Politecnico di Torino

Tesi di Laure Magistrale

Mappatura e analisi delle startup nel settore dell'intelligenza artificiale in Italia

Relatore: Alessandra Colombelli
Candidato: Mihail Pascari

Anno accademico: 2019/2020

Indice

Mappatura e analisi delle startup nel settore dell'intelligenza artificiale in Italia**Errore. Il segnalibro non è definito.**
Abstract 3
Introduzione 4
Startup Innovative e Intelligenza Artificiale 6
Cosa sono le startup 6
L'importanza delle startup innovative 7
Startup di Intelligenza artificiale 10
Letteratura e framework proposti da terzi 14
Metodologia 19
Ipotesi di lavoro 19
Macro-rappresentazione della procedura 20
Classificazione top-down 22
Classificazione bottom-up: 23
Informazioni riguardo il codice 24
Dati 26
AIDA: database dati quantitativi economico- finanziari riguardo società italiane 26
Registro Imprese: database dati qualitativi riguardo società italiane 26
Merge dei dati: dataset startup innovative italiane 26
Pagine scaricate 27
Analisi dei risultati 28
Relazioni tra copertura siti web e caratteristiche azienda 28
Classificazioni a confronto 31
Conclusioni 42
Validità del framework di mapping delle startup di Intelligenza Artificiale 42
Miglior metodo di classificazione 43
Criticità del framework proposto e potenziali miglioramenti 44
Espansioni e utilizzi 45

Bibliografia .... **Errore. Il segnalibro non è definito.**

Abstract

Le startup che fanno uso di intelligenza artificiale stanno guidando la quarta rivoluzione industriale, allo stesso tempo sollevano un rilevante numero di dibattiti riguardo come questa tecnologia debba essere integrata nelle nostre vite minimizzando gli effetti collaterali. In questo studio l'autore esamina il mercato italiano delle startup di intelligenza artificiale. Lo fa andando a ricercare e mappare le aziende che rientrano in questo insieme, per poi studiarne lo spettro dei domini delle soluzioni proposte dagli imprenditori. Sono state analizzate le 12 mila startup presenti alla data dell'analisi nel Registro Imprese nella sezione Startup Innovative, dalle quali 260 sono state classificate come creatrici di algoritmi intelligenti per il loro business. Per la classificazione è stata utilizzata sia una classifica ottenuta analizzando la letteratura, sia una ottenuta con tecniche di web crawling. Con questo lavoro viene offerto un metodo ripetibile e scalabile per mappare e analizzare il mercato italiano delle startup di intelligenza artificiale. La distribuzione di iniziative imprenditoriali tra i vari domini di applicazione è eterogenea, alcuni sono più frequentemente intrapresi rispetto ad altri. Questo studio può contribuire a futuri studi che non dovranno ripetere da zero la raccolta e la classificazione delle informazioni.

Introduzione

Il legame tra la presenza di startup innovative in un certo ecosistema e la crescita economica di questo, è stato confermato in maniera solida dalla letteratura che si occupa di ricercare i driver di sviluppo e incremento del benessere della collettività (Matriciano, 2020) (Isenberg, 2011).

Quando si parla di Startup Innovative è necessario non metterle sullo stesso piano di quelle alle quali ci si riferisce con il generico termine "startup" che indica semplicemente un'azienda neonata. Difatti le aziende neonate con il semplice fatto di entrare in un mercato non apportano significativi contributi all'economia in termini di profitti, posti di lavoro e crescita in quanto spesso presentano attività effimere sul piano della sopravvivenza (Colombelli et al., 2016).

L'attuale letteratura non ha definito delle dinamiche nè consolidato delle definizioni inerenti le startup innovative. Riuscire a codificare cosa sia innovativo a priori sembra non essere una strada percorribile, inoltre è arduo riuscire a comprendere all'interno di ecosistemi complessi ricchi di agenti interdipendenti quali siano le cause di un successo o fallimento di un soggetto. E' forte l'esigenza di ricerca e proposte di metodi che portino a politiche di supporto alle aziende che potenzialmente possono creano un grande valore per se stessi e per l'intero sistema di cui fanno parte (Colombelli et al., 2020).

L'intelligenza artificiale è una di quelle tecnologie che permettono delle innovazioni radicali in diversi settori, pertanto è interesse della collettività che vengano studiate e possibilmente mappate. Queste stanno guidando la quarta rivoluzione industriale, allo stesso tempo sollevano un rilevante numero di dibattiti riguardo come questa tecnologia debba essere integrata nelle nostre vite minimizzando gli effetti collaterali.

In questo studio l'autore propone una metodologia di mapping delle attività innovative delle startup di intelligenza artificiale italiane basata su web scraping. La logica alla base della classificazione necessita di una tassonomia delle attività del mercato analizzato. Una nomenclatura è necessaria al fine identificare in maniera univoca certi tipi di tecnologie o business. Allo scopo verrà studiata e criticata una classificazione top-down proposta dalla Commissione Europea (Samoili et al., 2020). In completamento a questo verrà proposta una classificazione alternativa ottenuta a partire dall'analisi dell'insieme delle parole presenti sui siti web delle aziende che fanno intelligenza artificiale.

A seguito della descrizione della metodologia di lavoro è presente l'analisi di una prima implementazione del framework. I risultati suggeriscono che l'utilizzo di dati web-based in combinazione con dati forniti dal Registro Imprese e AIDA possono essere utilizzati per mappare l'ecosistema delle startup innovative di intelligenza artificiale in Italia. Inoltre, è stato messo in evidenza il fatto che, per la classificazione delle attività, una tassonomia derivata da un'analisi delle parole utilizzate dalle startup stesse restituisce una mappatura più granulare rispetto ad una ottenuta usando una tassonomia delle attività proposta dall'alto.

Startup Innovative e Intelligenza Artificiale

Cosa sono le startup

Il termine startup è stato cognato negli anni ’90 in un articolo che analizzava i processi di sviluppo di prodotti innovativi da parte di aziende software (Carmel, 1994). Nella Figura1 possiamo vedere come negli ultimi 16 anni vi sia stato un crescente interesse da parte degli autori di libri, articoli e paper nei confronti dell’argomento.

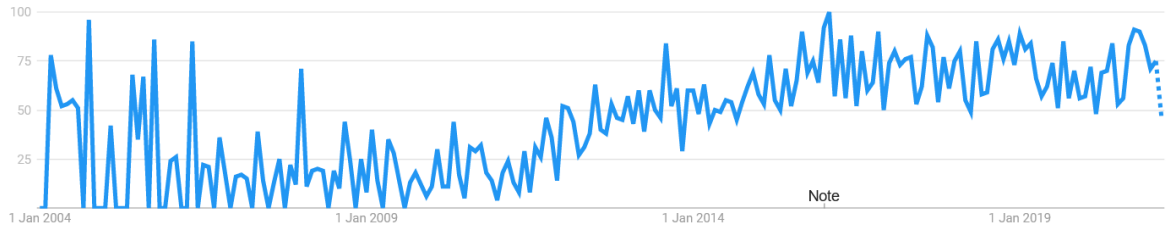


Fig1 Google Trends – Interesse nel tempo per il termine “startup” nei libri e nella letteratura

Le startup sono delle organizzazioni umane che vengono create allo scopo di risolvere un problema. L’associazione con l’innovazione che comunemente si ha di esse non è errata. Sono riconosciute come la miglior forma di organizzazione per un business che abbia come obbiettivo quello di creare un prodotto o un servizio innovativo. Ciò è collegato alla loro natura agile e improntata alla risoluzione di un problema in modo alternativo. Citando Erik Ries, imprenditore e autore di diverse analisi dei meccanismi dell’innovazione, le startup sono “un’organizzazione umana pensata per creare nuovi prodotti o servizi in condizioni di incertezza”.

Una delle più famose e definizioni di startup viene attribuita a Steve Blank, persona di spicco nella scena delle startup mondiali, ritenuto uno dei punti di riferimento quando si parla di formalizzare i processi di innovazione nelle startup. Per lui una startup è “un’organizzazione

temporanea che ha l'obiettivo di trovare un business model ripetibile e scalabile". Questa frase tocca i diversi aspetti delle startup fondamentali da comprendere per potervi interagire in maniera propria. Il primo termine "organizzazione" si riferisce alla necessità di coordinamento di risorse, in molti casi scarse se parliamo di startup europee, da parte del team. L'organizzazione di una startup richiede spesso l'impostazione di quelli che sono dei puri esperimenti scientifici. È pertanto necessario usare risorse finanziare in modo parsimonioso, in modo da poter effettuare più sperimentazioni possibili. È necessario poter accedere e coordinare l'uso di diversi strumenti tecnologici necessari alla sperimentazione. È fondamentale avere delle risorse umane con la giusta propensione alla sperimentazione, e con la giusta preparazione all'utilizzo di strumenti tecnologici e all'interpretazione dei dati risultanti dalla sperimentazione. Il secondo termine "temporanea" è coerente con l'analogia dell'esperimento scientifico. Deve esserci un disegno da parte del team su quelli che sono gli step di sperimentazione, su quali siano i risultati e in che tempi questi ci si attende possano essere raggiunti. Dopo un certo tempo la startup deve morire o evolversi in una realtà stabile che facendo tesoro delle sperimentazioni riesce a creare valore per i suoi clienti. Una terza alternativa sarebbe quella che nella realtà accade più frequentemente, e che si lega all'ultima parte della frase attribuita a Blank "trovare un business model ripetibile e scalabile": dopo una fase di sperimentazione iniziale la startup muta diverse volte la sua natura fino a trovare la forma che permette l'evoluzione in una realtà che crea valore, oppure fino a morire. E' pertanto fondamentale che vi sia sufficiente coscienza ed esperienza che porti il team ad avere la capacità di fare pivoting il giusto numero di volte (Blank, 2010).

L'importanza delle startup innovative

In molti casi tra policy maker e academici è diffusa l'idea che un grande numero di nascite di aziende sia necessariamente un fenomeno da considerarsi positivo dal punto di vista della collettività. La nascita di un'azienda non è sufficiente per ipotizzare la conseguente nascita di

opportunità di impiego e crescita economica. Bisogna considerare una serie di meccanismi per i quali molti imprenditori non sono altro che imprenditori per necessità, coloro che per sfuggire alla disoccupazione creano un business senza farsi guidare dall'osservazione di reali opportunità di innovazione, oppure imprenditori che imitano realtà già esistenti. Inoltre, anche tra gli imprenditori che effettivamente creano aziende guidati dall'intento di innovare sono rari i casi di innovazione radicale, quel genere di innovazione che attraverso la creazione di nuove realtà distrugge gli schemi preesistenti portando un periodo di sviluppo e crescita economica all'ambiente che ha permesso tale processo creativo. In quest'ottica la nascita di nuove aziende e il loro ingresso nel mercato in quanto tale potrebbe portare a effimere realtà e solleva seri dubbi sulla visione alternativa che anticipa il presunto ruolo dell'ingresso come veicolo per l'aggiornamento tecnologico, la produttività crescita e creazione di occupazione. Focalizzandoci sul concetto di Startup Innovativa invece che sulle generiche startup, le prospettive in termini di aumento della produttività, crescita economica e creazione di opportunità di lavoro cambiano positivamente in maniera consistente. È stato dimostrato come le startup nate principalmente per apportare una forma di innovazione abbiano maggiori probabilità di sopravvivenza e migliori performance. Più specificatamente, gli autori dello studio (Cefis, 2005) sono riusciti a dimostrare che essere un'azienda innovativa incrementa le chanche di soppravivenza dell'11% in più rispetto alla controparte non innovativa. In sintesi è consolidata l'idea che la proponsione all'innovazione sia affermata come un driver di crescita, e in maniera più specifica come un buon predittore delle probabilità di sopravvivenza e performance sopra la media in termini di profittabilità, export e creazione di posti di lavoro (Colombelli et al., 2016).

Le startup innovative insieme ad altri attori costituiscono le principali fonti di sviluppo economico. Sono stati condotti studi che hanno messo in evidenza la dipendenza della crescita

economica dalla presenza di business innovativi (Matriciano, 2020). Avendo un impatto positivo sull'occupazione e la crescita le startup innovative diventano delle entità preziose da coltivare e osservare per i governi e tutti gli attori del tessuto innovativo. Investire al fine di incentivare l'imprenditoria innovativa è uno degli strumenti che i governi possono utilizzare per generare più occupazione. Non è solo il welfare il ritorno che si ha come risultato, bensì ci sono considerevoli ritorni in termini di tasse che possono essere riutilizzati per alimentare l'ecosistema innovativo (Isenberg, 2011).

Data questa importanza largamente riconosciuta delle startup innovative come motore dell'innovazione, dell'occupazione e della crescita economica, negli ultimi anni sono state comuni le iniziative politiche che in tutto il mondo mirano all'incentivo della loro nascita e crescita. In questo scenario è possibile osservare un insieme di approcci e criteri eterogenei nel definire cosa siano le startup innovative. E' del tutto mancante un metodo sistematico che permetta di poter capire chi siano i produttori di innovazione. E' pertanto necessario l'intervento di ricercatori del settore che supportino il processo di rifinimento degli approcci utilizzati per analizzare la creazione e lo sviluppo delle startup innovative. Questo avrebbe ripercussioni positive anche sul design e l'efficacia delle politiche pubbliche (Audretsch et al., 2020).

Una migliore osservazione del mercato e pertanto delle migliori policy potrebbero risolvere alcuni problemi che impediscono alle startup innovative di crescere e creare valore. Nel suo articolo Colombelli (Colombelli et al., 2020) analizza l'effetto del DECRETO-LEGGE 18 ottobre 2012, n. 179 sulla scelta di strategie di protezione e appropriazione del valore generato da un'innovazione. La decisione di usare o meno metodi formali (esempio: brevetti) e informali (esempio: segreti industriali, tempi di entrata nel mercato, accesso ad asset complementari) al fine di proteggere i propri diritti, può fare la differenza al fine di sopravvivere. Dati i fallimenti di mercato inevitabili, far valere i propri diritti derivanti dall'aver innovato drena le energie e

le risorse delle startup. Adottare adeguate misure di protezione e appropriazione per queste realtà è costoso. Effettuando un'analisi su dati raccolti tramite survey nel 2016 da più di 1600 startup beneficiarie del decreto legge, è stata trovata un'evidenza di riguardo il fatto che policy di tipo finanziario hanno incentivato l'adozione di metodi di protezione formali e informali. Le policy sul mercato del lavoro hanno incentivato solo strategie di tipo formale (Colombelli et al., 2020).

Al fine di poter investire e monitorare l'effetto delle politiche messe in atto, è importante definire degli strumenti che possano catturare le attività delle startup innovative su larga scala, senza ritardi temporali e con bassi costi. Strumenti per la rilevazione dell'attività innovativa delle imprese esistono già, e nella gran parte dei casi si tratta di metodi basati su interviste o questionari distribuiti e raccolti con tempi lunghi e costi del personale alti.

Startup di Intelligenza artificiale

Con intelligenza artificiale (AI – artificial intelligence) si indica un sistema in grado di apprendere come imparare, in altre parole è una serie di istruzioni (algoritmo) che permette ad un computer di scrivere degli algoritmi senza essere istruito per farlo. Gli umani hanno la capacità innata di dedurre per analogie informazioni e regole che condizionano un certo sistema dalla semplice attività di osservazione del contesto. Un sistema AI può eseguire unicamente azioni basate su lettura e scrittura di dati, e non ha una conoscenza a priori delle relazioni che intercorrono tra i dati. È questa l'etimologia di "artificial", ossia l'aggettivo indica la provenienza non fisica, extra-sensoriale puramente basata su dati. (Corea, 2017).

L'intelligenza artificiale è figlia di diverse scienze. Eredita elementi dalle più antiche congetture degli antichi filosofi greci, cercando di sfruttare possibili definizioni di concetti come conoscenza, coscienza, sapere, codifica della conoscenza e ultima ma non meno importante etica e giustizia. Fa uso di teoremi di matematici che vanno dal 1700 ad oggi,

soprattutto dell'algebra lineare e della statistica. Soprattutto è figlia di tutti quei matematici e informatici che durante il '900 l'hanno pensata e modellata, uno su tutti Turing che ha addirittura ipotizzato dei test per poter mettere alla prova macchine in grado di mimare i comportamenti umani. I test consistono in una serie di domande con l'obiettivo di distinguere un agente umano da una macchina facendo leva sulla mancanza di capacità di interpretare un certo contesto da parte della macchina. (Corea, 2017) (Turing, 1950)

Gli storici del settore hanno individuato diverse epoche che questa tecnologia ha dovuto attraversa per arrivare al largo consenso di cui gode al giorno d'oggi. Nella seconda metà degli anni '50 sono state definiti in maniera formale i primi concetti in quello che è passato alla storia come l'incontro di Dartmouth. Finanziamenti del progetto DARPA negli USA e altre pubblicazioni e lavori hanno alimentato diversi momenti di popolarità effimera che portavano entusiasmo nell'ambiente per brevi periodi, sino alla prima metà degli anni novanta. Le tecniche utilizzate sino ad allora erano basate sui cosiddetti knowledge-based approach, senza riscuote mai il successo atteso probabilmente anche per una mancanza di risorse di cui oggi disponiamo in quantità maggiore (memoria, capacità computazionale, connessioni, ecc). Nella seconda metà degli anni '90 il paradigma più popolare è diventa il data-driven approach. Questa nuova famiglia di metodi ha innescato una certa positività degli investitori e in generale l'interesse degli addetti ai lavori.

Il 2012 viene ricordato come una pietra miliare del settore, in quanto in quell'anno è stato raggiunto un traguardo tecnico: una rete neurale è stata usata con successo per il riconoscimento di immagini mantenendo un tasso di fallimento inferiore al 15%. Questo ha alzato il livello dell'interesse e delle aspettative che imprenditori e policy makers hanno nei confronti dell'intelligenza artificiale guardando ai prossimi anni. Se prima di allora era un argomento molto popolare nei laboratori informatici, da quel momento ha cambiato il modo di immaginare la società di molte persone (Krizhevsky, 2012)

Storicamente questo periodo è stato indicato dagli storici contemporanei come la quarta rivoluzione industriale, della quale l'intelligenza artificiale e tecnologie affini sono le protagoniste indiscusse. In breve tempo i giganti del mondo hanno investito fondi per ordini di grandezza equiparabili alla spesa nel militare. Stati Uniti e Cina sono chiaramente in testa a questa corsa e mirano ad ottenere prima della controparte risultati che comporterebbero ritorni economici e sociali difficili da comparare a quelli che si sono avuti nelle precedenti rivoluzioni industriali. Gli altri stati non sono svantaggiati rispetto i due colossi solo in termini economici, bensì soffrono anche sul piano dell'accesso alla risorsa primaria per l'AI: i dati. Negli Stati Uniti hanno casa i colossi dei social più usati in occidente, i marketplace che indiscutibilmente controllano sono padroni del mercato occidentale. In Cina hanno quartiere generale non solo i social che hanno un monopolio sul territorio del gigante asiatico, ma il 43% delle transazioni e-commerce mondiali avviene su server di società cinesi. Inoltre, lo stato asiatico è leader mondiale per numero di pagamenti digitali, lascia indietro gli USA di ben 11 volte (Craglia et al., 2018).

In poco tempo questo settore, da essere confinato nei laboratori, è diventato campo di battaglia dei giganti della macroeconomia. In otto anni, dopo essere stata sdoganata la tecnologia, complici i piani di investimento dei governi e le strategie delle aziende più innovatrici molte soluzioni di Intelligenza Artificiale sono passate da progetti pilota a vere e proprie soluzioni a livello di produzione dalle quali alcune aziende non potrebbe più prescindere. Le applicazioni che nella gran parte dei casi sono già a regime all'interno dell'organizzazione sono quelle riguardanti l'elaborazione intelligente dei dati, sistemi di ranking o scoring che forniscono raccomandazioni agli umani ed elaborazioni del linguaggio umano. Le categorie che presentano ancora progetti in fase sperimentale il più delle volte sono quelli di automazione dei sistemi robotici e quello della guida autonoma (Turbaro et al., 2018)

È molto interessante il modo in cui le grandi aziende accedono alle innovazioni in questo settore. La strategia maggiormente diffusa è la cosiddetta acqui-hiring, questa prevede che una compagnia molto grande decida di inglobare una certa tecnologia scoperta da piccole realtà innovative attraverso l'acquisizione dell'intera startup. Solitamente in questa strategia il focus non è sui flussi di cassa che la compagnia possa apportare al proprio conto economico, bensì sul l'attenzione viene posta sul riuscire a convincere i founder a diventare manager del progetto o addirittura rimanere soci lasciando entrare la società più grande. Questo è dovuto al fatto che in un settore come questo in cui il maggior driver di successo è la qualità del gruppo e la cultura formatasi all'interno di questo, non è vantaggioso privarsi del team e del suo valore intrinseco. Una delle maggiori sfide per i governi che mirano a favorire la nascita di startup che possano apportare innovazioni creando o implementando intelligenza artificiale è la formazione di persone con le giuste conoscenze e competenze. Altri problemi consistono nell'avversità all'innovazione da parte delle aziende stabili, il che porta ad una certa resistenza da parte del management a cambiamenti che renderebbero questi soggetti più competitivi (Brock & von Wangenheim, 2019; Corea, 2017)

Riassumendo, le startup di intelligenza artificiale si muovono in un settore che gode di grandi attenzioni e che viene messo sotto pressione dalle aspettative degli stakeholder. In tutti gli stati sviluppati e alcuni di quelli in via di sviluppo, il governo propone incentivi finanziari e fiscali alle aziende che creano o usano soluzioni di intelligenza artificiale in quanto queste sono state individuate come potenziali casi di "Disruptive Innovation". Questa locuzione inglese sta ad indicare il fenomeno di nascita di realtà che rompendo gli schemi stabiliti in un certo mercato, comportano un'evoluzione dello stesso dalla quale si innescano periodi di crescita e arricchimento per l'intero ecosistema che accoglie tali cambiamenti.

Letteratura e framework proposti da terzi

L'innovazione radicale ha la capacità di rimodellare profondamente l'economia e aprire nuovi periodi di crescita, è pertanto vitale per l'intera società sviluppare un sistema in grado di misurare il grado di innovazione delle attività negli ambienti in cui questa deve essere intensa.

Misurare il livello di innovazione delle attività ad un livello sufficientemente accurato permettebbe ai ricercatori di analizzare quali siano i fattori trainanti e quanto siano efficaci le policy di innovazione. Purtroppo, vi sono evidenze che mostrano quanto i sistemi attualmente usati per l'indicizzazione dell'innovazione siano poco precisi e che soffrano di un ritardo sistematico nel dare una vista sullo stato del sistema innovativo. Degli esempi di tali sistemi attualmente in uso sono i questionari, le interviste e gli indici basati su numero e qualità dei brevetti (OECD, 2009) (Squicciarini, 2013) (Nagoaka, 2010).

Comunemente l'innovazione viene misurata in base ad indicatori costruiti su dati raccolti tramite survey su larga scala. Un esempio è il Community Innovation Survey (CIS), un'indagine biennale che viene condotta trasversalmente ai paesi membri dell'Unione Europea. Il contributo dell'Italia l'indagine viene svolta da ISTAT e viene pubblicata come Italian Innovation Survey (IIS). In entrambi i casi i questionari forniscono informazioni sia sulle imprese innovative sia riguardo quelle non innovative, inclusi i loro costi di Ricerca e Sviluppo. Il criterio di innovazione viene attribuito a seconda del suo grado di novità (nuovo per l'azienda, nuovo per il mercato, nuovo per l'industria o per il mondo) e in base al tipo di innovazione (innovazione di prodotto, processo, marketing oppure organizzativa). Purtroppo, questi indicatori soffrono di alcuni difetti. Ad esempio, l'IIS viene erogato solamente alle imprese dell'industria e dei servizi con almeno dieci addetti, escludendo pertanto tutte le imprese con un numero di impiegati inferiore, cosa molto comune nel caso delle startup.

Pertanto, la vera dimensione delle attività innovative rimane sconosciuta e può essere al massimo stimata attraverso modelli statistici. Con questo metodo i settori che presentano meno aziende ma che hanno del potenziale innovativo vengono tagliati fuori dalle osservazioni. Di conseguenza i metodi attualmente utilizzati per rilevare possibili innovazioni nel tessuto economico, soffrono di imprecisioni a livello settoriale e tecnologico. Bisogna porre una certa attenzione su quelli che sono i costi di esecuzione e tempi delle operazioni di distribuzione, raccolta e pre-processing delle interviste. Inoltre, per definizione queste interviste avranno sempre un ritardo di rappresentazione del mercato oggetto di studio (Kleinknecht, 2002).

Come alternativa ai questionari alcuni metodi impiegano le informazioni riguardo i brevetti (domande di brevetto, citazioni e licenze). Questo però porta ad una visione ridotta, infatti viene osservata solo la parte di innovazioni per le quali è stata ritenuta necessaria una strategia brevettuale. Inoltre, un brevetto non si traduce necessariamente in un utilizzo della tecnologia, spesso rimangono come pure invenzioni senza diventare innovazioni. Inoltre, i dataset di brevetti soffrono anch'essi di lag temporali sistematici (Squicciarini, 2013). Il tempo che intercorre tra la richiesta di registrazione e la disponibilità al pubblico in alcuni casi può essere di un anno o più (OECD, 2009).

Sfruttando l'odierna infrastruttura di reti di computer è ormai prassi comune per miliardi di aziende e privati comunicare e disseminare informazioni producendo crescenti quantità di dati. Per loro natura questi dati sono accessibili in forme non strutturate, memorizzati in sistemi decentralizzati, accessibili tramite protocolli differenti e tutto ciò comporta specifici requisiti per sistemi che hanno l'obiettivo di raccoglierli, pre-processarli e analizzarli. Il web mining, ossia l'applicazione delle tecniche di data mining a dati raccolti dal web, ha l'obiettivo di scoprire caratteristiche, pattern, tendenze e correlazioni da dati non strutturati pubblicati da diverse fonti. E' comprovato il fatto che queste tecniche portino a risultati di valore in diversi campi della ricerca scientifica (Askitas, 2015).

Per questo studio, che ha come obiettivo la definizione di un sistema di mapping dell'innovazione e in generale per la ricerca in campo economico, è particolarmente interessante quell'area del web formata dai siti web aziendali. Le aziende usano questi come mezzi di presenza sul web al fine comunicativo e ormai in molti casi anche al fine di vendere online. E' implicito pertanto che per ogni aziende online da questi fonti si possano ricavare informazioni riguardo i prodotti, i servizi, la credibilità, i traguardi raggiunti, le decisioni prese, le strategie intraprese e le relazioni con altre aziende (Gök, 2015). Ricavare informazioni riguardo le aziende leggendo i loro siti web comporta dei vantaggi in termini di dimensioni delle ricerche che si possono effettuare, costi ed eliminazione del ritardo temporale. Allo stesso tempo sorgono delle criticità in quanto non è banale raccogliere dati, armonizzarli e analizzarli. Allo stato attuale non vi sono dei framework di riferimento che permettano questo genere di raccolta di informazioni.

L'interesse per questo genere di strumenti è vivo, in letteratura esistono degli esempi che in parte ricoprono domande di ricerca simili e che generalmente tendono a suggerire un modello che permetta l'uso dei dati web-based per ottenere delle indicazioni riguardo l'innovatività di parti del tessuto economico. Vengono impiegate sia tecniche di web content mining che tecniche di web structure mining. Le tecniche di web mining consentono l'analisi dei testi e dei contenuti multimediali. Questo approccio è stato applicato da (Youtie, 2012) allo scopo di esplorare i siti web di 30 piccole imprese nel settore delle nanotecnologie, cercando informazioni riguardo la la transizione da fase di discovery a commercializzazione del prodotto. (Arora, 2013)ha usato un approccio simile per analizzare le strategie di entrata nel mercato di alcune aziende che vendono prodotti tecnologici a base di grafene. In entrambi i casi gli studi hanno identificato i diversi stadi di innovazione esistenti. (Gök, 2015) con un approccio basato su ricerca di keyword ha studiato le attività di Ricerca e Sviluppo di 296 aziende del Regno Unito. Il risultato web-based ha restituito insight più ricchi rispetto alla

ricerca basata su studio dei brevetti. Inoltre in questo studio è stato evidenziato come la raccolta dati tramite web scraping non soffra di un bias dovuto ad un cambio di comportamento da parte del rappresentante dell'azienda in quanto intervistato. (Beaudry, 2016) con un approccio basato su ricerca di keyword nel sito web aziendale è riuscito a mappare le attività di innovazione delle aziende canadesi nei settori aeronautico, spazio e difesa e nanotecnologie. L'autore ha potuto confermare delle correlazioni tra gli indicatori ricavati con questo metodo e gli indicatori tradizionali. (Nathan, 2017) ha combinato dati amministrativi, dati dai media e dai siti web al fine di sviluppare un nuovo metodo di misurazione dell'innovatività delle piccole e medie imprese. Per tale studio stati usati dati forniti da un'azienda che per business raccoglie dati riguardanti i siti web e dei media al fine di modellare cicli di vita delle aziende. Questo ha permesso di ottenere un numero di previsioni di lanci di prodotti o servizi tre volte maggiore a quello previsto con il metodo dei brevetti. Nathan conclude indicando le tecniche web-based come complementari alle tecniche in uso attualmente, e che questi servono per avere informazioni più ricche. Un altro risultato dello studio di Nathan è nell'aver scoperto che le aziende tecnologiche effettuano molti più lanci di prodotti o servizi rispetto alle non tecnologiche. Un altro studio che non fa uso di tecnologie di web mining ma che è interessante da nominare è quello di (Kim, 2012). In questo studio viene usato il text mining su un insieme di dati tratti da papers e brevetti al fine di trovare tecnologie emergenti e il loro stato di sviluppo.

Gli studi elencati basano il loro lavoro sull'idea di poter analizzare e rilevare le attività innovative a partire dalla ricca offerta di dati presenti sui siti web aziendali, potendo così dedurre informazioni su quello che è tutto un ecosistema innovativo. Tuttavia, nessuno di questi metodi si è ancora affermato come riferimento, o metodo da adottare al fine di effettuare ricerche nel campo dell'innovazione. Inoltre, mancano anche dei riferimenti in quanto a caratteristiche della fonti di dati stessa, il sito web. Sarebbe interessante avere dei dati riguardo

il modo in cui vengono strutturati i siti web aziendali genericamente (dimensione, profondità, tipo di informazioni presenti, tecnologie utilizzate, frequenza di aggiornamento, lingue supportate).

Metodologia

Per descrivere il metodo di lavoro si farà riferimento alla seguente terminologia. Sito web, sito internet e le sue versioni inglesi verranno usate per confermare la presenza e l'attività di un'azienda in internet. Viene dato per inteso che ogni sito è un insieme di pagine (esempio "www.nome-azienda.it", "www.nome-azienda.it/prodotti"). Le pagine sono organizzate gerarchicamente, quella al livello massimo è detta homepage o pagina principale ("www.nome-azienda.it"), mentre quelle ai livelli successivi vengono dette sotto pagine ("www.nome-azienda.it/prodotti"). La prima pagina scaricata da un sito web (che corrisponde all'indirizzo web fornito nei dataset ai quali si fa riferimento per far partire la ricerca, solitamente la homepage) verrà indicata come pagina di partenza, pagina iniziale, pagina di origine.

Ipotesi di lavoro

Per le startup odierne la comunicazione è al centro delle strategie di crescita, in questo contesto la presenza online permette loro di pubblicare informazioni riguardo I loro prodotti e servizi. La **prima ipotesi** alla base del lavoro consiste nel fatto che uno degli obiettivi di questi soggetti è mettere in evidenza gli aspetti innovativi di quanto vendono. Un piano di comunicazione non si limita alla pubblicazione di quanto viene offerto, bensì un obiettivo parallelo è quello di fornire informazioni riguardo la credibilità dell'azienda, i traguardi raggiunti, i principi, i partner e il relativo impegno (Kinne & Axenbeck, 2020). Pertanto, **la seconda ipotesi** che viene fatta ai fini della costruzione del framework è che i processi di innovazione e le eventuali partnership con forti innovatori vengano pubblicizzate in maniera esplicita (Kinne & Axenbeck, 2020).

I dati sono pertanto pubblici, sono decentralizzati e senza un costo di accesso. Un aspetto

negativo è l'assenza di una struttura costante dei dati trasversalmente a tutti I siti analizzati, il che porta ad avere dei costi di pre-processing e armonizzazione dei dati. Vi è pertanto bisogno di uno strumento per visitare ciascun sito in maniera automatica, scaricare I dati disponibili per poi poterli strutturare in modo che siano consultabili in modo omogeneo.

Tassonomia Commissione Europea

E' necessario menzionare il lavoro dal quale sono stati tratti la maggior parte dei termini per la classificazione top-down, essendo questo un paper pubblicato con l'intento di dare un punto di riferimento comune a tutti colore che devono discernere tra le attività di aziende di intelligenza artificiale. Questo lavoro è intitolato "AI Watch - Defining Artificial Intelligence. Towards an operational definition and taxonomy of artificial intelligence." (Samoili et al., 2020). Il report propone delle definizioni operative di intelligenza artificiale da adottare all'interno delle ricerche fatte dalla Commissione Europea sull'intelligenza artificiale. Ciò che viene proposta nella pratica è una tassonomia e un insieme di keyword. Questi sono corrispondenti in modo da poter incasellare tra loro quelli che sono i principali argomenti di ricerca e i topic trasversali agli argomenti principali.

Macro-rappresentazione della procedura

Nelle Fig. 1,2 viene descritto il processo che vede l'utilizzo di questo framework per ricavare una classificazione delle startup di Intelligenza Artificiale italiane basandosi sulle informazioni trovate nei loro siti web.
In maniera analoga a quanto previsto dalle metodologie basate su survey i dati di partenza sono estratti da dataset delle aziende italiane. Questi includono informazioni economico-finanziarie, qualitative e anagrafiche (esempio: settore, età, posizione geografica). È da qui che viene preso per ciascuna azienda l'indirizzo di partenza (URL).

La Fig.1 si riferisce alla procedura per l'individuazione delle startup AI basandosi sulla classificazione stilata dalla Commissione Europea (Samoili et al., 2020). Nella Fig.2 viene rappresentata la procedura, che utilizzando la classificazione top-down precedentemente ottenuta, permette di far emergere una classificazione bottom-up a partire dal contenuto dei siti delle startup marchiate come AI.

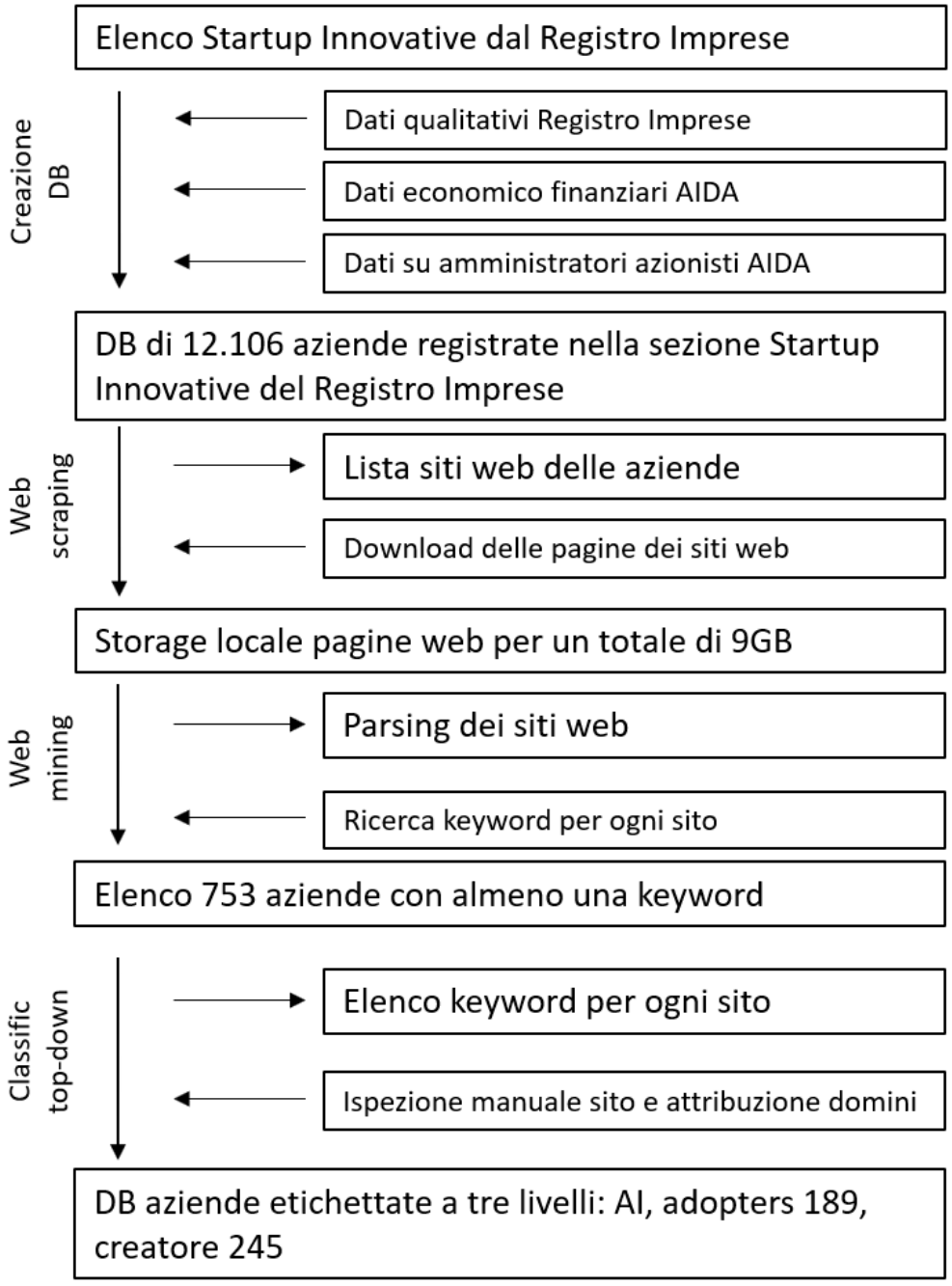


Fig.2 Classificazione top-down

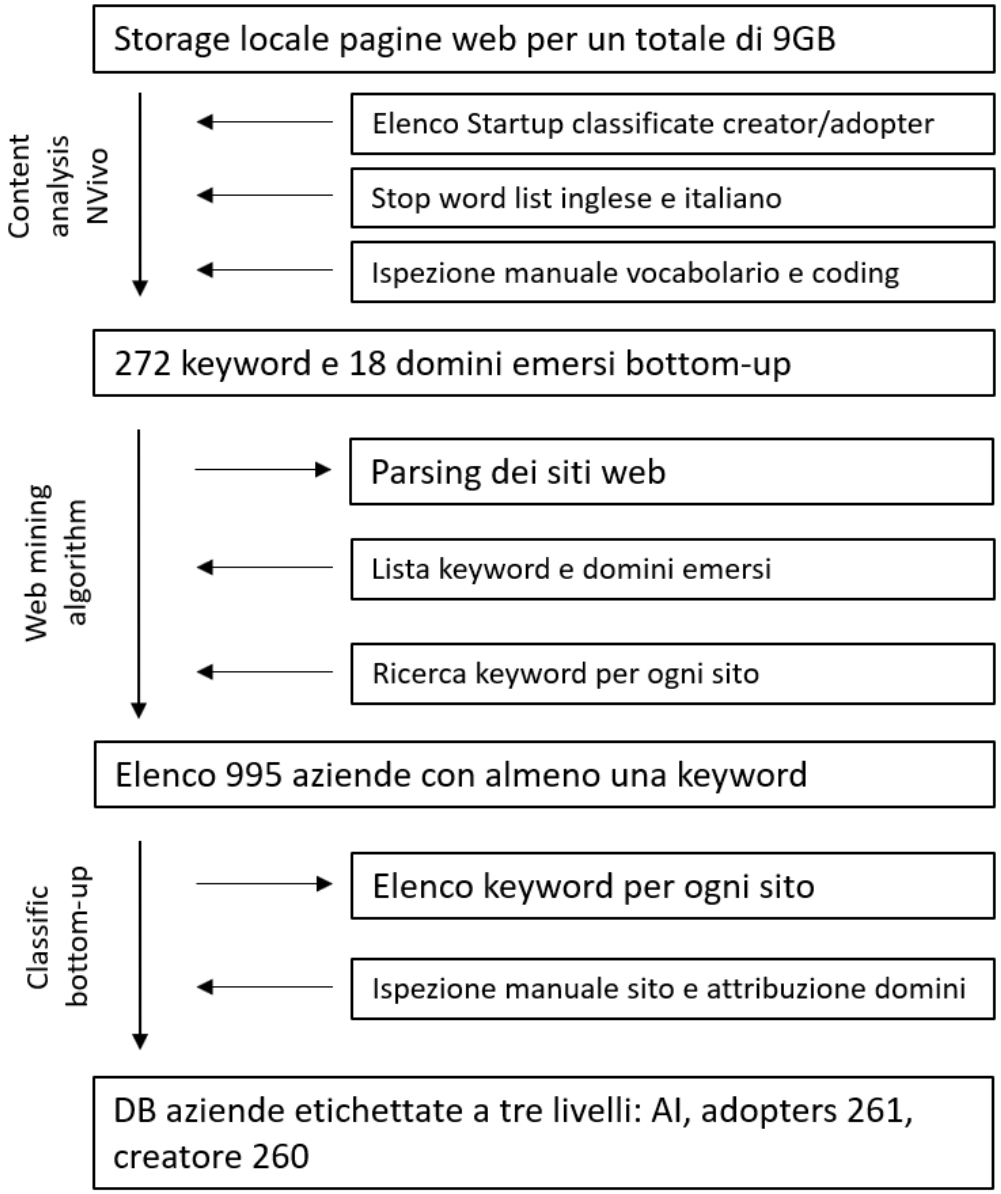


Fig.3 Classificazione bottom-up

Classificazione top-down

Partendo da una cernita della letteratura e dei report affini all'argomento affrontato si raccolgono in una lista le keyword utilizzate per riferirsi alle tecnologie sotto il cappello dell'AI, alle applicazioni, ai framework e concetti rilevanti al fine di comprendere il mercato. Infine, si sommano anche tutte quelle keyword individuati da lavori di cernita simili.

Usando lo script di web scraping per ogni sito delle 12000 aziende viene scaricata una copia locale, per poi effettuare una ricerca delle keyword nel file così ottenuto con lo script di web mining. Questo permette di ottenere un elenco di tutti i siti che hanno almeno una parola chiave, con l'elenco dei match di fianco. Successivamente, attraverso una scannerizzazione manuale, per ogni sito positivo viene effettuata una classificazione indicando se si è o meno difronte ad un falso positivo, quali siano i domini AI e quale attività all'interno di tali domini. Viene indicato anche qualora sia un caso di produttori della tecnologia (creator) o utilizzatori (adopter). Viene ritenuto che essi siano dei creator solo nel caso in cui in maniera oggettiva indicano loro stessi tale situazione. Un esempio può essere l'uso di locuzioni di significato equivalente a "abc srl ha sviluppato l'algoritmo", "algoritmo di proprietà di abc srl", "la nostra soluzione proprietaria". In opposizione vi sono casi equivalenti alle seguenti frasi: "abc srl impiega le migliori soluzioni AI allo stato d'arte", "abc srl fa uso di servizi di AI", "abc srl insieme al partner tecnologico ha sviluppato l'algoritmo."

Classificazione bottom-up:

Utilizzando Nvivo, un software di analisi testuale dei dati, è stato estrapolato l'insieme di tutte le parole usato in ciascuno sito web di una delle startup AI reso disponibile offline nella fase di classificazione top-down. Dopo una scrematura manuale del vocabolario mirata ad eliminare parole poco significative, è stato avviato un processo di creazione di cluster delle keyword così individuate. I cluster vengono creati da Nvivo sulla base di co-occorrenze. Ogni cluster permette la definizione di quelle che saranno le attività ed i domini della classificazione bottom-up: ad ogni parola ad essi partenente viene applicata un'etichetta (in gergo viene detto coding), questo permette di avere anche dei gradi di correlazione tra cluster.

Attraverso lo script di web mining viene ricercata ogni nuova keyword nei file locali di tutte le 12000 startup italiane. Si ottiene un elenco di tutti i siti che hanno almeno una parola chiave, con l'elenco dei match di fianco. Successivamente, attraverso una scannerizzazione manuale, per ogni sito positivo viene effettuata una classificazione indicando se si è o meno difronte ad un falso positivo, quali siano i cluster di appartenenza in caso positivo. Quello che si ottiene al termine delle operazioni è un dataset in cui il numero di aziende classificate come AI è di 260. Questo e altri dati quantitativi vengono mostrati e discussi nel capitolo "Analisi dei risultati" di questo documento.

Questa tecnica di mapping soffre di lag inferiori rispetto la creazione di un classificazione top-down e una successiva survey. Le fonti di ritardo nella ricezione di un'informazione riguardo potenziali innovazioni si riducono alla decisione dell'azienda di posticiparne la pubblicazione e ad eventuale frequenza di esecuzione del web scraper. È facile immaginare espansioni di questo sistema in prima battuta su altre tipologie di startup, successivamente su altri segmenti di aziende più mature. Questo permetterebbe la costruzione di mappe dell'innovazione relativamente poco costose (sia per chi raccoglie che per chi viene osservato), capillari e in breve tempo, probabilmente da combinare con le tecniche di survey per avere sempre dei confronti sul medio e lungo periodo.

Informazioni riguardo il codice

Il web scraping è una tecnica attraverso la quale viene usato un software per estrarre dati da un sito web in maniera automatica. Per la scrittura del codice è stato scelto Python perché così il progetto è facilmente modificabile da chiunque abbia conoscenze di base di logica e programmazione. Un'eventuale interfaccia grafica permetterebbe l'uso di istanze di questo codice da parte di personale di ricerca totalmente estraneo alla programmazione. Per gestire le chiamate http ai siti web è stata usata la libreria Requests.

Deve poter essere eseguito frequentemente su grandi quantità di siti, in modo da poter creare un panel dei siti web. In questo può tornare utile l'infrastruttura IT del Politecnico per parallelizzare diverse esecuzioni contemporanee e velocizzare il download dei contenuti, dato che per completare la richiesta e il download di tutte le pagine di un sito web in alcuni casi sono richiesti tempi dell'ordine dei minuti.

Il web mining è l'insieme delle tecniche utili ad estrarre informazioni da grandi quantità di testo. Anche per questo modulo è stato scelto Python come linguaggio di programmazione. Il modo in cui vengono salvati i file deve essere ottimizzato per storage & retrieve agile in maniera trasversale alle aziende e al tempo, per questo la libreria principale è Beautifulsoup. Quest'ultima permette il cosiddetto parsing delle pagine web, ossia esse vengono lette tenendo conto di quali parti siano visibili all'utente e di come siano strutturati i contenuti. In generale questo è un modulo che varia a seconda delle analisi che si vogliono fare, ed è quello che si adatta meglio per essere sostituito da un software di Reti Neurali che faccia context analysis.

Dati

Per questo studio sono stati utilizzati dati provenienti da AIDA e Registro Imprese.

AIDA: database dati quantitativi economico- finanziari riguardo società italiane

Il Politecnico ha un accordo grazie al quale fornisce accesso ai dati per i ricercatori. Tramite AIDA BvD fornisce principalmente delle informazioni di carattere quantitativo economico-finanziari riguardo le società di capitale italiane.

Unica condizione di ricerca utilizzata dall'autore per individuare il dataset di aziende per lo studio è che queste siano iscritte alla sezione Startup Innovative del Registro Imprese. In totale il portale ha restituito 12106 risultati per i quali sono stati esportati i dati economici e finanziari per gli anni di attività che vanno dal 2013 al 2020.

Registro Imprese: database dati qualitativi riguardo società italiane

Il sito del registro imprese fornisce libero accesso alla lista aggiornata delle startup innovative, fornendo insieme all'elenco una serie di informazioni principalmente di carattere qualitativo.

Merge dei dati: dataset startup innovative italiane

I dati offerti da AIDA e Registro Imprese sono parte fondamentale del framework di lavoro, sia per avviare la ricerca che per poter poi ottenere informazioni che nascono dall'incrocio dei dati della classificazione con quelli di carattere economico finanziario. L'autore pertanto ha definito delle routine che permettano di scaricare localmente tali dati e pulirli, mantenendo uno schema costante del dataset. Questo permette un certo grado di automazione delle procedure attraverso l'uso di do-file STATA. Vengono importati in file dta tutti i dati scaricati dalle fonti elencate precedentemente per poi essere pre-processati e sottoposti ad una procedura di unione (merge) usando come indice univoco il codice fiscale delle aziende. In

entrambi i dataset è presente una colonna che indica l'eventuale sito web dell'azienda osservata. Delle 12106 aziende 2119 non presentano un indirizzo. Queste ricevono una procedura che parte da una ricerca di "ragione sociale" e "ragione sociale + partita iva" sui principali motori di ricerca allo scopo di trovare il sito web. Nei casi in cui venga individuato questo viene inserito nella tabella delle startup.

Pagine scaricate

La copertura di aziende fornite in input al web scraper è di 81,6% (9881 indirizzi presenti tra AIDA e Registro Imprese). Il 7,34% (725) degli URL sono risultati non raggiungibili in quanto il server non funzionava al momento della richiesta della pagina o in quanto il sito non presentava contenuti. Il numero totale di pagine scaricate è di 85.877, con un peso di oltre 9Gb di dati. Si tratta di file "plain text" nei quali il contenuto è riportato in HTML. Una delle funzioni del modulo di web mining è quella di "parsing" delle pagine, ossia leggendo i file testuali ai fini dell'analisi viene ricreata la pagina web originaria. Il tempo di esecuzione dell'intero processo di classificazione, top-down e bottom up, ha richiesto 40 ore di esecuzione avvalendosi di una macchina di fascia media (processore Intel i7 terza generazione, 8Gb RAM, SSD Samsung EVO840, connessione 50Mbps).

Analisi dei risultati

Relazioni tra copertura siti web e caratteristiche azienda

La copertura di aziende fornite in input al web scraper è del 81,6% (9881 indirizzi presenti tra AIDA e Registro Imprese), tuttavia non è costante al variare di alcune variabili anagrafiche. I grafici 2-5 illustrano rispettivamente la copertura di siti web al variare di Settore, Classe di produzione, Età dell'azienda e Posizione Geografica.

Tra i settori non vi sono differenze rilevanti in quanto ad URL disponibili. Anche nel caso della classe di produzione non vi sono differenze rilevanti. La colonna F, che rappresenta il caso di aziende con più di 5 milioni di euro, è poco significativa perché presenta meno di dieci osservazioni, tutte con un sito disponibile. Dal grafico 4 possiamo osservare che per le fasce d'età 1 e 2 l'assenza di siti web è contenuta al di sotto di cifre significative. Non si po' dire la stessa cosa per la terza fascia, che contiene aziende con un'età maggiore ai 4 anni. Quest'ultima presenta una percentuale di missing del sito web del 30%. Questa sistematicità viene confermata dal test chi-quadro nella Tabella 1.

Infine, nel grafico 5 abbiamo una situazione simile per tutte le posizioni ad esclusione di Lazio, Sicilia e Val d'Aosta che sono le uniche a superare il 20% di casi in cui manca un indirizzo web. Nel caso della Val d'Aosta i dati non sono significativi in quanto presenta un'esigua popolazione di startup, solo 22. Diverso è il caso della Sicilia e soprattutto della regione Lazio che ospitano rispettivamente 367 e 1.067 aziende. Questi risultati sono coerenti con quelli deducibili dal dataset pubblicato da Eurostat che fornisce dati riguardo la copertura di siti web per le aziende europee [EUROSTAT Websites and Functionality].

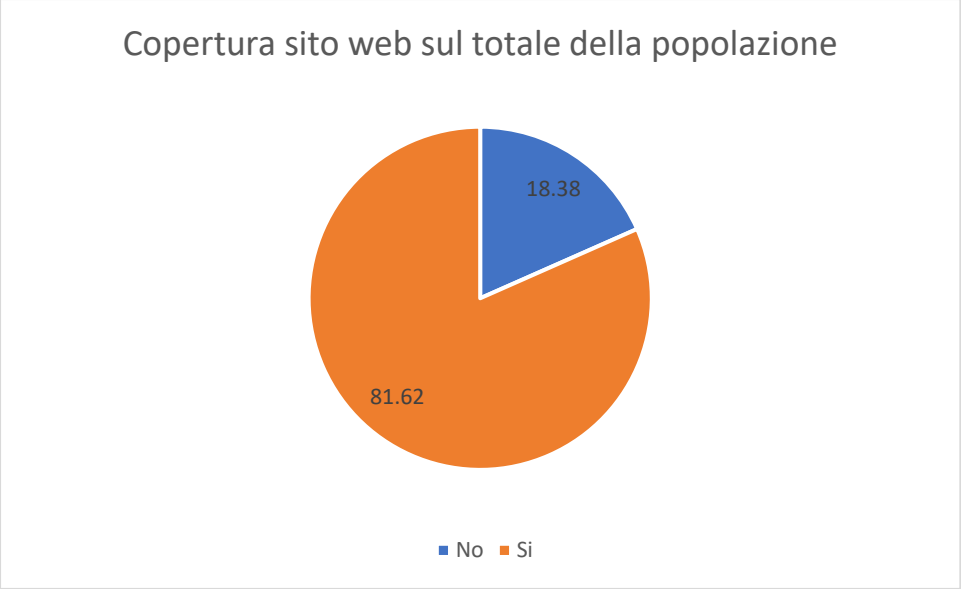


Grafico1: Copertura sito web sul totale della popolazione

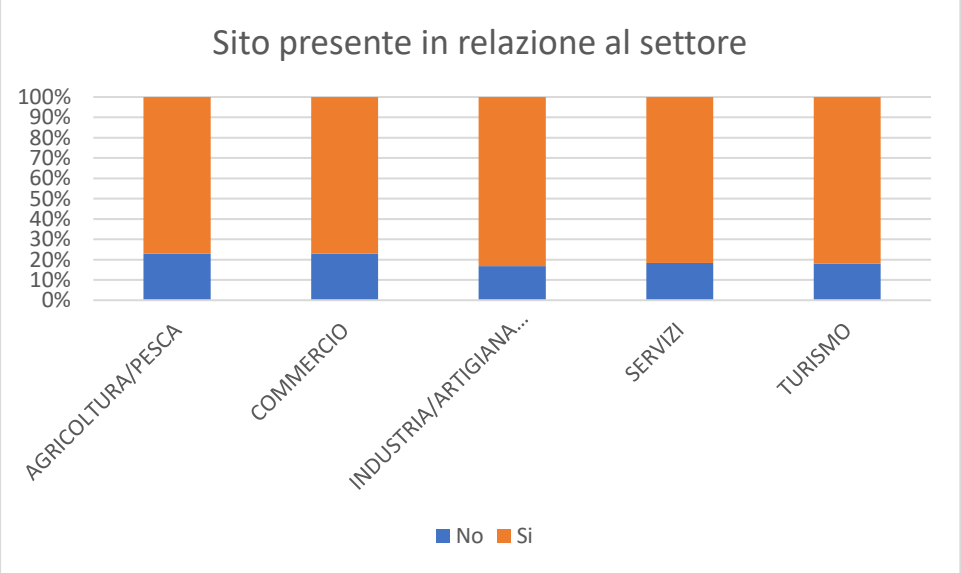
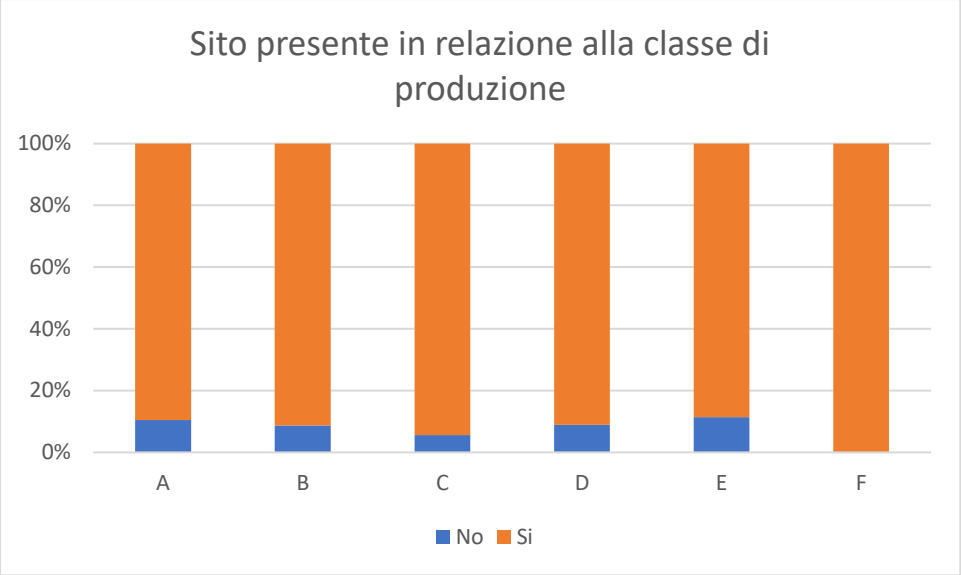


Grafico2: Copertura sito al variare del settore di attività



Grafic3: Copertura sito al variare della classe di produzione

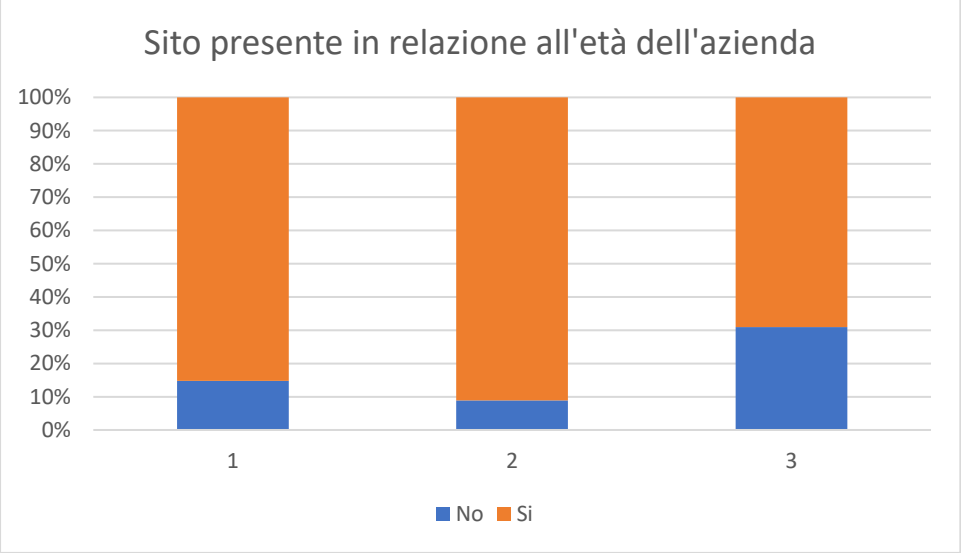
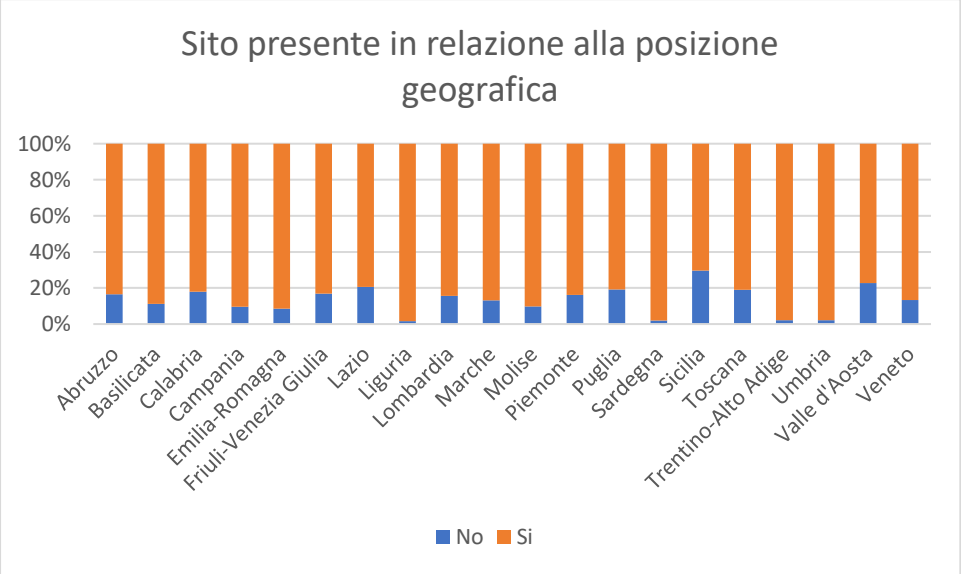


Grafico 4: Copertura sito al variare dell'età dell'azienda

	Classe d'età azienda			
sito_presente	1	2	3	Total
0	299	482	1442	2,225.00
1	1,726.00	4,924.00	3224	9,881.00
	15%	9%	0.309044149	18%
Total	2,025.00	5,406.00	4666	12,106.00
Pearson chi2(3) = 828.3948 Pr = 0.000				

Tabella 1: Copertura sito al variare dell'età dell'azienda con test chi-quadro



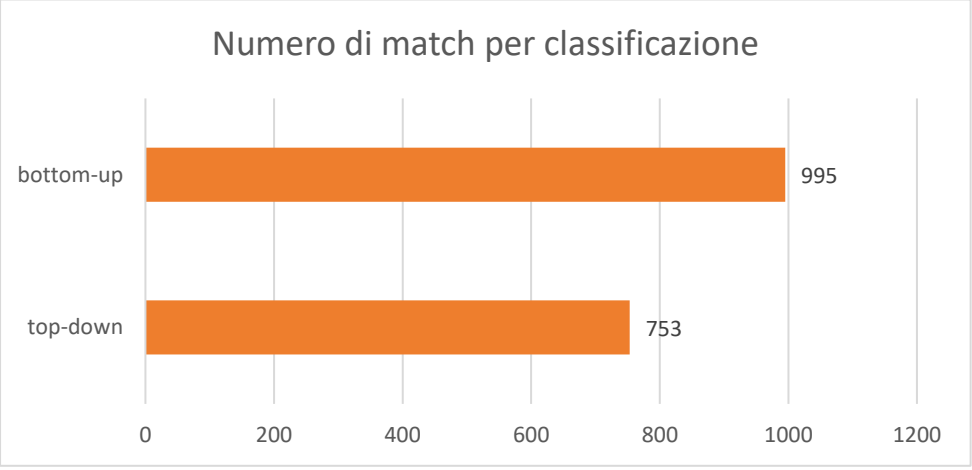
Grafioc5: Copertura sito al variare della posizione geografica

Classificazioni a confronto

La popolazione di startup studiate è 12.106 e di queste 9881 presentano un sito web che è stato fornito al web crawler per la classificazione top-down in prima istanza, per la classificazione bottom-up in seconda. Per ciascuna classificazione i dati scaricati sono stati usati come input

del web miner, che ha fornito una lista di potenziali aziende che hanno a che fare con la creazione o l'uso di soluzioni AI, sulla base del ritrovamento di keywords nei testi del sito web. Per ogni classificazione la lista di keyword differisce. La top-down usa una serie di keyword fornite dalla Commissione Europea (Samoili et al., 2020) in quello che è un suo lavoro di creazione di una tassonomia di riferimento, con l'aggiunta di termini individuati dall'autore nella letteratura o presso fonti rilevanti in materia. La bottom-up usa un vocabolario creato da una cernita manuale di tutte le parole usate nei siti delle startup etichettate come inerenti all'intelligenza artificiale.

Questi due diversi processi hanno prodotto risultati che differiscono in termini di numero di match, quindi di potenziali startup di intelligenza artificiale, ma anche in termini di numero di adopters e creators. Il grafico 6 mostra come il processo di classificazione bottom-up restituisca 995 match superando in maniera rilevante i 753 match dell'approccio top-down. Il grafico 7 mostra come questo numero di molto maggiore della classificazione bottom-up si traduca in 15 creators in più, 72 adopters in più (+38%) e 155 falsi positivi in più(+48%).



Grafioc6: numero di match per tipologia di classificazione

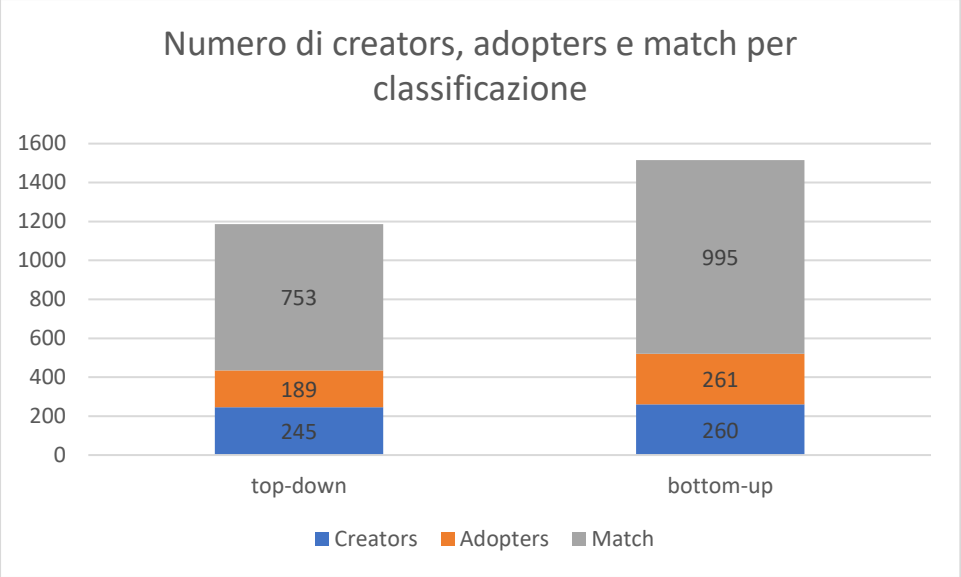


Grafico7: numero di match, creator e adopter per tipologia di classificazione

È stato stimato che il valore del mercato italiano dell'intelligenza artificiale nel 2019 si aggira intorno agli 84 milioni di euro (Osservatorio_Artificial_Intelligence, 2019). Stimando lo stesso valore a partire dalla sommatoria dei ricavi 2019, ai quali possiamo accedere grazie ad AIDA, otteniamo i valori riportati nel grafico 8. Tra le due stime non si va oltre i 14 milioni di differenza. Tale differenza potrebbe essere imputata al fatto che in questo studio vengono considerate solo le aziende iscritte alla sezione Startup Innovative del Registro Imprese, invece nello studio condotto dall'Osservatorio Artificial Intelligence i dati si riferiscono ad un insieme più ampio di tipologie di aziende.

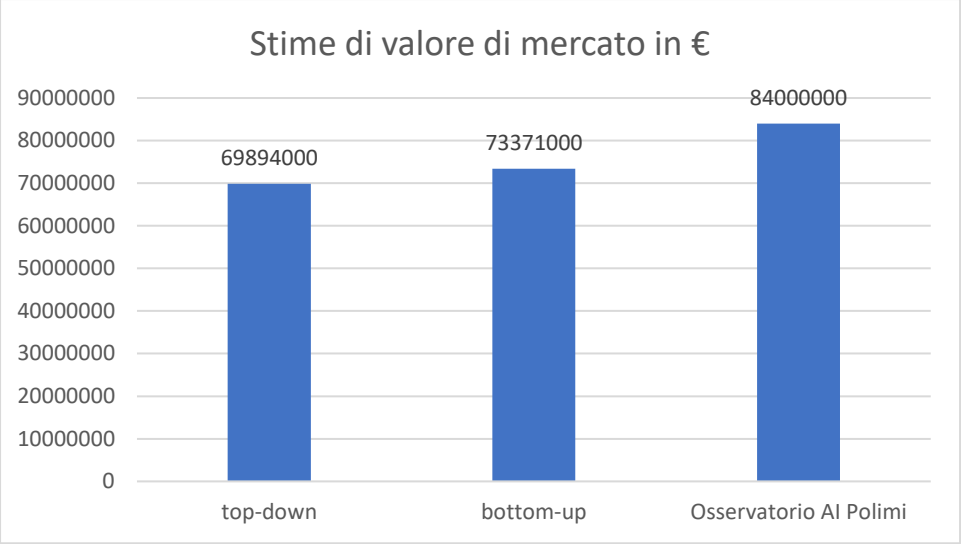


Grafico 8: stima del valore del mercato italiano dell'intelligenza artificiale nel 2019 per classificazioni e per report Osservatorio Artificial Intelligence del Politecnico di Milano

Al termine delle rispettive procedure le due classificazioni restituiscono non solo una lista di potenziali match ma anche a quale categoria questi appartengono. Questo avviene tramite un processo di cernita manuale dei match. Anche in questo caso le due classificazioni forniscono dei numeri diversi, questa volta in termini di domini e keyword. La classificazione top-down presenta 7 domini e un totale di 73 keyword, le quali sono state in gran parte identificate attraverso il report della Commissione Europea come significative in un contesto AI. Dalla classificazione bottom-up sono emersi 18 cluster di parole, che poi sono diventati i domini. In totale le parole identificate come keyword sono 272. In entrambi i casi le parole sono sia in lingua italiana e inglese, in particolare nel caso della top-down sono stati utilizzati termini italiani quando a priori sembrava sensato.

In entrambe le classificazioni, per ogni startup identificata come creator, sono stati assegnati almeno un dominio di attività. In un numero rilevante di osservazioni è stato necessario utilizzare più di un dominio, in quanto la realtà presenta soggetti che non usano il proprio know-how ad un solo fine. Pertanto, è necessario osservare le relazioni tra domini che emergono nei due casi. Nel grafico 9 possiamo osservare la rappresentazione delle relazioni tra domini che

sono emerse dalla classificazione top-down, mentre nel grafico 10 abbiamo uno schema analogo per la classificazione bottom-up. Bisogna tenere a mente che i domini mai comparsi in osservazioni multi-dominio non vengono rappresentati in quanto non verrebbero connessi ad una controparte. Inoltre, D1 e D2 non implicano relazioni gerarchiche ma è puramente connesso all'ordine nel quale i domini sono stati osservati sul sito-web. È immediato osservare che il primo schema presenta un numero basso di domini e relazioni rispetto al secondo schema. Il secondo schema mostra una rappresentazione meno incentrata su pochi domini, e rappresenta più relazioni per ognuno di questi.

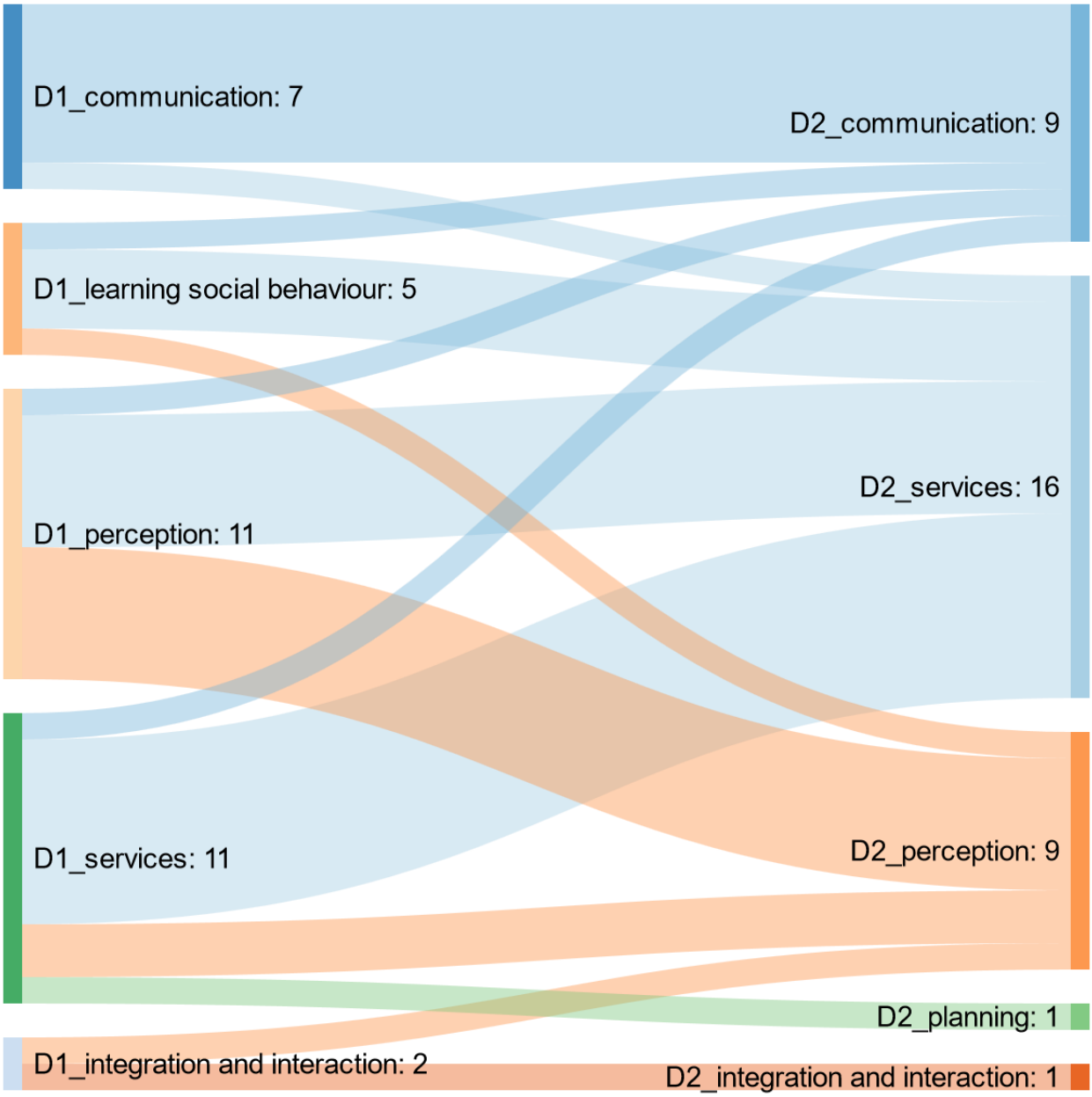


Grafico9: relazioni tra domini della classificazione top-down

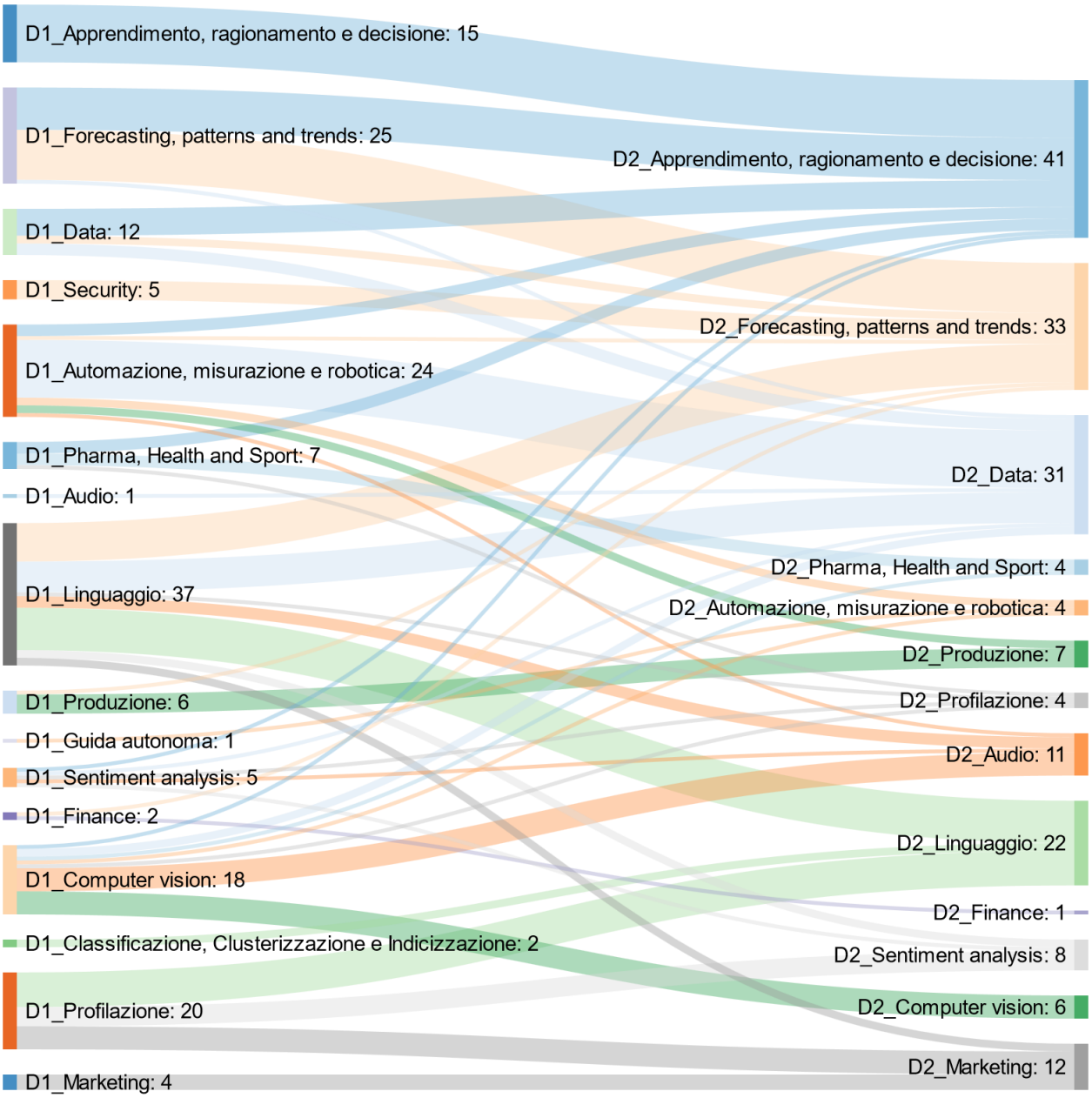


Grafico10: relazioni tra domini della classificazione bottom-up

Una diversa distribuzione dei match tra i domini la si può notare direttamente dai grafici 11 e 12. Nel primo caso il dominio più grande si avvicina al 50% dei match, il che indica molti match che non avevano dominio più rappresentativo sono stati etichettati come facenti parte del dominio “services”. Quello che invece traspare dal secondo caso è che 5 domini hanno catturato il 60% dei match circa, il che indica che vi è una maggior capacità di rappresentazione del secondo metodo. I 13 domini meno popolari si dividono il restante 40% dei match.

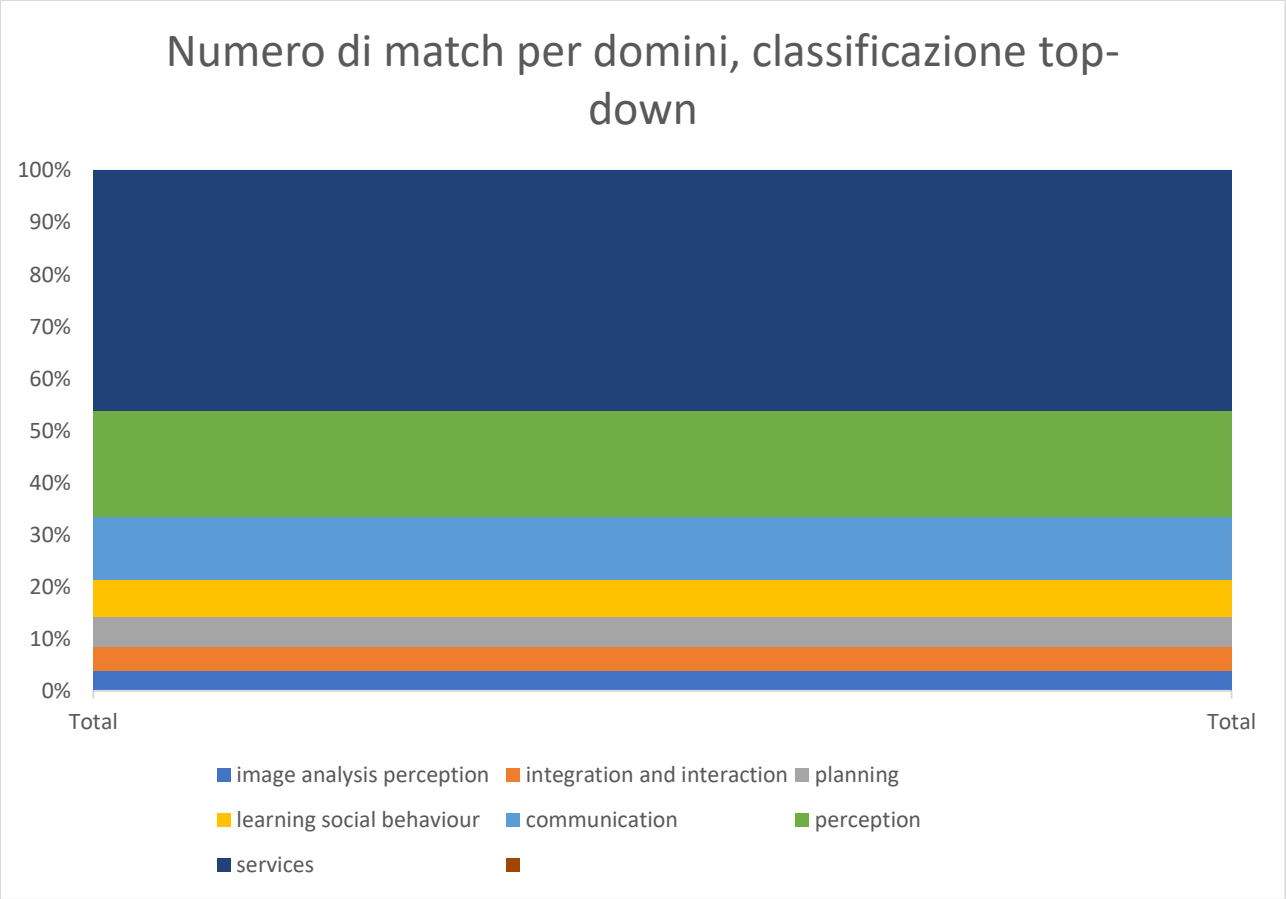


Grafico11: numero di match per ogni dominio della classificazione top-down

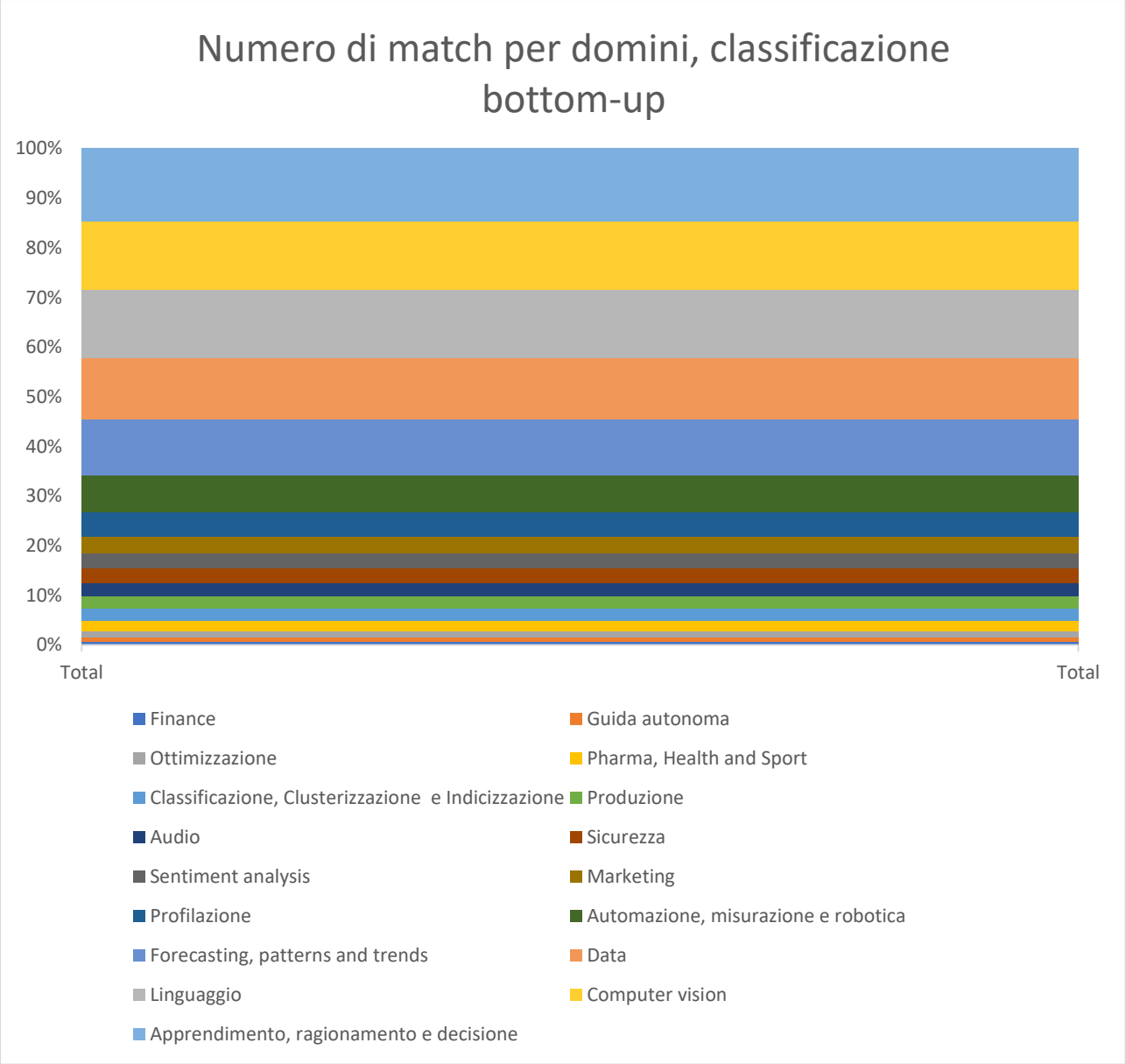


Grafico12: numero di match per ogni dominio della classificazione bottom-up

Il numero di parole della classificazione top-down è 73, oltre a quelle comunicate dalla Commissione Europea l'autore ha arricchito l'elenco con delle parole trovate nella letteratura di riferimento e in articoli del settore da fonti affidabili. Come nel caso dei domini è importante capire come queste parole abbiano svolto il loro compito di rappresentazione del mercato.

Nelle figure 3 e 4 usando delle cloudword sono state rappresentate le occorrenze delle keyword. La dimensione del carattere e l'intensità del colore di questo sono direttamente proporzionali

alla frequenza percentuale relativa della keyword rispetto l'insieme totale. Nel caso della top-down, figura 3, un osservatore può cogliere subito il fatto che 6 parole circa prevalgono in maniera netta su tutto il resto dell'insieme. Anche nel caso della bottom-up è presente un gruppo di keyword prevalenti, ma si tratta di un gruppo più ampio e soprattutto non vi è lo stesso distacco tra queste e il resto. Questo confronto fa mergere anche nel caso delle keyword una miglior rappresentazione da parte della tecnica bottom-up, vengono messi in evidenza un numero maggiore di parole chiave e il passaggio verso le meno rilevanti è più graduale.

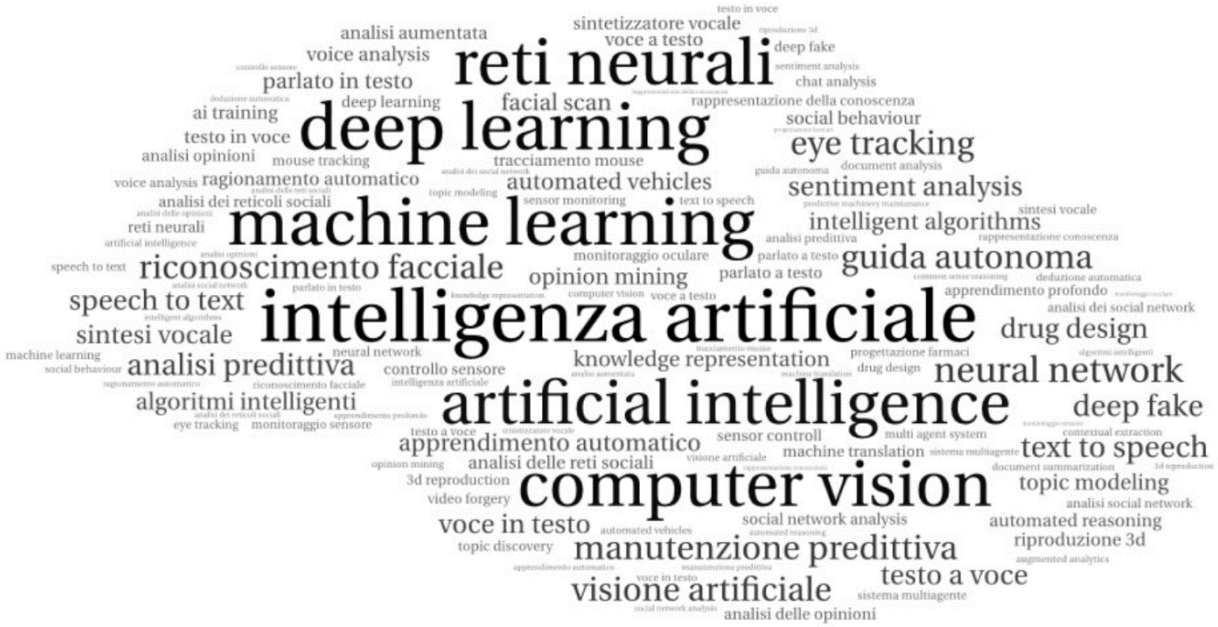


Figura 3: numero di match, creator e adopter per classificazione top-down

Grafico14: numero di match, creator e adopter per classificazione bottom-up

Conclusioni

Validità del framework di mapping delle startup di Intelligenza Artificiale

Studiando le relazioni tra la copertura di indirizzi web e l'anagrafica dell'azienda solo in due casi sono state individuate delle sistematicità che non si possono definire irrilevanti.

Il primo caso riguarda l'assenza di un indirizzo web per le aziende con un'età maggiore ai 4 anni. Queste presentano una percentuale di missing del sito web del 30%. Le ragioni alla base di questa sistematica mancanza possono essere diverse. Una riguarda il meccanismo di registrazione e aggiornamento dei dati da parte del Registro Imprese o di AIDA, è plausibile che non essendo stato indicato come dato obbligatorio nei primi anni dalle camere di commercio dove avveniva la registrazione nella Sezione Startup Innovative, questo manca per alcuni casi di startup più mature.

Il secondo caso riguarda l'assenza di un indirizzo web per le aziende provenienti dal Lazio o dalla Sicilia. Infatti, queste regioni presentano più del 20% di startup senza un sito web. Anche in questo caso le ragioni sono molteplici e probabilmente fuori dal controllo di chi gestisce la ricerca.

Questo framework è stato concepito facendo le seguenti assunzioni. La **prima ipotesi** alla base del lavoro consiste nel fatto che uno degli obiettivi delle startup è mettere in evidenza gli aspetti innovativi di quanto vendono. La **seconda ipotesi** che viene fatta ai fini della costruzione del framework è che i processi di innovazione e le eventuali partnership con forti innovatori vengano pubblicizzate in maniera esplicita. Dato che nell'82% dei casi un sito web è disponibile, è ragionevole considerare il primo assunto come valido. In secondo luogo, essendo state individuati 18 domini e un totale di 272 keyword, significa che le aziende effettivamente

compiono uno sforzo comunicativo al fine di poter esplicitare precisamente la loro value proposition.

Dato che i problemi riguardo la mancanza del sito web per alcuni casi è risolvibile, visto che le ipotesi alla base sono rispettate la metodologia descritta in questo elaborato di laurea può essere ritenuta sensata e utile. Di seguito vengono identificate altre criticità e potenziali miglioramenti.

Miglior metodo di classificazione

Il capitolo dell'analisi dei risultai fornisce dei dati per poter comprendere quale delle due classificazioni rappresenti meglio la realtà delle Startup Innovative italiane nel mercato dell'intelligence artificiale. Bisogna valutare non solo la capacità rappresentativa, bensì anche l'accessibilità delle informazioni e il costo di implementazione di una classificazione o l'altra. Non è infatti irrilevante, nella procedura bottom-up, la fase di selezione delle parole rilevanti dall'insieme totale delle parole utilizzate dai siti web di partenza. E' costosa in quanto richiede un operatore impegnato per un tempo che cresce più che proporzionalmente al crescere dei siti da ispezionare. Infatti, anche se l'operatore usufruisce dell'ausilio di software di analisi qualitativa dei dati come NVivo, il tempo di lettura e interpretazione delle frasi richiede un tempo minimo per poter generare la base del coding (individuazione delle parole rilevanti) che poi viene automatizzato dal software. In termini di tempo la classificazione bottom-up ha richiesto diversi giorni di lavoro.

Il pool iniziale di startup, almeno per la prima esecuzione della bottom-up, è tratto dai risultati della top-down eseguita nei giorni precedenti. Non vi fosse stata questa sarebbe stato necessario individuare un'altra fonte che indicasse un insieme iniziale, probabilmente un ulteriore lavoro di ricerca manuale. Il che comporta una certa dipendenza della bottom-up dalla top-down, per una prima esecuzione.

Le analisi dei risultati ottenuti con le due classificazioni parlano chiaro: la bottom-up restituisce un mapping di qualità maggiore, più vicino alla realtà. In primo luogo, sono stati individuati il 38% di adopters in più. Inoltre, il numero di domini emerso è maggiore e questi sono più rappresentativi, mostrando una distribuzione meno concentrata in pochi casi. Le relazioni tra domini, quindi i diversi modi di vendere servizi di una singola osservazione, vengono rappresentati in maniera più granulare. Questo viene confermato anche dal confronto degli insiemi di keyword.

Pertanto, considerando i punti sopraelencati, la bottom-up offre una maggior qualità. Ma probabilmente il metodo top-down può ancora essere utile in casi di necessità di esecuzioni più rapide, o in casi in cui la bottom-up non abbia ancora una base di startup da cui partire.

Criticità del framework proposto e potenziali miglioramenti

Per le aziende che non presentano un sito web in AIDA o nel Registro Imprese è possibile effettuare una procedura manuale di ricerca tramite motore di ricerca, usando come keyword il codice fiscale da solo, la denominazione da sola, o una combinazione dei due dati.

I siti web presenti in AIDA o nel Registro Imprese ma non raggiungibili nel momento in cui il web crawler ha richiesto le pagine ai server possono essere inaccessibili per diversi motivi. Il più frequente è che il sito non sia più pubblico o non lo sia mai stato se non per una pagina bianca all'unico scopo di occuparne il dominio. Nel caso in cui invece il motivo dell'inacessibilità sia una restaurazione o un guasto momentaneo, il problema verrebbe risolto nel momento in cui la procedura venisse ripetuta in un periodo di aggiornamento successivo.

Il lavoro manuale è al centro di entrambe le procedure presentate. Un operatore che analizzi i match e che indichi il dominio e l'attività di riferimento per la startup è imprescindibile

attualmente. Questo lavoro può essere sostituito da un algoritmo intelligente che effettui questa task di classificazione in maniera autonoma, più veloce, più precisa e tracciabile. Il lavoro svolto manualmente non verrebbe buttato, al contrario si tratterebbe di una base di dati fondamentale per fare training di algoritmi di machine learning che possono visitare ed etichettare autonomamente le aziende.

All'interno della procedura bottom-up vi è un ulteriore task manuale sostituibile: l'individuazione delle parole rilevanti dall'insieme totale di parole utilizzate nell'insieme di startup di partenza. Un algoritmo di reti neurali allenato al fine di individuare il significato dei testi e il contesto potrebbe svolgere il lavoro.

I grandi cluster della bottom-up possono essere ulteriormente suddivisi in altri cluster più rappresentativi. Questo è fattibile svolgendo un'analisi sui singoli cluster emersi dalla procedura bottom-up, cercando di individuare sottogruppi di parole identificabili con una particolare tecnologia o trend.

Espansioni e utilizzi

Una espansione interessante consisterebbe nell'aggiungere un modulo in grado di effettuare operazioni di crawling sui social, in particolare Linkedin che essendo in un periodo di crescente popolarità può diventare un punto di riferimento per le startup quando si tratta di gestire le relazioni con altre aziende o potenziali lavoratori. Si avrebbe un fonte di dati in più parecchio interessante e complementare a quella già presente dei database di Registro Imprese e Aida.

In questo studio l'autore ha proposto un framework focalizzato sull'individuazione delle startup coinvolte nell'implementazione o creazione di tecnologie di intelligenza artificiale. Estendere questa ricerca ad altri driver tecnologici e innovativi sarebbe un palese apporto di valore. Degli esempi su tutti possono essere tratti dagli obiettivi che la Commissione Europea si è preposta per il cosiddetto "decennio digitale". Il primo riguarda il design e la produzione

di processori e semiconduttori di nuova generazione con l'intenzione di arrivare nella seconda metà degli anni '20 a poter reggere la richiesta di potenza computazionale. Strettamente connessa a questo primo obiettivo vi sono anche le strategie annunciate dalla CE riguardo le super computing technologies e più in particolare il quantum computing (Commissione_Europea, 2020).

Nel dicembre 2012 InfoCamere, in accordo al DECRETO-LEGGE 18 ottobre 2012, n. 179 "Ulteriori misure urgenti per la crescita del Paese", alle aziende iscritte al Registro Imprese ha aperto la possibilità di iscriversi alla Sezione Startup Innovative se rispettano determinati requisiti in termini di dimensione, età e organico. Tale politica ha riscosso successo, pertanto a gennaio 2015 è stata ripetuta un'apertura analoga (DECRETO-LEGGE 24 gennaio 2015, n. 3 Misure urgenti per il sistema bancario e gli investimenti) in questo caso per le PMI Innovative. In AIDA sono disponibili i dati su entrambi i tipi di aziende, inoltre il decorso più frequente della Startup Innovativa è proprio diventare PMI Innovativa. Avere un mapping di questi due insiemi permetterebbe, congiuntamente all'allargamento dei temi trattati, di avere una visione costantemente aggiornata sulle piccole realtà innovative del territorio italiano. Per avere una visione completa il passo successivo potrebbe essere quello di inserire nello studio il resto delle società che sono presenti nel database di AIDA e Registro Imprese, avendo così una visione completa delle attività innovative del tessuto economico italiano.

Estendere la ricerca oltre i confini italiani richiederebbe in prima istanza l'accesso ad una fonte di dati economico finanziari riguardo aziende estere. Inoltre, essendovi quadri normativi diversi sarebbe necessario uno studio di questi. In letteratura per la Germania qualcuno ha già proposto un framework di classificazione dell'intero tessuto economico tedesco o per alcune aziende di Berlino [Kinne 2020, Horne 2020]. Le maggiori criticità per un tentativo di espansione del progetto verso l'estero potrebbe essere la lingua locale. Questo problema verrebbe meno

attraverso l’impiego di un modulo di classificazione basato su reti neurali invece che effettuato manualmente da un operatore.

Appendice

Elenco keyword classificazione top-down tratte da documento Commissione Europea

Dominio	Keyword	Fonte
Images analysis PERCEPTION	3D REPRODUCTION (3D recostruction)	(Samoili et al., 2020)
	FACIAL SCAN	(Samoili et al., 2020)
	Photo editing	(Samoili et al., 2020)
	Eye-tracking e mouse tracking	(Samoili et al., 2020)
	Audio processing	(Samoili et al., 2020)
	COMPUTER VISION	(Samoili et al., 2020)
LEARNING social behaviour	*SOCIAL BEHAVIOUR*	(Samoili et al., 2020)
SERVICES	*WEB VULNERABILITY*	(Samoili et al., 2020)
	AI TRAINING	(Samoili et al., 2020)
	SENSOR MONITORING	(Samoili et al., 2020)
	CONSULTING	(Samoili et al., 2020)
	AUGMENTED ANALYTICS	(Samoili et al., 2020)
	DRUG DESIGN	(Samoili et al., 2020)
	PREDICTIVE MACHINERY MAINTANANCE	(Samoili et al., 2020)
COMMUNICATION	*VOICE ANALYSIS*	(Samoili et al., 2020)
	DOCUMENT ANALYSIS	(Samoili et al., 2020)
	NLP	(Samoili et al., 2020)
	TOPIC DISCOVERY AND MODELING	(Samoili et al., 2020)
	CONTEXTUAL EXTRACTION	(Samoili et al., 2020)
	SENTIMENT ANALYSIS	(Samoili et al., 2020)
	SPEECH-TO-TEXT AND TEXT-TO-SPEECH CONVERSION	(Samoili et al., 2020)
	DOCUMENT SUMMARIZATION	(Samoili et al., 2020)
	MACHINE TRANSLATION	(Samoili et al., 2020)
	CHAT ANALYSIS	(Samoili et al., 2020)
REASONING	*Knowledge representation*	(Samoili et al., 2020)
	Common sense reasoning	(Samoili et al., 2020)
	Automated reasoning	(Samoili et al., 2020)
PLANNING	*Planning and scheduling*	(Samoili et al., 2020)
	Searching	(Samoili et al., 2020)
	Optimisation	(Samoili et al., 2020)
INTEGRATION and INTERACTION	*Multi-agent system*	(Samoili et al., 2020)
	Robotics and Automation	(Samoili et al., 2020)
	Connected and Automated vehicles	(Samoili et al., 2020)

Elenco keyword classificazione top-down trovate in letteratura ed enciclopedie online

Dominio	Keyword	Fonte
Images analysis PERCEPTION	riproduzione 3d	Traduzione da Commissione Europea
	riconoscimento facciale	Traduzione da Commissione Europea
	Deep fake, video forgery	(Wikipedia, 2021)
	monitoraggio oculare e tracciamento mouse	Traduzione da Commissione Europea
	voce a testo, voce in testo, parlato in testo, parlato a testo, testo a voce, testo in voce	Traduzione da Commissione Europea
	sintesi vocale, sintetizzatore vocale	(Wikipedia, 2021)
	visione artificiale	Traduzione da Commissione Europea
LEARNING social behaviour	*Social network analysis, analisi social network, analisi delle opinioni , analisi dei reticoli sociali*	(Wikipedia, 2021)
SERVICES	*Neural network, deep learning, machine learning*	(Wikipedia, 2021)
	Intelligent algorithm	(Wikipedia, 2021)
	Sensor controll	(Wikipedia, 2021)
	Analisi predittiva	Traduzione da Commissione Europea
	Manutenzione predittiva	Traduzione da Commissione Europea
	intelligenza artificiale, algoritmi intelligenti, reti neurali	(Wikipedia, 2021)
	analisi aumentata, progettazione farmaci	Traduzione da Commissione Europea
	controllo sensore, monitoraggio sensore	Traduzione da Commissione Europea
COMMUNICATION	*Opinion mining*	(Wikipedia, 2021)
REASONING	*apprendimento automatico*	Traduzione da Commissione Europea
	apprendimento profondo	Traduzione da Commissione Europea
	rappresentazione della conoscenza, rappresentazione conoscenza, ragionamento automatico, deduzione automatica	(Wikipedia, 2021)
INTEGRATION and INTERACTION	*sistema multiagente*	Traduzione da Commissione Europea
	guida autonoma	Traduzione da Commissione Europea

Bibliografia

Arora. (2013). Entry strategies in an emerging technology: A pilot web-based study on graphene firms. *Scientometrics*.

Askitas. (2015). The Internet as a data source for advancement in social sciences. *International Journal of Manpower*.

Beaudry. (2016). Validation of a web mining technique to measure innovation in high technology Canadian industries. . *CARMA 2016–1st International Conference on Advanced Research Methods and Analytics*.

Blank. (2010). *Not All Those who Wander are Lost: Posts from an Entrepreneurial Career.* Cafepress.com.

Calvino, F., Criscuolo, C., & Menon, C. (2016). No country for Young Firms?: Start-up Dynamics and National Policies. *OECD Science, Technology and Industry Policy Papers*.

Carmel. (1994). Time-to-completion in software package startups,. *Proceedings of the Twenty-Seventh Hawaii International Conference on System Sciences*.

Cefis. (2005). A Matter of Life and Death: Innovation and Firm Survival. *Industrial and Corporate Change*.

Commisisone_Europea. (2016). *isoc_ciweb.* Tratto da Commissione Europea: https://ec.europa.eu/eurostat/statistics-explained/index.php/Digital_economy_and_society_statistics_-_enterprises#Access_and_use_of_the_internet

Commissione_Europea. (2020). *decade digitale.* Tratto da Commissione Europea: https://ec.europa.eu/digital-single-market/en/news/member-states-join-forces-european-initiative-processors-and-semiconductor-technologies

Criscuolo, C., Gal, P., & Menon, C. (2014). The Dynamics of Employment Growth: New Evidence from 18 Countries. *OECD Science, Technology and Industry Policy Papers*.

Gök. (2015). Use of web mining in studying innovation. *Scientometrics*.

Henderson, M. (1993). Underinvestment and Incompetence as responses to Radical Innovation: Evidence from the Photolithographic Alignment Equipment Industry. *RAND Journal of Economics*.

Isenberg, D. (2011). *The Entrepreneurship Ecosystem Strategy as a New Paradigm for Economic Policy: Principles for Cultivating Entrepreneurship.* The Babson Entrepreneurship Ecosystem Project.

Kim. (2012). Technology trends analysis and forecasting application based on decision tree and statistical feature analysis. *Expert Systems with Applications*.

Kleinknecht. (2002). *The non-trivial choice between innovation indicators.* Economics of Innovation and New Technology.

Krizhevsky. (2012). ImageNet Classification with Deep ConvolutionalNeural Networks. *ImageNet LSVRC-2010*.

Matriciano. (2020). The effect of R&D investments, highly skilled employees, and patents on the performance of Italian innovative startups. *Technology Analysis & Strategic Management*.

Matriciano, D. (2020). The effect of R&D investments, highly skilled employees, and patents on the performance of Italian innovative startups . *Technology Analysis & Strategic Management*.

Nagoaka. (2010). Patent Stastics as an Innovation Indicator. *Handbook of Economics of Innovation*.

Nathan. (2017). Innovative Events . *(No. 429; Centro Studi Luca d'Agliano Development Studies Working Paper)*.

OECD. (2009). *Patent statistics manual.* OECD.

Osservatorio_Artificial_Intelligence. (2019). *ARTIFICIAL INTELLIGENCE: UNA PRIMA FOTOGRAFIA DEL MERCATO ITALIANO.* Politecnico di Milano.

Squicciarini. (2013). *Measuring patent quality.* OECD Science.

Turing. (1950). Computing machinery and intelligence. *Mind*.

Wikipedia. (2021). Tratto da Wikipedia.

Youtie. (2012). Pathways from discovery to commercialisation: Using web sources to track small and medium-sized enterprise strategies in emerging nanotechnologies. . *Technology Analysis & Strategic Management*.