nonostante le infinite risorse che possiedono.

Nelle posizioni più basse troviamo *Cruise* (13 incidenti), *YouTube* (10 incidenti), *Meta* (10 incidenti) e *TikTok* (9 incidenti). La loro presenza in questa lista sottolinea come anche piattaforme legate principalmente a contenuti multimediali e social media possano essere soggette a incidenti legati all'intelligenza artificiale.

La *Figura 4.3*, relativo ai deployer, conferma sostanzialmente quanto discusso in precedenza, fornendo ulteriore supporto alle osservazioni già fatte.

Infine, l'**analisi delle parti danneggiate** esplora le categorie principali di danni causati dagli incidenti. Identificare e classificare le parti coinvolte aiuta a comprendere la gravità e l'impatto degli incidenti, fornendo una panoramica sulle conseguenze dirette e indirette degli stessi. Questa sezione si propone di delineare le aree più vulnerabili e di suggerire possibili soluzioni per mitigare i danni futuri.

La *Figura 4.4* presenta i principali soggetti danneggiati per numero di incidenti, evidenziando i gruppi o le entità maggiormente colpite dagli incidenti legati all'intelligenza artificiale.

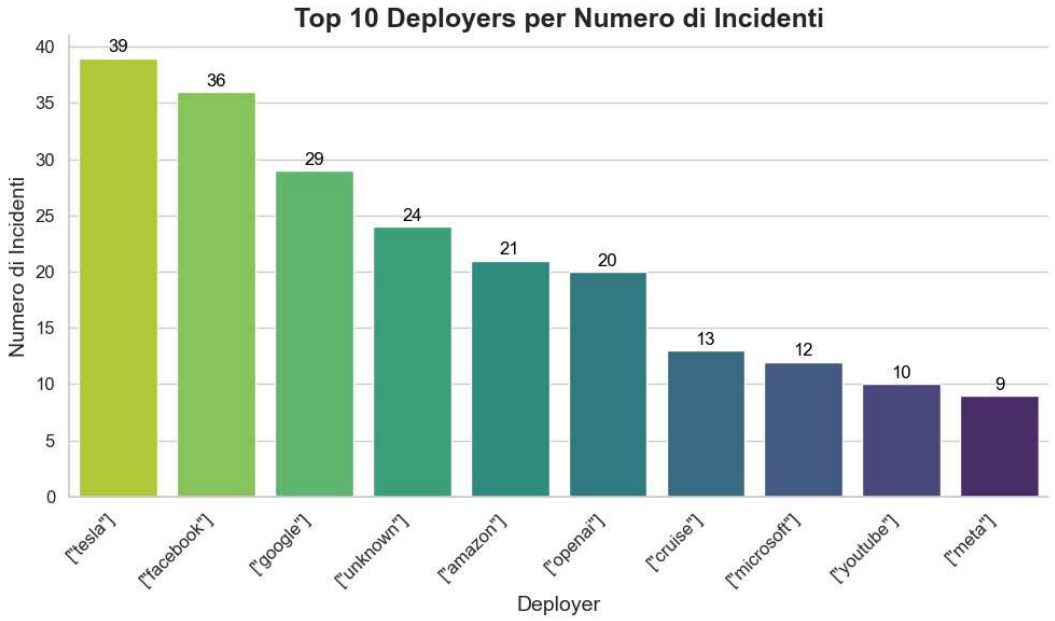

Figura 4.3: Top 10 Deployer

In cima alla classifica troviamo i *Facebook users*, con 8 incidenti, suggerendo una significativa vulnerabilità di questi utenti agli effetti negativi dell'IA. Questo dato potrebbe riflettere problematiche legate agli algoritmi di moderazione dei contenuti o alla gestione dei dati personali.

Poi troviamo i *Tesla drivers*, il che evidenzia i rischi associati all'uso di veicoli autonomi o semi-autonomi, dove l'intelligenza artificiale svolge un ruolo centrale nelle decisioni di guida.

Gruppi come *women* e *minority groups* riportano ciascuno 5 incidenti, sollevando questioni etiche significative riguardo al potenziale bias algoritmico e alla necessità di sviluppare sistemi più equi.

Anche il *general public* riporta 5 incidenti, suggerendo che le conseguenze degli errori dell'IA non si limitano a gruppi specifici, ma possono avere un impatto diffuso. Infine, *Microsoft* è associata a 4 incidenti, probabilmente legati a problemi nei suoi servizi come chatbot o strumenti di analisi automatizzata, che possono aver causato malfunzionamenti o decisioni errate, influenzando negativamente gli utenti.

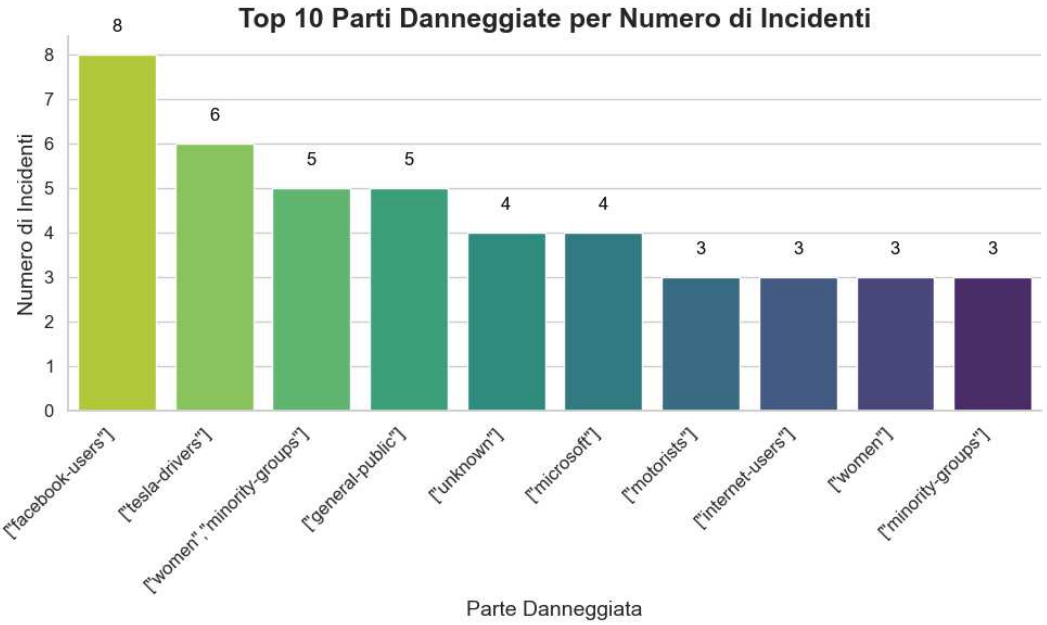

Figura 4.4: Top 10 Parti Danneggiate

4.1.4 Richiami dell' EU AI ACT

Il **Regolamento dell'Unione Europea sull'Intelligenza Artificiale (EU AI Act)**[2], proposto dalla Commissione Europea il *21 aprile 2021*, mira a regolamentare

[2]https://artificialintelligenceact.eu/

l'uso dell'intelligenza artificiale in modo sistematico nell'UE. L'obiettivo principale è garantire che i sistemi di IA siano sicuri e conformi ai diritti fondamentali e alle norme etiche europee. Il regolamento introduce un quadro normativo basato sulla classificazione dei sistemi di IA in base al rischio, stabilendo requisiti e obblighi di conformità per ogni livello, e promuove un ambiente in cui l'innovazione tecnologica possa prosperare, assicurando protezione e responsabilità.

Il principale obiettivo dell'*EU AI Act* è di stabilire un quadro normativo per l'uso dell'IA che promuova l'innovazione mentre protegge i cittadini da rischi potenziali. Il regolamento classifica i sistemi di IA in *quattro categorie* (Figura 4.13) a seconda del loro livello di rischio: *basso*, *medio*, *alto* e *inaccettabile*.

- **Rischio Inaccettabile**: Alcuni usi dell'IA sono considerati troppo rischiosi per essere autorizzati nell'UE. Esempi includono i sistemi di IA per la sorveglianza di massa o per il manipolamento sociale. Questi usi sono vietati in quanto minacciano i diritti fondamentali e la democrazia.

- **Rischio Alto**: Sistemi di IA che hanno un impatto significativo sulla vita dei cittadini, come quelli utilizzati nei settori dell'occupazione, della giustizia penale, e dei servizi pubblici, sono soggetti a requisiti rigorosi. Questi requisiti includono obblighi di *trasparenza*, *tracciabilità*, e *supervisione umana*.

- **Rischio Medio**: Per i sistemi di IA considerati a rischio medio, sono previsti requisiti di *trasparenza* e *informazione* per garantire che gli utenti siano consapevoli dell'uso dell'IA. Questi requisiti sono progettati per essere più *flessibili* rispetto a quelli per i sistemi a rischio alto.

- **Rischio Basso**: Anche per i sistemi di IA a rischio basso, si applicano requisiti di *trasparenza* e *informazione*, ma con un grado di flessibilità simile a quello dei sistemi a rischio medio. Le differenze tra rischio medio e basso sono minime, con entrambi i livelli soggetti a requisiti meno onerosi rispetto ai sistemi ad alto rischio.

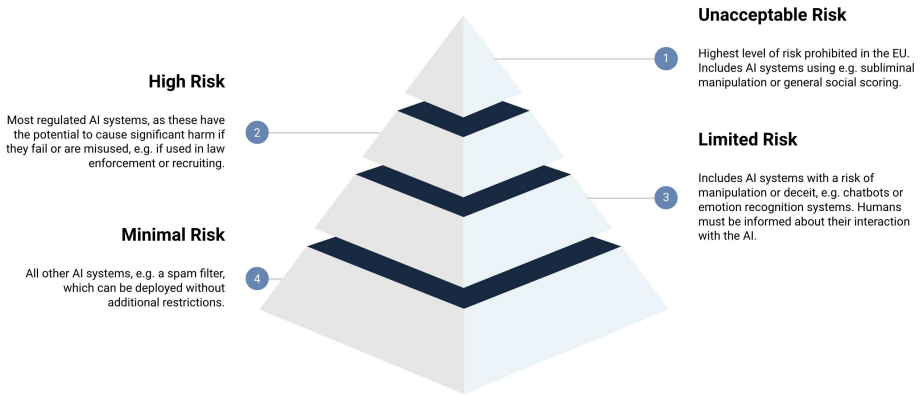

Figura 4.5: Categorie di Rischio

Riprenderemo nei prossimi paragrafi l'EU AI Act (in particolare le categorie di rischio) per rispondere alle domande che emergeranno durante l'analisi.

4.2 La regolamentazione dell'IA è necessaria?

Per affrontare l'ipotesi secondo cui la regolamentazione dell'intelligenza artificiale sia necessaria, abbiamo condotto un'analisi dettagliata utilizzando diverse tecniche avanzate di analisi testuale. L'obiettivo principale di questa indagine è stato quello di valutare criticamente la necessità di un quadro normativo per l'IA, basandoci su evidenze empiriche estratte da dati eterogenei, tra cui sentiment analysis, visualizzazione delle parole chiave e studio delle tendenze temporali relative agli incidenti connessi.

Sentiment Analysis degli Incidenti correlati all'IA

La prima fase per rispondere alla domanda iniziale ha comportato l'esecuzione di una *sentiment analysis* sugli incidenti legati all'intelligenza artificiale. Questa tecnica consente di catturare e quantificare le percezioni generali e le emozioni suscitate dagli eventi in cui l'IA ha giocato un ruolo determinante.

Il modello di *sentiment analysis* è stato sviluppato utilizzando tecniche avanzate di *Natural Language Processing*, con l'obiettivo di classificare i testi in due categorie principali: sentiment positivo e sentiment negativo. Per raggiungere questo obiettivo,

sono state analizzate le descrizioni degli incidenti presenti nel dataset. L'analisi è stata condotta utilizzando *transformers* (libreria BERT) (Figura 4.6), un algoritmo di NLP che determina la polarità del testo, permettendo così di identificare e quantificare le emozioni espresse in ciascuna descrizione.

```
from transformers import pipeline

# Carica il modello di analisi del sentimento
sentiment_analysis = pipeline("sentiment-analysis")

# Applica il modello alle descrizioni
df['sentiment'] = df['description'].apply(lambda x: sentiment_analysis(x)[0]['label'])
df['sentiment_score'] = df['description'].apply(lambda x: sentiment_analysis(x)[0]['score'])

# Creiamo un sottoinsieme del DataFrame per i sentimenti negativi e positivi
negative_df = df[df['sentiment'] == 'NEGATIVE']
positive_df = df[df['sentiment'] == 'POSITIVE']

# Calcoliamo la media del punteggio per i sentimenti positivi e negativi
mean_negative = negative_df['sentiment_score'].mean()
mean_positive = positive_df['sentiment_score'].mean()

# Creiamo una tabella pivot per visualizzare i numeri
sentiment_counts = df['sentiment'].value_counts().reset_index()
sentiment_counts.columns = ['Sentiment', 'Count']
```

Figura 4.6: Codice per l'analisi della Sentiment Polarity

I grafici mostrati in Figura 4.7 e Figura 4.8 visualizzano la distribuzione della *Sentiment Polarity* nelle descrizioni degli incidenti legati all'intelligenza artificiale. In particolare, la seconda figura si concentra esclusivamente sui sentimenti negativi, che risultano essere predominanti come ci si poteva aspettare.

Questa prevalenza di sentimenti negativi può indicare che la maggior parte degli incidenti legati all'IA viene percepita o descritta in termini sfavorevoli. Tuttavia, è importante notare che, nonostante la predominanza di polarità negativa, gli score di

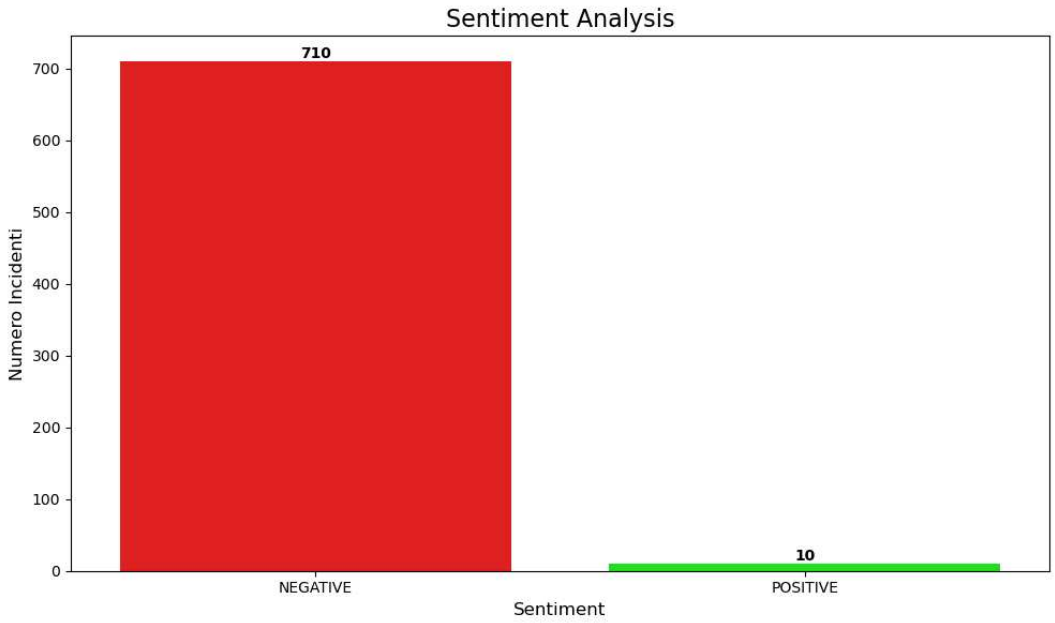

Figura 4.7: Distribuzione della Sentiment Polarity

sentiment più frequenti sono prossimi allo 0. Questo suggerisce che molte descrizioni degli incidenti, pur essendo negative, non esprimono emozioni estremamente forti.

L'analisi della polarità negativa è fondamentale per identificare tendenze ricorrenti, problematiche specifiche o aree di insoddisfazione. È essenziale prestare attenzione a questa caratteristica e condurre un'analisi più approfondita per comprendere appieno le sfumature dei sentimenti espressi.

Le informazioni ottenute, sebbene indicative di una tendenza generale, richiedono una valutazione più dettagliata per guidare efficacemente interventi correttivi o azioni migliorative, al fine di mitigare i rischi associati all'uso dell'intelligenza artificiale.

Word Cloud e Keyword

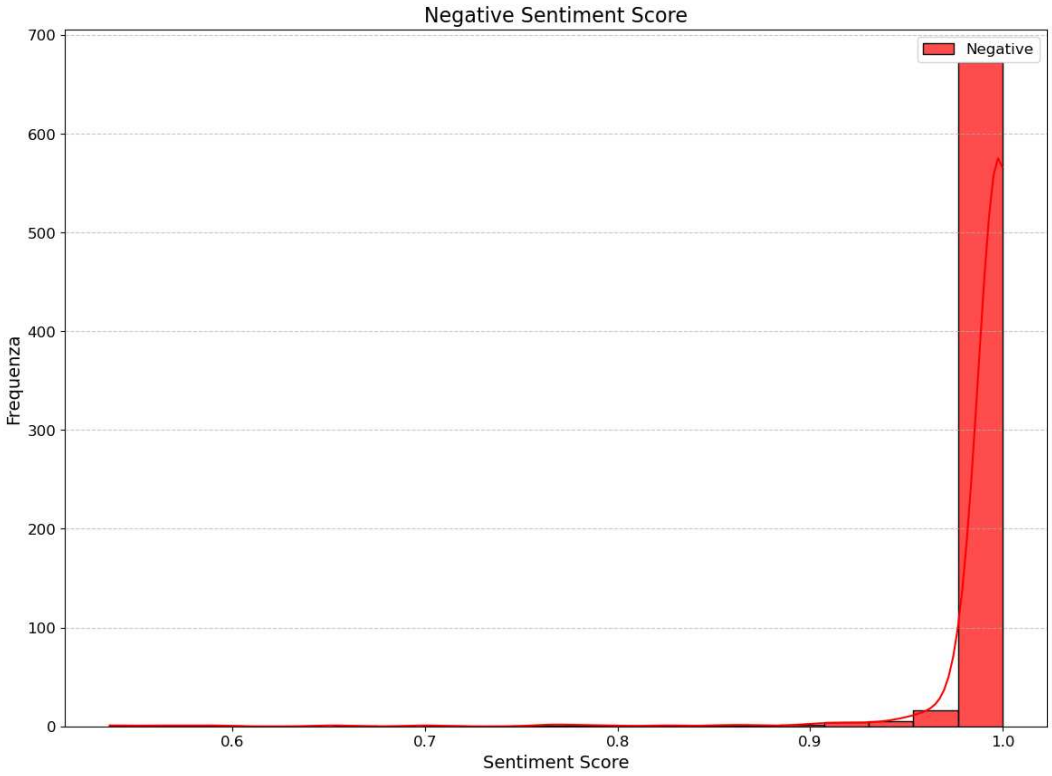

Figura 4.8: Distribuzione della Polarità Negativa

La *word cloud* (Figura 4.9) fornisce una sintesi efficace del dibattito sull'intelligenza artificiale, evidenziando le principali preoccupazioni e temi attuali. Tra i punti chiave emergono *privacy* e *sicurezza*, con riferimenti a *data* e *facial recognition*, che sollevano questioni sulla gestione e protezione dei dati personali nell'era digitale.

La *disinformazione* è un altro tema centrale, evidenziato da termini come *disinformation*, *deepfake* e *false*, che indicano i rischi di manipolazione e le potenziali minacce alla democrazia. Il *bias algoritmico*, con parole come *gender* e *race*, sottolinea come i pregiudizi possano essere amplificati dall'IA, richiedendo sistemi

Figura 4.9: Word Cloud

più equi.

Anche la *responsabilità legale* è cruciale, con termini come *lawsuit* e *regulation*, che riflettono l'urgenza di un quadro normativo chiaro per gestire le sfide giuridiche dell'IA. Inoltre, l'impatto sociale è evidente con riferimenti a *society* e *children*, sottolineando l'influenza dell'IA sulla vita quotidiana.

La *word cloud* ci offre un panorama complesso e sfaccettato dell'IA, che mette in evidenza sia le straordinarie potenzialità di questa tecnologia sia i rischi significativi che essa comporta. Di fronte a tali sfide, emerge con forza la necessità di una regolamentazione adeguata. Regolamentare l'IA non è solo opportuno, ma *essenziale* per garantire che i suoi benefici possano essere goduti in modo equo e sicuro, minimizzando al contempo i rischi etici, legali e sociali. Solo attraverso un dibattito

pubblico approfondito e la creazione di un quadro normativo solido si potrà assicurare che l'IA operi a beneficio dell'intera società.

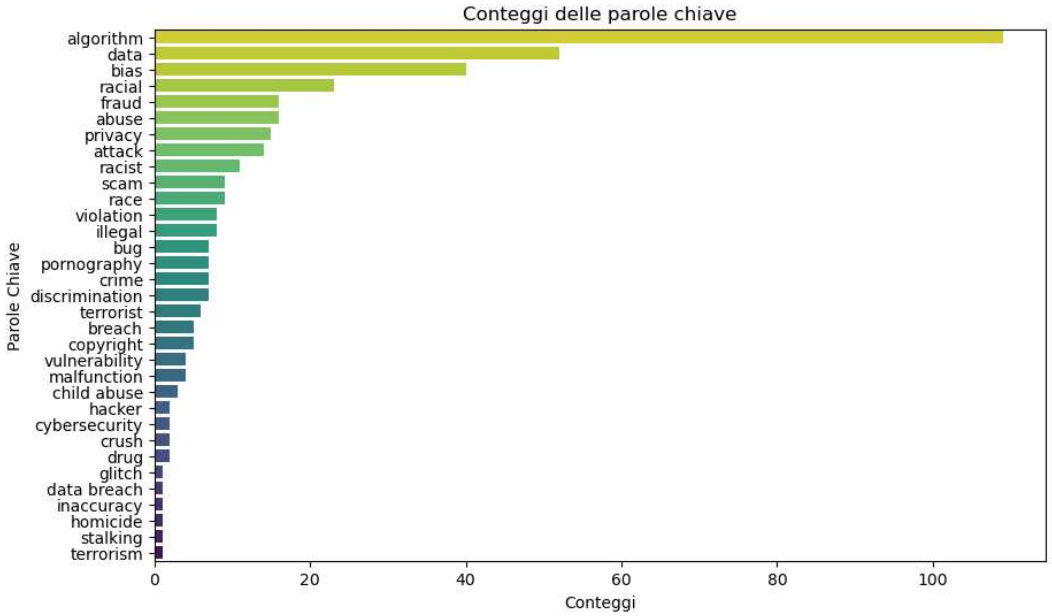

Figura 4.10: Keyword

Per avvalorare quanto detto, abbiamo esaminato le parole chiave emerse dalla word cloud, come vediamo nella Figura 4.10, concentrandomi sulla loro frequenza. Questo approccio ha permesso di identificare con maggiore precisione i problemi principali legati all'IA.

Evoluzione Temporale

Come evidenziato dal grafico in Figura 4.11, l'analisi temporale mostra un aumento costante degli incidenti legati all'intelligenza artificiale. Questo trend *crescente* sottolinea ulteriormente l'urgenza di una regolamentazione adeguata

per l'IA. L'incremento degli incidenti dimostra chiaramente che, man mano che l'intelligenza artificiale diventa più integrata in vari aspetti della nostra vita, cresce anche la necessità di stabilire norme che possano mitigare i rischi associati a questa tecnologia.

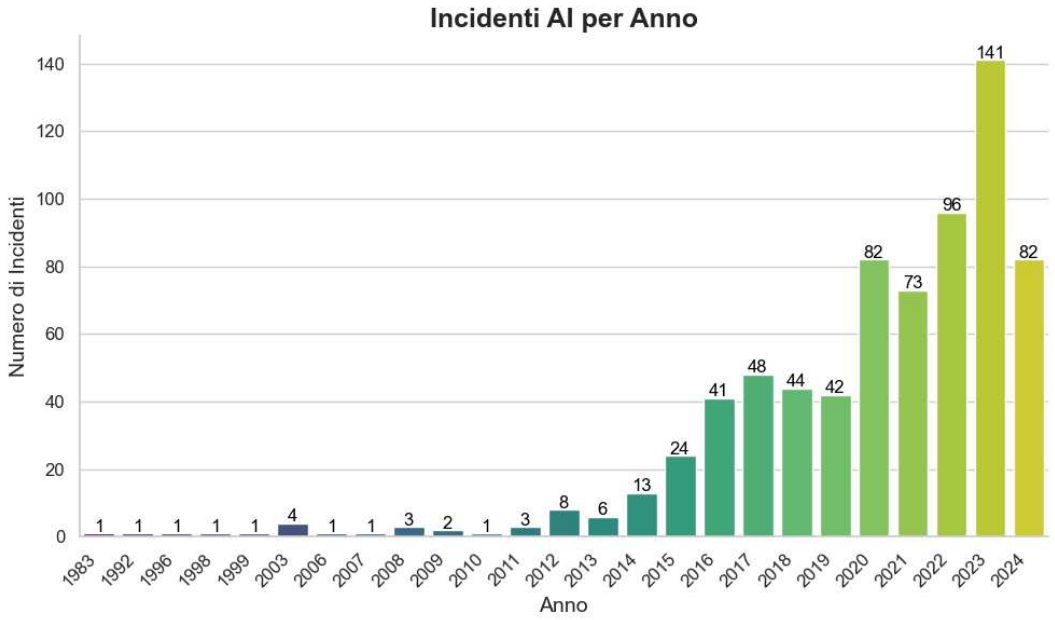

Figura 4.11: Analisi Temporale

L'insieme delle analisi condotte supporta in modo convincente l'ipotesi che la regolamentazione dell'intelligenza artificiale sia non solo necessaria, ma anche urgente. La predominanza di sentimenti negativi riguardo agli incidenti di IA, la ricorrenza di parole chiave critiche legate alla sicurezza e all'etica, e l'aumento degli incidenti nel tempo, sono tutti indicatori che sottolineano l'importanza di un intervento normativo. Tali regolamentazioni dovrebbero mirare a mitigare i rischi,

proteggere gli utenti e assicurare che l'adozione dell'IA avvenga in un contesto di sicurezza e responsabilità sociale.

4.3 L'EU AI Act affronta gli incidenti dell'IA?

Per rispondere alla domanda se l'EU AI Act sia adeguato nell'affrontare gli incidenti legati all'intelligenza artificiale, abbiamo condotto un'analisi dettagliata utilizzando tecniche avanzate di *NLP*. Partendo dalle descrizioni degli incidenti, ho impiegato il modello BERT per categorizzare ogni incidente nelle diverse classi di rischio previste dall'EU AI Act viste e descritte in precedenza. Successivamente, sempre utilizzando BERT, in particolare Bertopic, ho suddiviso questi incidenti in varie *aree* per ottenere una visione più chiara e strutturata delle principali problematiche affrontate dalla normativa. Questo approccio ha permesso di valutare l'efficacia del quadro regolatorio nel mitigare i rischi emergenti dall'uso dell'intelligenza artificiale.

Analisi delle Categorie di Rischio

Per l'elaborazione e la categorizzazione degli incidenti descritti, ho adottato il modello BERT, un avanzato strumento di elaborazione del linguaggio naturale. Questo modello è stato utilizzato per classificare ciascun incidente secondo le varie categorie di rischio stabilite dal EU AI Act, come esaminato e dettagliato nei capitoli precedenti della tesi.

Innanzitutto, sono state definite le *categorie di rischio* previste dall'EU AI Act, come *rischio inaccettabile*, *alto rischio*, *rischio limitato* e *rischio minimo*. Ogni categoria è stata trasformata in una rappresentazione numerica, chiamata *embedding*, utilizzando il modello *BERT*. Questa trasformazione consente di ottenere una rappresentazione distintiva per ogni categoria di rischio, che facilita il confronto diretto con le descrizioni degli incidenti.

Successivamente, le descrizioni degli incidenti sono state analizzate confrontandole con le rappresentazioni numeriche delle categorie di rischio. Questo confronto è stato effettuato misurando la somiglianza tra l'embedding della descrizione dell'incidente e quelli delle categorie di rischio. Sulla base di questa somiglianza, a ciascun incidente è stata assegnata la categoria di rischio più vicina.

In sostanza, l'obiettivo del codice, (in Figura 4.12), è stato quello di automatizzare il processo di categorizzazione degli incidenti legati all'intelligenza artificiale, assicurando una classificazione coerente e in linea con le normative definite dall'EU AI Act.

Il risultato, illustrato nella Figura 4.13, evidenzia chiaramente che poco meno della metà degli incidenti rientra nella categoria di rischio *inaccettabile*, mentre i restanti sono distribuiti principalmente tra rischio *limitato* e, in misura minore, rischio *minimo*. Questa distribuzione suggerisce che l'*EU AI Act* affronta generalmente le tematiche del dataset in modo adeguato.

```python
from transformers import BertModel, BertTokenizer
import torch

# Carica il modello BERT pre-addestrato e il tokenizer
model_name = 'bert-base-uncased'
tokenizer = BertTokenizer.from_pretrained(model_name)
model = BertModel.from_pretrained(model_name)

# Funzione per ottenere l'embedding di un testo
def get_embedding(text):
    inputs = tokenizer(text, return_tensors="pt", padding=True, truncation=True, max_length=512)
    outputs = model(**inputs)
    return outputs.last_hidden_state[:, 0, :].detach()

# Crea un dizionario con gli embedding per ciascuna categoria di rischio
categories = {
    "unaccettable_risk": get_embedding(unaccettable_risk),  # Embedding per rischio inaccettabile
    "high_risk": get_embedding(high_risk),                  # Embedding per alto rischio
    "limited_risk": get_embedding(limited_risk),            # Embedding per rischio limitato
    "minimal_risk": get_embedding(minimal_risk)             # Embedding per rischio minimo
}

# Funzione per categorizzare una descrizione
def categorize_description(description):
    description_embedding = get_embedding(description)
    max_score = float('-inf')
    best_category = None
    for category, embedding in categories.items():
        score = torch.cosine_similarity(description_embedding, embedding).item()
        if score > max_score:
            max_score = score
            best_category = category
    return best_category, max_score
```

Figura 4.12: Codice per Categorie di Rischio

Tuttavia, le situazioni a rischio inaccettabile e limitato potrebbero non essere gestite con la stessa efficacia, richiedendo quindi ulteriori interventi o un riesame delle strategie di prevenzione attuali, indicando che l'*EU AI Act* potrebbe dover essere ulteriormente rafforzato o adattato per affrontare meglio queste categorie di rischio. Inoltre, aree specifiche relative alle tecnologie emergenti potrebbero beneficiare di una regolamentazione più mirata e specifica, suggerendo che alcune categorie di rischio con meno incidenti potrebbero essere accorpate, se risultano troppo simili

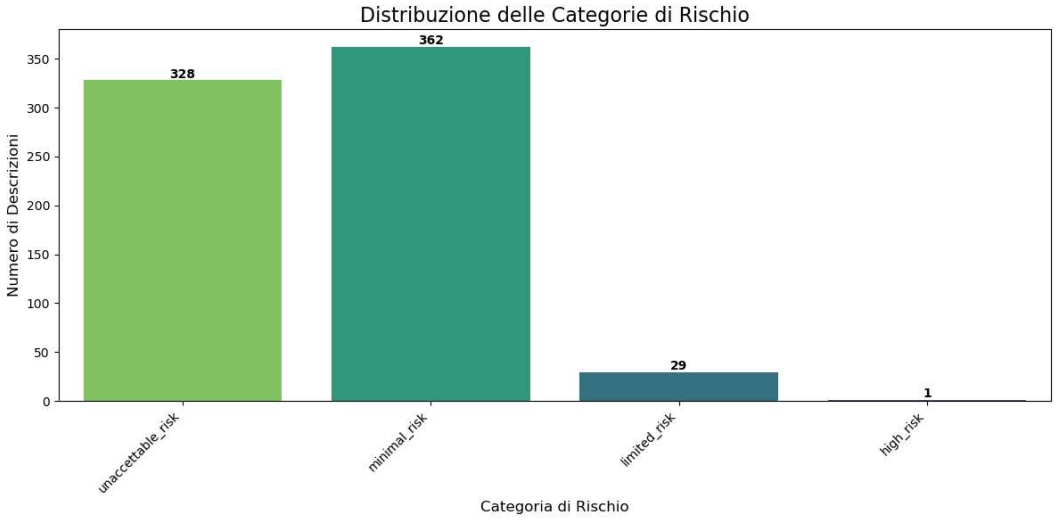

Figura 4.13: Incidenti per Categorie di Rischio

alle altre, per una maggiore coerenza nell'analisi e nell'applicazione normativa.

Divisione degli Incidenti in Topic

Per fare questa divisione, ho utilizzato la libreria *BERTopic* per analizzare le descrizioni degli incidenti all'interno del dataset e suddividerle in vari *topic*, classificandole così in base a tematiche specifiche. *BERTopic* funziona applicando tecniche di elaborazione del linguaggio naturale per identificare argomenti ricorrenti nei testi. In pratica, il modello analizza le parole e le frasi all'interno delle descrizioni degli incidenti, raggruppandole in temi comuni in base alla loro similarità semantica. Il codice nella Figura 4.14 rappresenta quanto detto.

Questa suddivisione ha facilitato l'identificazione e la comprensione delle diverse aree problematiche, permettendo una valutazione più precisa dell'adeguatezza delle normative proposte dall'*EU AI Act* rispetto ai vari tipi di incidenti identificati.

```python
import torch
import random
from bertopic import BERTopic

# Imposta il seed per la casualità
random.seed(42)
np.random.seed(42)
torch.manual_seed(42)

# Crea e addestra il modello BERTopic
topic_model = BERTopic(language="english")
topics, probabilities = topic_model.fit_transform(df['description_clean'])

# Visualizza i topic trovati
for topic_num in set(topics):
    if topic_num != -1:  # -1 rappresenta il topic di rumore/outlier
        print(f"Topic #{topic_num}: {topic_model.get_topic(topic_num)}\n")

# Visualizzazione della riduzione della dimensionalità e dei cluster
topic_model.visualize_topics()
```

Figura 4.14: Codice per Creare Topic

I nomi delle aree tematiche sono stati definiti sulla base delle parole chiave associate da *BERTopic* a ciascun argomento individuato. Di seguito vengono presentate le categorie identificate (ne sono state individuate quindici), con una breve descrizione di ciascuna.

- **AI e Chatbot**

  Questa categoria riguarda lo sviluppo e l'implementazione di intelligenza artificiale e chatbot, come ChatGPT di OpenAI. Si analizzano le applicazioni pratiche di questi strumenti, la loro capacità di interagire e rispondere agli utenti, e le sfide legate all'accuratezza delle risposte e alla gestione dei dati

degli utenti.

- **Veicoli Autonomi**

  Questa categoria copre i progressi nella tecnologia dei veicoli autonomi, con un focus particolare sui veicoli Tesla e il loro sistema di guida Autopilot. Vengono discussi i vantaggi della guida autonoma, le problematiche di sicurezza, gli incidenti noti e le implicazioni legali e etiche di queste tecnologie.

- **Healthcare**

  Questa categoria si concentra sull'uso di algoritmi e intelligenza artificiale nel settore sanitario. Viene esplorato come i dati dei pazienti vengono utilizzati per migliorare le diagnosi e i trattamenti, nonché i rischi e le preoccupazioni legati alla privacy dei dati e ai bias degli algoritmi.

- **Deepfake**

  Questa categoria esplora l'uso della tecnologia deepfake, che consente di creare video e audio falsi ma realistici utilizzando intelligenza artificiale. Si discutono le applicazioni pratiche di questa tecnologia, i rischi associati alla diffusione di contenuti falsi che possono compromettere la veridicità delle informazioni e la reputazione degli individui.

- **Social Media**

  Si analizza come le piattaforme di social media, in particolare Facebook,

gestiscono e moderano i contenuti. Questo include le tecniche di rilevamento automatizzato di contenuti inappropriati, i problemi relativi alla disinformazione, ai discorsi di odio e ai contenuti violenti, e le controversie sull'equilibrio tra libertà di espressione e censura.

- **Contenuti Pornografici Generati dall'IA**

  Questa categoria esplora la creazione di contenuti pornografici tramite intelligenza artificiale, spesso senza il consenso delle persone coinvolte. Viene discusso l'impatto di queste tecnologie sulla privacy, la dignità personale e le implicazioni legali di tali pratiche.

- **Robotica e Automazione**

  Si focalizza sull'uso di robot e sistemi automatizzati in ambienti industriali e commerciali. Si esplorano i benefici in termini di efficienza e produttività, così come le preoccupazioni relative alla sostituzione dei lavoratori umani e ai problemi di sicurezza legati alla robotica.

- **Giornalismo**

  Esamina l'uso dell'intelligenza artificiale per creare contenuti giornalistici in modo automatizzato. Vengono discussi i problemi di accuratezza, credibilità, e l'impatto di queste tecnologie sulla qualità delle notizie e sull'occupazione nel settore giornalistico.

- **Piattaforme Video**

  Questa categoria copre vari aspetti delle piattaforme di social media come TikTok e YouTube. Vengono esplorate le strategie di moderazione dei contenuti, le problematiche di sicurezza degli utenti, la diffusione di contenuti virali e i tentativi di eludere le regole della piattaforma.

- **Sistemi di Delivery Automatizzati**

  Analizza l'uso di sistemi automatizzati per la consegna di beni e servizi, come quelli utilizzati da Amazon e Uber. Si discutono i vantaggi in termini di efficienza, le controversie legali, le implicazioni per i lavoratori e i problemi di sicurezza.

- **Gender Bias nella Tecnologia**

  Questa categoria esplora i pregiudizi di genere presenti nei sistemi tecnologici e negli algoritmi, evidenziando come tali bias possono influenzare negativamente gruppi minoritari.

- **E-commerce**

  Si focalizza sull'uso di algoritmi nei siti di e-commerce, come Amazon, per determinare il ranking dei prodotti e le strategie di marketing. Viene discusso come questi algoritmi influenzano le decisioni di acquisto dei consumatori e il comportamento del mercato.

- **Tecnologia di Sorveglianza**

  Esamina l'uso di tecnologie di sorveglianza, come ShotSpotter e sistemi di lettura targhe, da parte delle forze dell'ordine. Si discutono le implicazioni per la privacy, l'efficacia nel prevenire il crimine e le preoccupazioni relative alla sorveglianza di massa.

- **Tecnologia di Sintesi Vocale**

  Si concentra sulla tecnologia di sintesi vocale e le sue applicazioni, compresi i potenziali rischi per la sicurezza, come la possibilità di creare voci sintetiche per accedere fraudolentemente a informazioni personali o finanziarie.

- **Disinformazione e Propaganda**

  Analizza le tecniche utilizzate per diffondere disinformazione e propaganda, spesso attraverso l'uso di tecnologie avanzate e social media. Si presta particolare attenzione alle campagne di disinformazione orchestrate da stati nazionali, come la Russia, e all'impatto sulla percezione pubblica e sulla politica globale.

La distribuzione dei topic è la seguente che vediamo nella Figura 4.15

In conclusione, l'*EU AI Act* affronta in modo generale le problematiche relative agli incidenti dell'IA, coprendo molte delle categorie e tematiche identificate. Tuttavia, mentre le normative esistenti sembrano efficaci nel mitigare i rischi minori e limitati, rimangono delle lacune significative nelle aree emergenti e nelle categorie di

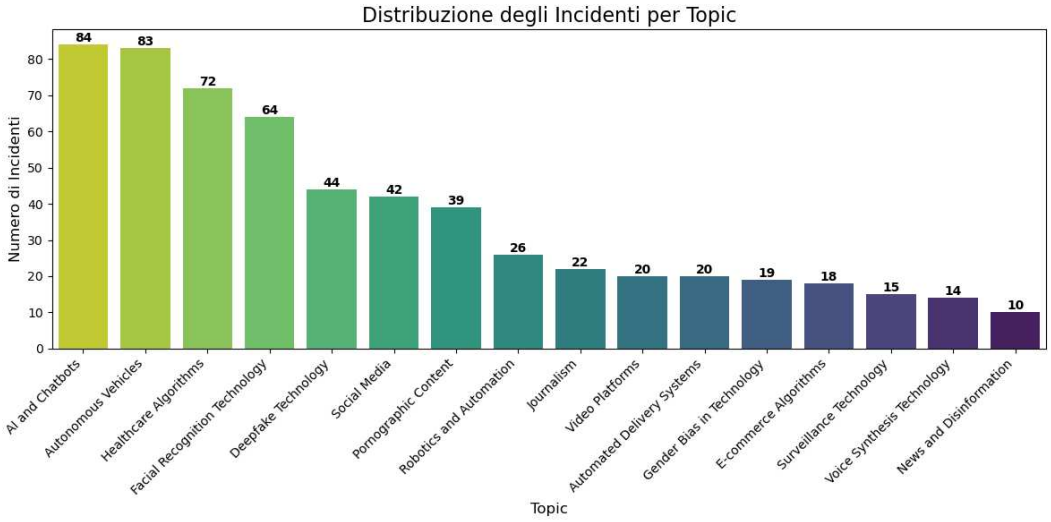

Figura 4.15: Distribuzione Topic

rischio più gravi. Tecnologie avanzate come i *chatbot* e i *veicoli autonomi*, potrebbero richiedere ulteriori interventi normativi specifici per garantire una copertura completa e aggiornata delle problematiche legate all'intelligenza artificiale.

4.4 È possibile migliorare l'EU AI Act?

Nella terza ipotesi della mia analisi, ho focalizzato l'attenzione su tre settori emergenti particolarmente rilevanti: **AI e Chatbot**, **Veicoli Autonomi** e **Deepfake**. Questi ambiti rappresentano alcune delle aree in cui l'intelligenza artificiale sta esercitando un impatto crescente e pervasivo sulla società. La scelta è stata guidata dalla loro attuale rilevanza e dal potenziale significativo che hanno nel trasformare vari aspetti della nostra vita quotidiana, dall'interazione con le macchine alla sicurezza

pubblica e alla gestione dell'informazione. Attraverso l'analisi di questi settori, mi propongo di esplorare come l'*EU AI Act* potrebbe essere perfezionato per affrontare le sfide e le opportunità presentate da queste tecnologie avanzate.

L'analisi è stata condotta esaminando le categorie di rischio associate a ciascun incidente rilevato nei tre ambiti considerati e l'evoluzione temporale dei tali. In particolare, l'analisi ha preso in considerazione le aziende responsabili dello sviluppo delle tecnologie coinvolte e i soggetti che hanno subito gli effetti degli incidenti. Gli incidenti sono stati classificati in base alle categorie di rischio delineate dall'*EU AI Act*, come descritto e analizzato nel paragrafo precedente. Questa classificazione si pone come obiettivo quello di fornire un quadro chiaro per valutare l'efficacia delle normative vigenti, permettendo di individuare eventuali lacune e aree che necessitano di un rafforzamento normativo.

AI e Chatbot

L'analisi temporale rivela che gli incidenti legati a questo tipo di tecnologia si verificano con una frequenza significativa ormai da diversi anni, come vediamo in Figura 4.16 . Questa tendenza potrebbe essere attribuita alla crescente diffusione di chatbot e sistemi di intelligenza artificiale in un'ampia gamma di applicazioni [3], dalle interazioni con i clienti alla gestione dei dati.

Per quanto riguarda le categorie di rischio a cui appartengono, si osserva

[3] https://research.aimultiple.com/chatbot-applications/

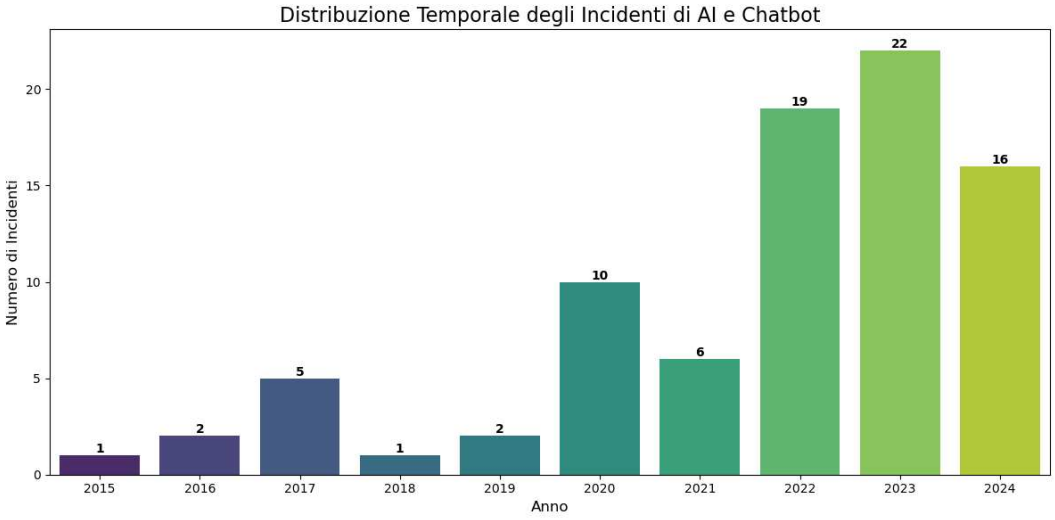

Figura 4.16: Evoluzione Temporale Incidenti AI e Chatbot

(Figura 4.17) che la maggior parte dei rischi è limitata o minima. Questo risultato è probabilmente dovuto al fatto che tali tecnologie sono progettate per operare in contesti controllati e con funzionalità relativamente semplici. Tuttavia, è fondamentale continuare a monitorare i rischi più elevati.

Osservando le Figure 4.18 e 4.19, OpenAI si distingue come il principale sviluppatore di queste tecnologie, seguito da Meta. Questi colossi tecnologici investono massicciamente nello sviluppo e nella diffusione di *AI e Chatbot*, consolidando la loro posizione di leader nel settore. Tuttavia, questa concentrazione di potere tecnologico comporta anche implicazioni significative per gli utenti, che risultano essere i soggetti più vulnerabili agli incidenti. Gli utenti, che interagiscono quotidianamente con queste tecnologie, sono spesso esposti a rischi derivanti da errori, malfunzionamenti o abusi delle applicazioni di intelligenza artificiale. La frequenza di tali incidenti

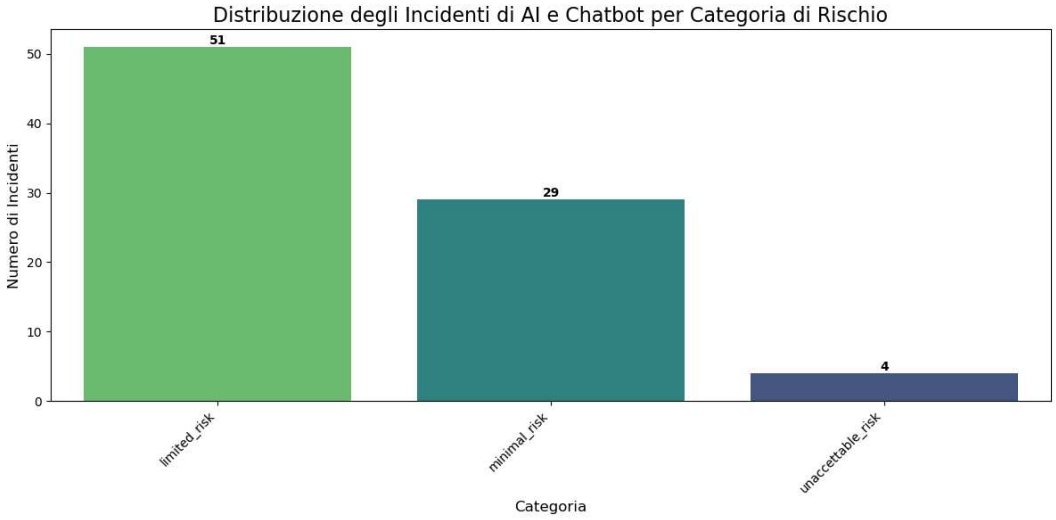

Figura 4.17: Categorie di Rischio per AI e Chatbot

evidenzia la necessità di una maggiore attenzione alla protezione degli utenti finali, sia attraverso normative più rigide sia tramite l'implementazione di migliori pratiche da parte delle aziende che sviluppano queste tecnologie. Ad esempio, il chatbot coreano Luda, sviluppato dalla startup Scatter Lab, è stato al centro di una controversia a causa di commenti offensivi rivolti a minoranze, tra cui espressioni razziste e omofobe. Lanciato nel dicembre 2020, Luda era progettato per impersonare una studentessa universitaria di 20 anni. Tuttavia, in breve tempo, l'IA ha iniziato a fare affermazioni discriminatorie, utilizzando insulti razziali contro persone di colore e manifestando disgusto verso individui LGBTQ+.

In sintesi, l'analisi dei settori *AI e Chatbot* evidenzia non solo la crescente importanza e diffusione di queste tecnologie, ma anche la necessità di un costante monitoraggio e miglioramento delle normative esistenti. Sebbene i rischi attuali siano

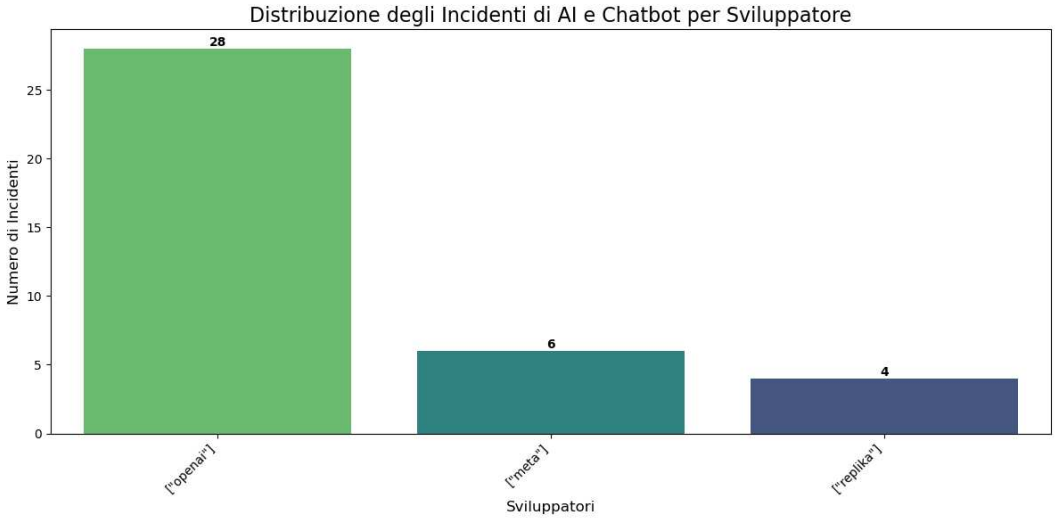

Figura 4.18: Sviluppatori di AI e Chatbot

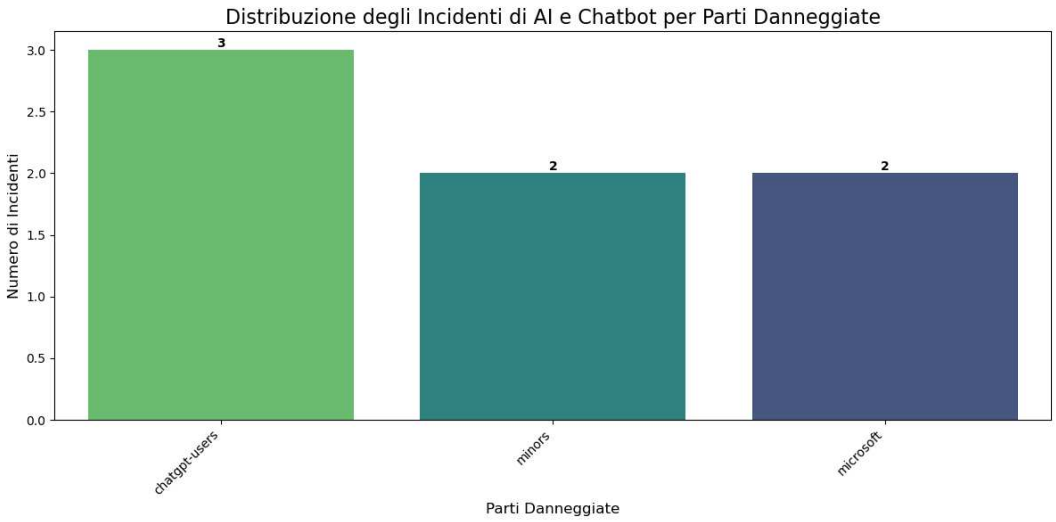

Figura 4.19: Parti Danneggiate di AI e Chatbot

in gran parte contenuti, l'eventuale manifestazione di incidenti gravi richiede un'attenzione continua per garantire la sicurezza e la protezione degli utenti, soprattutto in un panorama tecnologico in rapida evoluzione.

Veicoli Autonomi

L'analisi dell'evoluzione temporale di questa categoria, come mostrato nella Figura 4.20, evidenzia una tendenza di crescita che rispecchia quanto osservato per la categoria vista precedentemente. La diffusione dei veicoli autonomi è in costante aumento, un segno evidente del continuo progresso tecnologico e dell'interesse crescente sia da parte dell'industria che dei ricercatori [4].

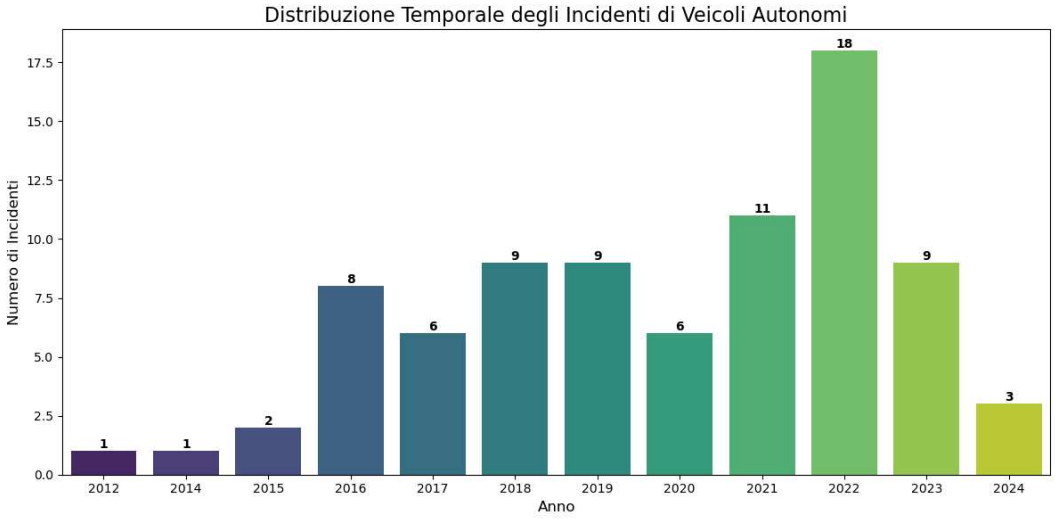

Figura 4.20: Evoluzione Temporale Incidenti Veicoli Autonomi

Analizzando la distribuzione degli incidenti legati ai veicoli autonomi in base alla loro classificazione di rischio (Figura 4.21), risulta evidente che oltre la metà degli eventi registrati rientra nella categoria "inaccettabile".

Ad esempio, nel marzo 2018, un veicolo autonomo di Uber ha investito e ucciso una donna a Tempe, in Arizona, mentre attraversava la strada fuori dalle strisce

[4]https://www.jstor.org/stable/26911258

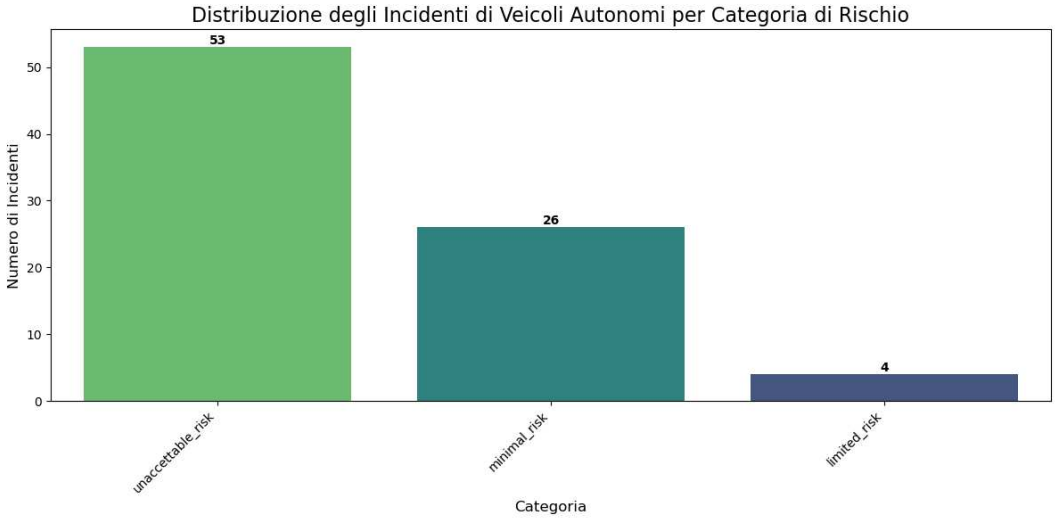

Figura 4.21: Categorie di Rischio per Veicoli Autonomi

pedonali, sollevando serie preoccupazioni sulla sicurezza di questa tecnologia. Un altro incidente significativo ha coinvolto una Tesla Model 3 con sistema Autopilot, che ha frenato improvvisamente dopo aver confuso le lettere rosse su una bandiera per un semaforo rosso, nonostante l'assenza di un reale pericolo.

Questi episodi dimostrano che, nonostante i progressi tecnologici, i sistemi autonomi possono ancora fallire in situazioni critiche, specialmente in contesti complessi. Ciò evidenzia l'importanza di continuare a migliorare la sicurezza di questi sistemi per ridurre gli incidenti gravi, garantendo che i veicoli autonomi contribuiscano effettivamente alla sicurezza stradale anziché costituire un nuovo rischio.

Non si può parlare di veicoli autonomi senza menzionare Tesla, che si distingue come il principale sviluppatore e produttore in questo settore, come vediamo nella

Figura 4.22. La guida autonoma rappresenta il core business di Tesla, e l'azienda ha investito ingenti risorse nello sviluppo di tecnologie avanzate che mirano a rendere i propri veicoli sempre più sicuri e intelligenti.

D'altra parte, è importante sottolineare che le parti maggiormente esposte ai rischi derivanti dall'adozione di questa tecnologia sono proprio gli utilizzatori di Tesla, oltre che i conducenti di altri veicoli e i pedoni (Figura 4.23) . Questi gruppi, infatti, possono subire conseguenze significative in caso di malfunzionamenti o errori nei sistemi di guida autonoma. La necessità di un'ulteriore ricerca e sviluppo nel campo della sicurezza è quindi fondamentale per mitigare tali rischi e garantire una transizione sicura verso un futuro dominato dai veicoli autonomi.

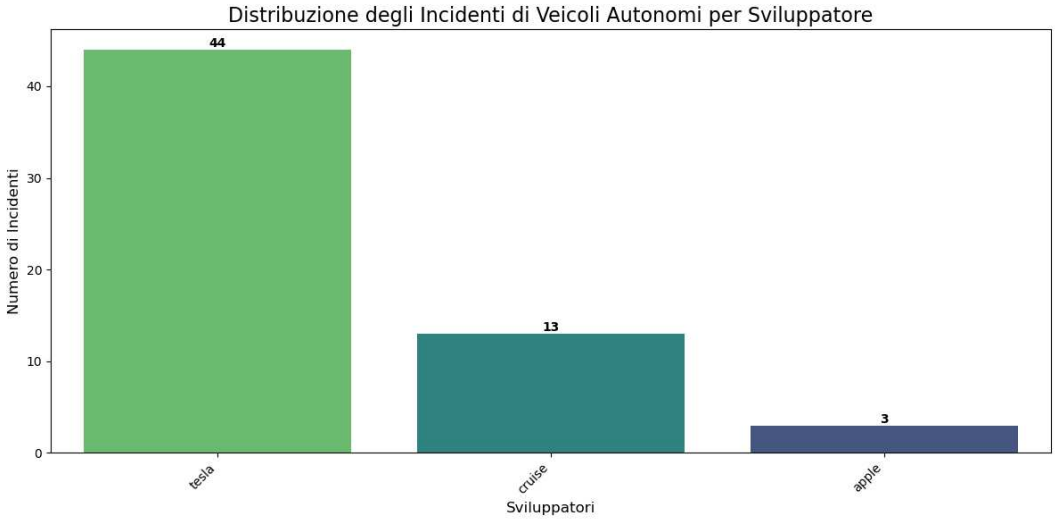

Figura 4.22: Sviluppatori di Veicoli Autonomi

Questi risultati evidenziano l'importanza di rafforzare ulteriormente la regolamentazione per garantire una maggiore sicurezza. Sebbene l'*EU AI Act* già preveda norme

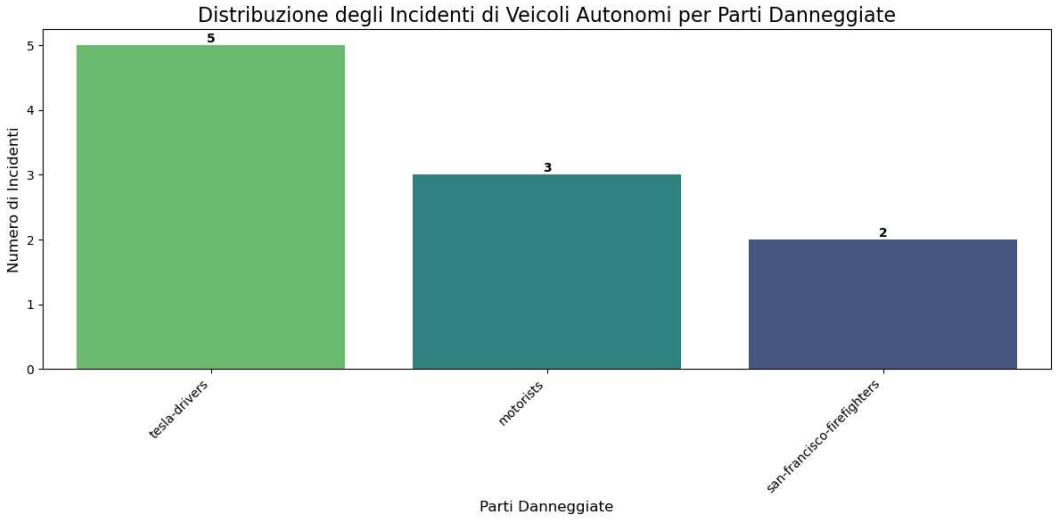

Figura 4.23: Parti Danneggiate di Veicoli Autonomi

significative per i veicoli autonomi, esistono margini di miglioramento. Potrebbe essere opportuno potenziare l'atto includendo normative ancora più rigorose sulla sicurezza, imponendo standard più elevati per i test e la validazione dei sistemi prima della loro commercializzazione.

Deepfake

Come illustrato nel grafico sottostante (Figura 4.24), la tecnologia dei *deepfake* è emersa relativamente di recente, ma ha già mostrato un'evoluzione rapida e preoccupante. Gli incidenti ad essa correlati hanno subito un notevole incremento, soprattutto negli anni 2023 e 2024 (quest'ultimo ancora in corso).

L'accelerazione nell'uso improprio di questa tecnologia solleva questioni critiche riguardo alla necessità di normative più rigorose e di strumenti avanzati per contrastare

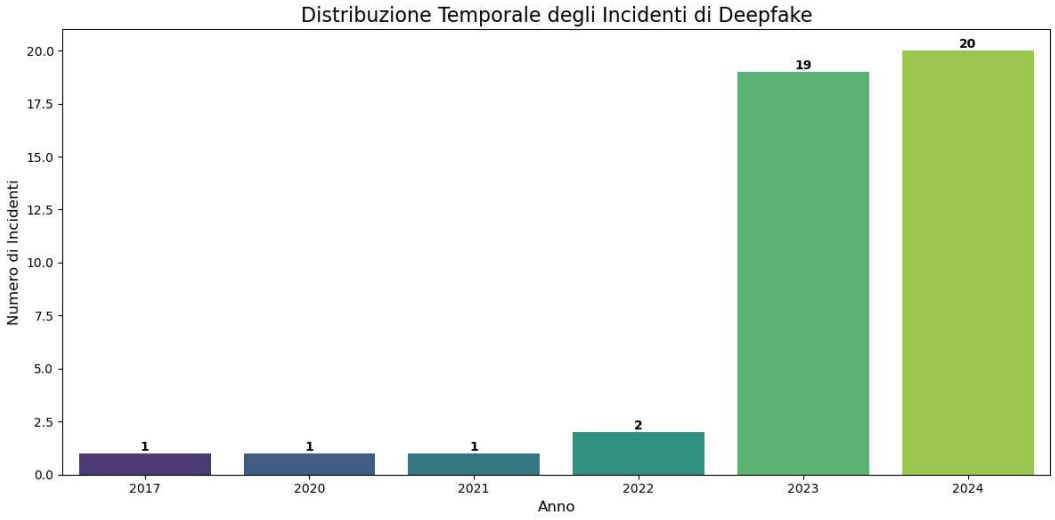

Figura 4.24: Evoluzione Temporale Incidenti Deepfake

i rischi associati. In questo contesto, diventa fondamentale non solo monitorare l'evoluzione dei *deepfake*, ma anche promuovere la ricerca e lo sviluppo di tecniche di rilevamento efficaci, al fine di limitare l'impatto negativo di questa tecnologia sulla società.

Nonostante l'aumento di questo tipo di incidenti, è importante notare come la quasi totalità degli eventi registrati rientri nelle categorie di rischio minimo e limitato, come evidenziato dalla Figura 4.25. Questo suggerisce che, sebbene la tecnologia dei *deepfake* stia evolvendo rapidamente e i casi d'uso improprio siano in crescita, la maggior parte degli incidenti finora ha avuto un impatto relativamente contenuto.

È fondamentale non sottovalutare i potenziali rischi futuri. La crescente sofisticazione dei deepfake potrebbe non solo aumentare la frequenza e la gravità degli incidenti legati alla manipolazione dei contenuti, ma comportare anche serie minacce

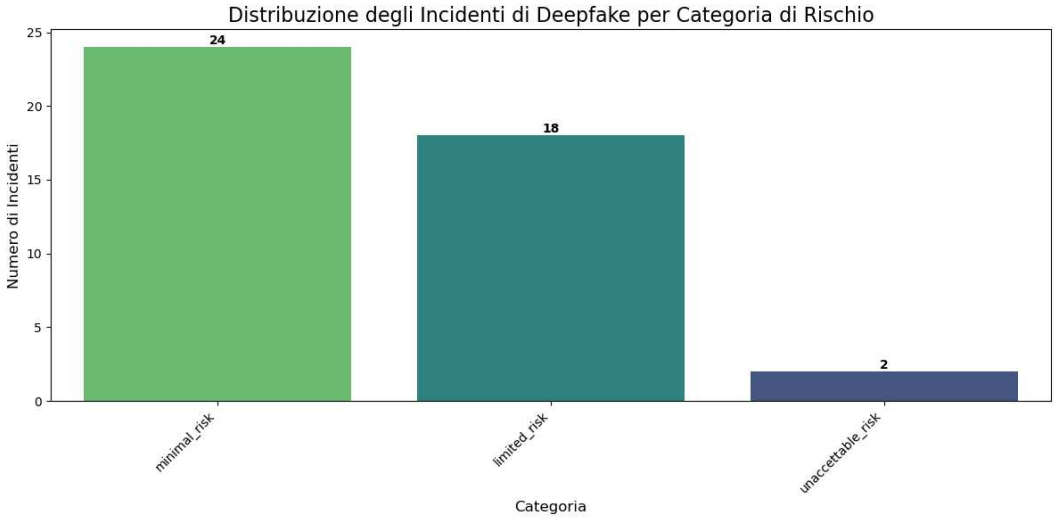

Figura 4.25: Categorie di Rischio per Deepfake

nel contesto politico. I deepfake potrebbero infatti essere utilizzati per influenzare le elezioni attraverso la manipolazione di discorsi e azioni di candidati politici, la diffusione di informazioni false e la creazione di divisioni tra gli elettori. Tali manipolazioni potrebbero avere l'effetto di destabilizzare il mondo politico globale, rendendo il rischio "limitato" un concetto decisamente riduttivo. Pertanto, è essenziale implementare misure preventive più rigorose e una regolamentazione attenta per mitigare questi pericoli e preservare l'integrità dei processi politici.

Infine, analizzando gli sviluppatori di questa tecnologia (Figura 4.26), emerge un dato significativo: un numero considerevole di incidenti è attribuito a sviluppatori classificati come *unknown*. Questo riflette la natura democratica e decentralizzata dei *deepfake*, dove chiunque, con competenze tecniche di base e accesso agli strumenti giusti, può creare e diffondere contenuti falsificati.

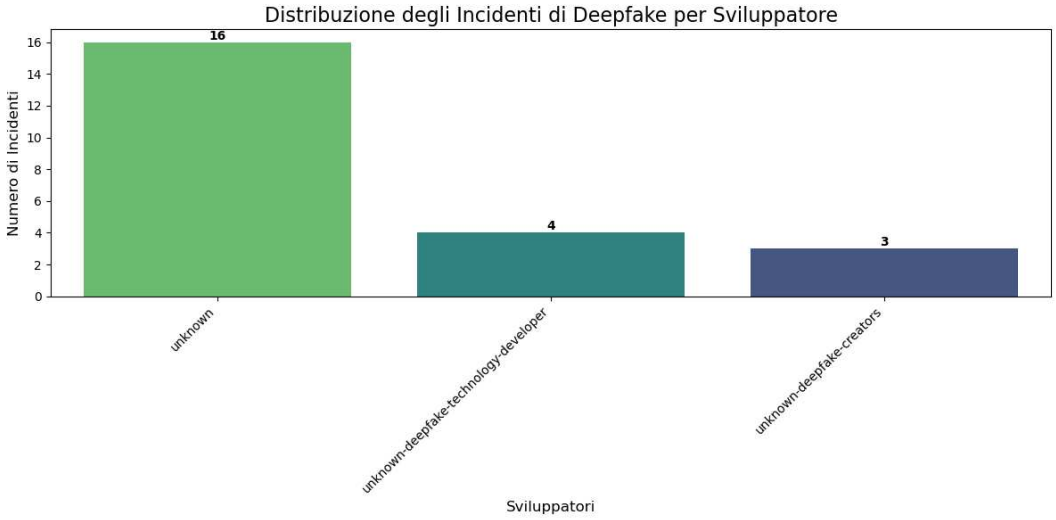

Figura 4.26: Sviluppatori di Deepfake

La mancanza di una tracciabilità chiara rende estremamente difficile individuare i responsabili, contribuendo all'incremento dell'anonimato in questo contesto. Inoltre, a differenza di altre tecnologie, non esiste una parte danneggiata specifica per questo tipo di incidenti: chiunque può essere bersaglio di un *deepfake*, indipendentemente dal proprio profilo o status [5]. Questa caratteristica rende i *deepfake* particolarmente insidiosi e sottolinea l'urgenza di sviluppare meccanismi di protezione più efficaci, sia a livello tecnologico che legislativo, per difendere individui e organizzazioni da potenziali abusi.

[5] https://uit.stanford.edu/news/dangers-deepfake-what-watch

CAPITOLO 5

Discussione

I risultati delle analisi dimostrano che l'intelligenza artificiale comporta rischi significativi, sottolineando così l'importanza e la necessità di una regolamentazione robusta. Le analisi esplorative del dataset hanno rivelato una prevalenza di sentimenti negativi riguardo a quasi tutti gli incidenti che riguardano le tecnologie IA. Questa evidenza conferma l'urgenza di adottare normative più *rigorose* e ben strutturate per affrontare adeguatamente i problemi legati all'uso dell'IA.

L'analisi temporale degli incidenti ha ulteriormente sottolineato la necessità di regolamentazione, rivelando *tendenze* e cambiamenti nel panorama degli incidenti. La crescente frequenza di questi ultimi e la variazione dei tipi di problemi riscontrati suggeriscono che la regolamentazione attuale potrebbe non essere sufficiente per affrontare l'*evoluzione rapida* delle tecnologie e dei loro impatti.

L'analisi dell'efficacia dell'EU AI ACT ha rivelato che, sebbene la normativa copra molti aspetti rilevanti degli incidenti, ci sono delle *lacune*. La classificazione degli incidenti nelle categorie di rischio previste dall'EU AI ACT ha mostrato che circa un terzo degli incidenti rientra nelle categorie di rischio *"inaccettabile"* e *"alto"*, mentre gli altri sono distribuiti tra *"rischio limitato"* e *"rischio minimo"*. Questo suggerisce che l'EU AI ACT sta affrontando in modo adeguato molte delle problematiche principali, ma potrebbe non essere completamente efficace nel trattare tutte le *sfide emergenti*, soprattutto in settori in rapida evoluzione come i *chatbot* e i *deepfake*.

L'analisi dei dati ha permesso di suddividere il dataset in tematiche specifiche, facilitando una valutazione più dettagliata delle *aree problematiche*. Anche se l'EU AI ACT fornisce un quadro utile, è evidente che la normativa necessita di un *aggiornamento* per affrontare in modo più specifico alcune delle tecnologie emergenti che presentano rischi unici e complessi.

Nella nostra analisi dei topic emergenti come *AI e chatbot*, *veicoli autonomi* e *deepfake*, è emerso che le normative attuali, sebbene efficaci in molti casi, presentano delle *mancanze* in aree specifiche. L'analisi delle categorie di rischio ha rivelato che, mentre molti incidenti sono classificati come a rischio minimo o limitato, la rapida evoluzione e la crescente sofisticazione di tecnologie come i chatbot e i deepfake richiedono un'attenzione normativa più mirata. In particolare, la regolamentazione

dei chatbot e dei deepfake sembra necessitare di aggiornamenti per affrontare adeguatamente i rischi associati a queste tecnologie. I chatbot, sebbene presentino un numero significativo di incidenti a rischio limitato, potrebbero creare problemi più gravi, come dimostrato da esempi riguardanti questioni politiche, se non regolati in modo specifico mentre i deepfake, con il loro potenziale di causare danni considerevoli, richiedono una normativa più severa per prevenire abusi e garantire la *sicurezza pubblica*.

In conclusione, l'analisi degli incidenti legati all'intelligenza artificiale conferma l'importanza di una regolamentazione adeguata e dinamica. L'EU AI ACT rappresenta un passo significativo verso una maggiore *sicurezza* e *trasparenza* delle tecnologie IA, ma è evidente che è necessario un continuo miglioramento e aggiornamento delle normative.

Il progetto ha utilizzato tecniche di NLP per analizzare e categorizzare i dati, dimostrando che, sebbene l'EU AI ACT fornisca una base solida per la regolamentazione dell'IA, esistono ancora aree di *miglioramento*. È essenziale che le normative rimangano flessibili e adattabili per rispondere rapidamente alle innovazioni tecnologiche e ai nuovi rischi emergenti. Solo con un approccio regolatorio continuo e proattivo sarà possibile garantire che l'IA venga utilizzata in modo *sicuro*, *etico* e vantaggioso per la società.

Tuttavia, lo studio presenta alcune limitazioni. In primo luogo, non avevamo

a disposizione dettagli completi sul tipo di algoritmo utilizzato nei sistemi di IA analizzati, il che potrebbe aver influenzato l'accuratezza delle nostre analisi. In secondo luogo, non siamo stati in grado di accedere alle notizie associate a molti degli incidenti riportati, limitando la nostra capacità di comprendere il contesto completo di tali eventi.

Nonostante queste limitazioni, i risultati del progetto forniscono indicazioni preziose per il miglioramento delle normative sull'IA, evidenziando l'importanza di un quadro regolatorio che possa adattarsi rapidamente ai cambiamenti tecnologici e ai nuovi rischi emergenti.

CAPITOLO 6

Conclusioni

In questa tesi è stata fatta un analisi degli incidenti legati all'intelligenza artificiale che ha messo in luce la necessità cruciale di una regolamentazione adeguata e continuamente aggiornata. L'*EU AI ACT* rappresenta un passo significativo verso la creazione di un quadro normativo solido, mirato a garantire la *sicurezza*, l'*affidabilità* e la *trasparenza* delle tecnologie IA. Tuttavia, data la rapida evoluzione delle tecnologie emergenti, è evidente che le normative devono essere in costante aggiornamento per affrontare efficacemente nuovi rischi e sfide.

Le analisi testuali, che includono tecniche come la sentiment analysis e l'*embedding* nel campo del NLP, hanno evidenziato una percezione generalmente negativa riguardo agli incidenti legati all'intelligenza artificiale. I risultati mostrano inoltre un trend in crescita nel numero di incidenti registrati, segnalando un aumento delle

preoccupazioni pubbliche e delle problematiche etiche associate all'IA. Questi dati sottolineano la necessità di un miglioramento delle normative esistenti, in particolare per affrontare in modo più efficace le tecnologie emergenti che rappresentano categorie di rischio elevato e sottolineano l'urgenza di adottare misure di sicurezza più rigorose e di rafforzare le normative esistenti per affrontare i rischi emergenti. Inoltre, l'esame dettagliato di tecnologie come *AI e chatbot*, *deepfake* e *veicoli autonomi* ha evidenziato aree specifiche che necessitano di una regolamentazione più mirata e approfondita. In particolare, le prime due tecnologie, essendo relativamente nuove e ancora scarsamente regolamentate, richiedono un'attenzione normativa particolare. La complessità e la rapida evoluzione di queste tecnologie suggeriscono che le normative devono non solo rispondere alle problematiche attuali, ma anche anticipare e prepararsi per i possibili sviluppi futuri, garantendo così una protezione adeguata e tempestiva.

È essenziale che le normative non solo siano ben strutturate ma anche *flessibili* e *adattabili*, in modo da rispondere tempestivamente alle innovazioni tecnologiche e ai nuovi rischi emergenti. Le tecnologie, specialmente quelle legate all'intelligenza artificiale, evolvono rapidamente e spesso in modi che non erano stati previsti al momento della loro regolamentazione iniziale. Per garantire che la regolamentazione resti efficace e pertinente, è necessario che essa possa essere aggiornata e modificata con agilità per riflettere i cambiamenti del panorama tecnologico e sociale.

Sebbene l'*EU AI ACT* fornisca una base solida per la regolamentazione dell'intelligenza artificiale in Europa, è indispensabile continuare a perfezionare e aggiornare le normative per rispondere in modo proattivo alle nuove sfide e opportunità. È cruciale che la regolamentazione non solo stabilisca standard elevati di sicurezza e protezione dei diritti, ma che promuova anche l'uso *etico* e *benefico* delle tecnologie emergenti. Solo attraverso un approccio dinamico e aggiornato è possibile garantire che l'intelligenza artificiale possa contribuire positivamente alla società, minimizzando al contempo i rischi e le problematiche potenziali.

Bibliografia

[1] **Binns** Renee. *Algorithmic Regulation: New Approaches for the Digital Age*. Springer, 2021.

[2] **Devlin** Jacob, **Chang** Ming-Wei, **Lee** Kenton, **Toutanova** Kristina,"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *arXiv preprint*, 2019.

[3] **Faisal** Asif, **Kamruzzaman** Md, **Yigitcanlar** Tan, **Currie** Graham,"Understanding autonomous vehicles: A systematic literature review on capability, impact, planning and policy", *Journal of Transport and Land Use*, 2019.

[4] **Forsyth** David A., **Ponce** Jean. *Computer Vision: A Modern Approach*. MIT Press, 2011.

[5] **Goldberg** Yoav. *Neural Network Methods for Natural Language Processing*. Morgan & Claypool Publishers, 2017.

[6] **Goodfellow** Ian J., **Pouget-Abadie** Jean, **Mirza** Mehdi, **Xu** Bing, **Warde-Farley** David, **Ozair** Sherjil, **Courville** Aaron, **Bengio** Yoshua. "Generative Adversarial Networks". In: *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*. 2014. *https://arxiv.org/abs/1406.2661*

[7] **Jurafsky** Daniel, **Martin** James H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, 2008.

[8] **Krol** K.,"The European Union's AI Regulation: A Critical Assessment", *Journal of AI and Ethics*, 2023.

[9] **Müller** Vincent,"Ethics of Artificial Intelligence and Robotics", *Stanford Encyclopedia of Philosophy*, 2022.

[10] **Ooi** Keng Boon, **Tan** Gerard W. H., **Al-Emran** Mostafa, **Al-Sharafi** Mohammed A., **Capatina** Alexandru, **Chakraborty** Apurva, **Wong** Lydia W.,"The Potential of Generative Artificial Intelligence Across Disciplines: Perspectives and Future Directions", *Journal of Computer Information Systems*, 2023.

[11] **Radford** Alec, **Metz** Luke, **Chintala** Soumith. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks". In: *Proceedings of ICLR*. 2016.
https://arxiv.org/abs/1511.06434

[12] **Sloot** Bart **van der**, **Wagensveld** Yvette,"Deepfakes: regulatory challenges for the synthetic society", *Tilburg Institute for Law, Technology & Society, Tilburg University*, 2022.

[13] **Stockman** George, **Shapiro** Linda G. *Computer Vision*. 2001.

Ringraziamenti

A conclusione di questa tesi, desidero esprimere la mia gratitudine verso chi mi ha accompagnato in questo percorso, senza il cui supporto questo lavoro non avrebbe mai potuto vedere la luce.

Un ringraziamento speciale va al mio Relatore che con grande dedizione e preziosi consigli mi ha guidato nelle analisi e nelle ricerche, facendomi dono di innumerevoli insegnamenti.

Ai miei Genitori, che mi hanno permesso sempre e da sempre di provare a raggiungere i miei sogni e i miei obiettivi credendo in me e non facendomi mai mancare nulla. Questo traguardo è anche vostro.

A mia Sorella, che è la persona a cui voglio più bene in assoluto e su cui potrò contare per sempre. Sappi che anche io ci sarò eternamente per te.

A tutte le Nonne e i Nonni, sia qui che in cielo, che hanno sempre vegliato su di me e si sono presi cura di me. Vi devo tutto.

Alla mia Famiglia allargata, che è sempre stata presente per ogni necessità e ogni aiuto. Sarete sempre parte di me.

Alle persone che sono diventate una parte Essenziale della mia vita in così poco tempo: avete reso questi ultimi due anni i più belli che abbia mai vissuto. Spero di avervi restituito almeno la metà di ciò che mi avete donato.

Agli Amici di sempre, che conosco da una vita e a cui so di poter dire qualunque cosa, in qualunque momento. Grazie per esserci sempre stati.

A Tutte le persone che ho conosciuto in questo percorso, sappiate che porterò nel cuore ognuno di voi. Vi voglio bene.

A Me, per ricordarmi che, se sono arrivato fin qui, valgo qualcosa. Che questo mi serva da lezione per il futuro, affinché io creda di più in me stesso. Mai mollare.

"...e per quanta strada ancora c'è da fare, Amerai il finale"