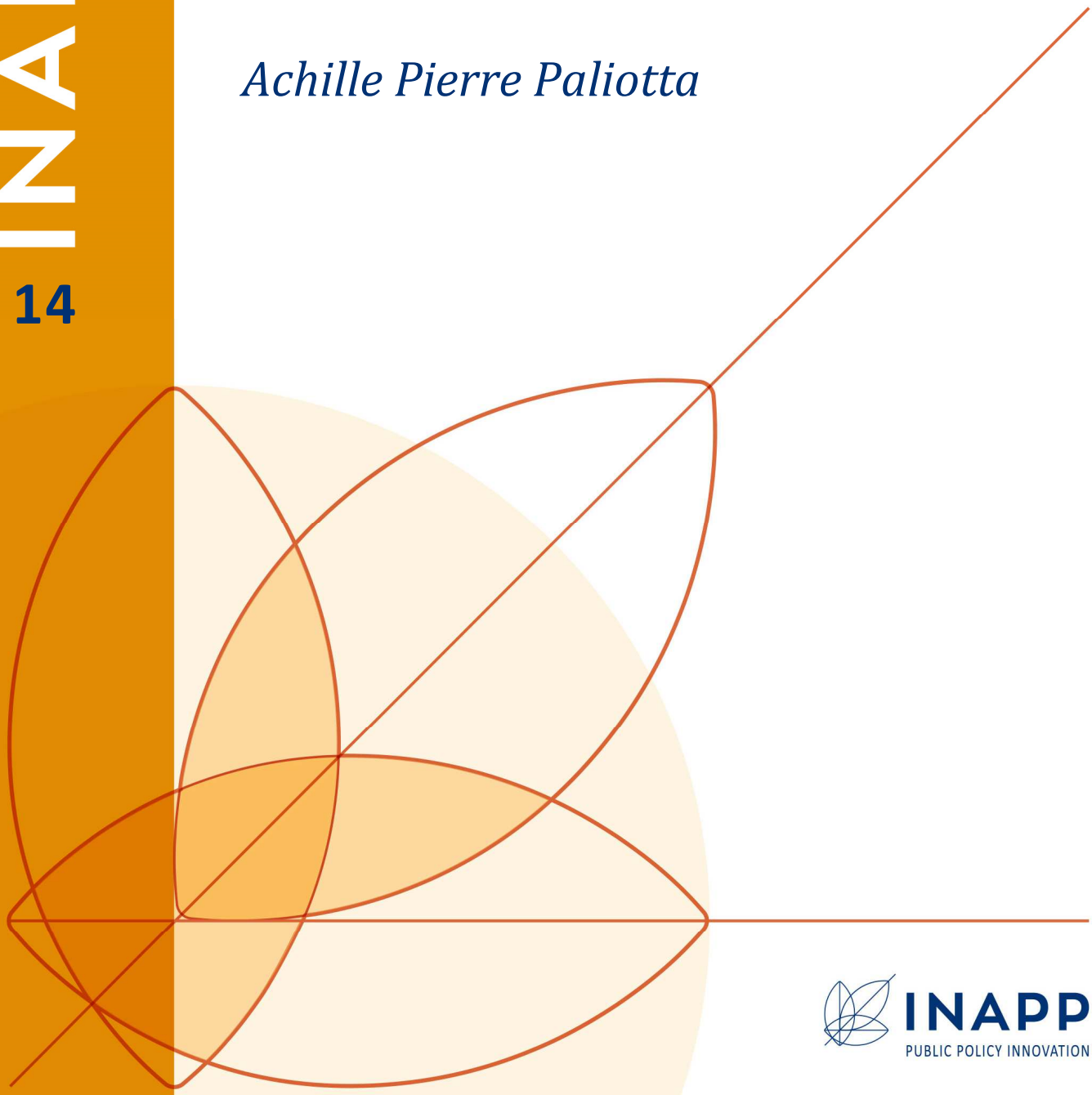
INAPP PAPER

14

LA WEB NEWS COVERAGE DI INDUSTRIA 4.0 IN ITALIA E GERMANIA

Achille Pierre Paliotta



INAPP
PUBLIC POLICY INNOVATION

L'Istituto Nazionale per l'Analisi delle Politiche Pubbliche (INAPP) è un ente pubblico di ricerca che si occupa di analisi, monitoraggio e valutazione delle politiche del lavoro, delle politiche dell'istruzione e della formazione, delle politiche sociali e, in generale, di tutte le politiche economiche che hanno effetti sul mercato del lavoro.

Nato il 1° dicembre 2016 a seguito della trasformazione dell'Isfol e vigilato dal Ministero del Lavoro e delle politiche sociali, l'Ente ha un ruolo strategico - stabilito dal Decreto Legislativo 14 settembre 2015, n. 150 - nel nuovo sistema di governance delle politiche sociali e del lavoro del Paese.

Inapp fa parte del Sistema statistico nazionale (SISTAN) e collabora con le istituzioni europee. Da gennaio 2018 è Organismo Intermedio del PON Sistemi di Politiche Attive per l'Occupazione (SPAO) per svolgere attività di assistenza metodologica e scientifica per le azioni di sistema del Fondo sociale europeo ed è Agenzia nazionale del programma comunitario Erasmus+ per l'ambito istruzione e formazione professionale. È l'ente nazionale all'interno del consorzio europeo ERIC-ESS che conduce l'indagine European Social Survey.

Riferimenti
Corso d'Italia, 33
00198 Roma
Tel. +39.06.85447.1
web: www.inapp.org

Contatti: editoria@inapp.org

La collana Inapp Paper è a cura di Claudio Bensi.

Il presente lavoro si inserisce all'interno di un progetto generale che ha l'obiettivo di produrre informazioni e strumenti finalizzati alla riduzione del mismatch tra domanda e offerta di lavoro.

Questo testo è stato sottoposto con esito favorevole al processo di peer review interna curato dal Comitato tecnico scientifico dell'Istituto.

Autore
Achille Pierre Paliotta

Testo chiuso: novembre 2018
Pubblicato: dicembre 2018

Coordinamento editoriale
Costanza Romano

Editing grafico ed impaginazione
Valentina Orienti

Le opinioni espresse in questo lavoro impegnano la responsabilità degli autori e non necessariamente riflettono la posizione dell'ente.

ISSN 2533-2996
ISBN 978-88-543-0169-6

ABSTRACT

LA WEB NEWS COVERAGE DI INDUSTRIA 4.0 IN ITALIA E GERMANIA

L'economia attuale è sempre più interconnessa con le nuove tecnologie grazie alla diffusione pervasiva delle macchine intelligenti. In tale contesto, nel 2011, si è sviluppato in Germania il fenomeno conosciuto come Industria 4.0. Tale Paese rappresenta, pertanto, il naturale benchmark con cui confrontare la situazione italiana. Il presente lavoro è uno studio comparativo della pubblica opinione, di tipo qualificato (news e blog) dei due paesi. Mediante la web news coverage è possibile fornire una visione esploratoria e riassuntiva (sicuramente sommaria ma computazionalmente potente) di quello che è oggetto di discussione in rete. Il raffronto tra i due paesi ha messo in luce che la situazione italiana, nel periodo marzo-aprile 2018, a differenza di quella tedesca, appare troppo legata all'attualità e ad eventi contingenti. Dall'analisi testuale degli articoli pubblicati (soprattutto news), difatti, non sempre si riescono a cogliere appieno gli aspetti tecnologici di profonda evoluzione di oggetti e strumenti connessi all'attività manifatturiera digitalizzata e alla fabbrica intelligente.

PAROLE CHIAVE: industria 4.0, impresa 4.0, internet of things, text mining, web scraping

THE WEB NEWS COVERAGE OF INDUSTRY 4.0 IN ITALY AND GERMANY

The current economy is increasingly connected with new technologies due to the spread of smart objects (IoT). In 2011 Industry 4.0 was developed in Germany. Therefore this country represents the natural benchmark to compare with the Italian situation. This work is a comparative study of expert opinion (news and blogs) of the two countries. Through the web news coverage it is possible to provide an exploratory and summarizing view (computatively very powerful) of what is the subject of discussion on the web. The comparison between the two countries has shown that the Italian situation, in the period March-April 2018, unlike the German one, seems related to current events too much.

KEYWORDS: industry 4.0, smart factory, internet of things, text mining, web scraping

PER CITARE IL PAPER: Paliotta A. P. (2018), La web news coverage di Industria 4.0 in Italia e Germania, Inapp Paper n. 14, Roma, INAPP

INDICE

INTRODUZIONE

L'economia attuale è sempre più intrecciata con le nuove tecnologie grazie alla diffusione pervasiva delle macchine intelligenti (Brynjolfsson e McAfee 2011) e si sostiene, da più parti, che si è oramai all'alba di una nuova rivoluzione industriale, la quarta (Industria 4.0).
Il processo di digitalizzazione è vieppiù caratterizzato dall'uso crescente di sensori dell'internet delle cose (*internet of things*, IoT), dai big data (Lee et al. 2014; Kitchin 2014) e dall'intelligenza artificiale (AI). Ciò consente di connettere persone, prodotti e servizi, nonché interi impianti di produzione, con i loro specifici processi interni ed esterni, tramite il web. In questo contesto i confini tra il mondo reale e quello digitale diventano sempre più sfumati e i dati vengono generati, raccolti e analizzati su vasta scala. Dopo una fase di analisi e interpretazione, questi dati costituiscono la base per implementare dei servizi intelligenti: servizi basati su dati che integrano prodotti fisici e consentono un elevato livello di personalizzazione in base alle specifiche esigenze e aspettative dei clienti. Con Industrie 4.0[1], a partire dal 2011, la Germania ha compiuto un passo significativo (Acatech 2013) verso la trasformazione digitale, nel settore dell'automazione della produzione, con il concetto di fabbrica intelligente ed è stata seguita da altri paesi, tra cui l'Italia, seppur con un certo ritardo.
Con il termine Industria 4.0, seppur generico, si intendono moltissime cose quali il disegno e la modellistica degli impianti (Zezulka et al. 2016), l'automazione snella (Kolberg e Zühlke 2015), la produttività collaborativa (Schuh et al. 2014), la produzione intelligente (Zhong et al. 2017), la produzione intelligente e integrata (Chen 2017). L'elemento centrale è, comunque, la connessione di internet e dei computer con il mondo fisico degli oggetti attraverso l'uso dei sistemi cyber-fisici (Lee et al. 2015). I nuovi sistemi così creati sono in grado di controllare, ottimizzare e configurare in modo autonomo i propri parametri di funzionamento.

Grazie alla posizione di guida, detenuta in questo settore industriale, la Germania rappresenta il naturale paese di riferimento con cui confrontare la situazione italiana. «Il modello tedesco può quindi sicuramente rappresentare un benchmark utile, visto che la struttura produttiva italiana è abbastanza simile a quella tedesca, in particolare per quanto riguarda il peso delle esportazioni manifatturiere. La grande differenza sta, come è noto, nella diversa dimensione media delle aziende, molto inferiore nel nostro paese» (Ballarino e Checchi 2013, 9). La struttura economica, basata sul manifatturiero, costituisce dunque l'ossatura portante dei due paesi, nonché l'ambito produttivo prioritario in cui implementare Industria 4.0, ragion per cui può essere utile mettere a raffronto i due paesi in quanto l'uno si trova in una fase più avanzata dell'altro, seppur in un quadro relativo alla sfera sociale e culturale, la blogosfera (Herring et al. 2005).

[1] Henning Kagermann, Wolf-Dieter Lukas e Wolfgang Wahlster coniarono il termine, in una comunicazione, tenuta alla Fiera di Hannover del 2011, in cui preannunciarono lo Zukunftsprojekt Industrie 4.0, riferito all'*internet of things* (*Internet der Dinge*, in tedesco) in ambito manifatturiero, goo.gl/TpejrW.

La tesi di fondo è, pertanto, quella di approfondire il framework culturale dei due paesi, entro cui si colloca un fenomeno innovativo qual è Industria 4.0, vale a dire una sorta di studio comparativo della pubblica opinione di tipo qualificato (news e blog). Lo sviluppo delle nuove tecnologie ha determinato la convergenza dei media e degli attori tradizionali verso la rete internet e i social networks[2] e ciò significa che tale copertura (*web news coverage*) è sempre più rilevante. La complessa relazione che unisce opinione pubblica, news media e decisori pubblici trova, infine, uno dei luoghi precipui di condensazione nei siti giornalistici e nei blog. Le agenzie dei media profondono sempre più risorse nell'implementazione digitale dei loro contenuti così, come i governi e gli enti istituzionali utilizzano la rete per promuovere politiche pubbliche, disseminare informazioni e favorire atteggiamenti coerenti (*nudging*) con tematiche di interesse collettivo. In questo senso il tool che qui si propone, può essere utile per monitorare quanto e come un determinato fenomeno possa essere recepito dalla pubblica opinione e possa sedimentarsi nella blogosfera.

In definitiva, le variabili culturali, nello specifico quella comunicativa, se non influiscono direttamente su quelle strutturali ed economiche, possono però contribuire a creare un ambiente sociale favorevole allo sviluppo di Industria 4.0, soprattutto riguardo al sistema dell'istruzione e della formazione, al fine di promuovere nei giovani competenze tecnico-specialistiche e un appropriato *mindset* così come richiesto dal sistema produttivo nazionale. Del resto, la rimodulazione del piano statale Industria 4.0 verso Impresa 4.0 segnala proprio questa difficoltà generale del Paese nel reperire il capitale umano qualificato adatto a gestire una fase di passaggio forse epocale. Tale stringente necessità viene riconosciuta non solo a livello nazionale[3] ma anche internazionale (Oecd 2017).

In ultimo, l'obiettivo conoscitivo, di mettere a raffronto le opinioni pubbliche di diversi paesi, è stato favorito, nel corso degli ultimi anni, dallo sviluppo progressivo di una serie di fattori politici, sociali e culturali (Norris 2007)[4]. Sono oramai disponibili, difatti, indagini complesse, svolte ad esempio a livello europeo, quali l'Eurobarometro[5] e la European Social Survey[6]; esse utilizzano, nel campo della ricerca sociale, metodi consolidati di raccolta dati mentre in questo studio di caso ci si è avvalsi di una

[2] In questa sede non è stato preso in considerazione il genere dei tweets sostanzialmente per due ragioni. La prima, è dovuta alla limitazione di caratteri per cui anche se Twitter accetta stringhe più lunghe di testo, tali messaggi vengono troncati. La seconda, è che Twitter, nel corso degli anni, ha accresciuto le misure di restrizione, all'accesso dei dati, cosicché molti tools di estrazione dati non sono attualmente più disponibili. Non si sono presi in considerazione, infine, i post diffusi su Facebook per le stringenti policies aziendali, che di fatto impediscono l'accesso a questa peculiare fonte di conoscenza sociale.

[3] Alfonso Fuggetta, Industria 4.0, il circolo vizioso tra competenze e domanda di lavoro, 5 gennaio 2018, goo.gl/8m6WPK.

[4] «As the world has become more interconnected through globalization, the social sciences have been tugged in its wake. (…) The growth of electoral democracies has also probably facilitated the study of public opinion, since this development facilitates freedom of expression for conducting independent social surveys and publishing the results of the analysis, also triggering the demand for commercial market research companies and non-profit social science institutes, free from political interference and overt state censorship» (Norris 2007, 530)

[5] L'Eurobarometro, promosso dall'Unione europea nel 1974, indaga in profondità, per ogni stato membro, le motivazioni, i sentimenti e le reazioni di determinati gruppi sociali verso un determinato argomento o concetto, ascoltando e analizzando il loro modo di esprimersi in gruppi di discussione o con interviste non direttive. Ogni sondaggio consiste di circa 1.000 interviste faccia a faccia per paese membro, goo.gl/jANEf4.

[6] L'European Social Survey (ESS) è un'indagine transnazionale condotta in tutta Europa fin dalla sua istituzione nel 2001. Ogni due anni vengono condotte interviste faccia a faccia con campioni trasversali selezionati. L'indagine misura gli atteggiamenti, le convinzioni e i modelli di comportamento delle diverse popolazioni in più di trenta nazioni, goo.gl/bHoime. L'Inapp, "full member" del consorzio ESS, ha realizzato nel 2017 l'indagine italiana relativa al Round 8 così come nel 2019 parteciperà al Round 9.

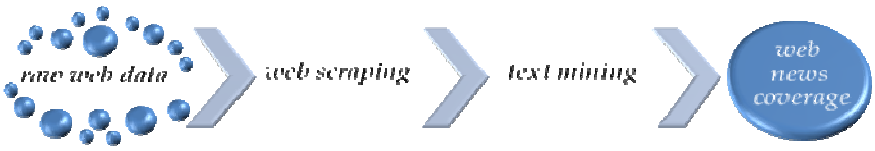
metodologia innovativa, l'estrazione dei dati dalla rete tramite *web scraping*[7] (Mitchell, 2015; Paliotta, 2016 e 2018a). Il vantaggio di poter far ricorso a questa metodologia, oltre a quelle del costo e della tempestività, è di poter effettuare una ricerca molto granulare dei fenomeni oggetto di studio (ricorso a parole chiave piuttosto che a complesse interviste) che non è possibile effettuare con i metodi tradizionali. Tutto ciò può spiegare, in sintesi, l'importanza che può annettersi all'estrazione delle informazioni dalla rete, riguardo a tematiche culturali, sociali e politiche (Rogers 2015; Paliotta 2018b)[8].
L'obiettivo di questo studio è, dunque, quello di rilevare come la tematica Industria 4.0 sia stata sviluppata nel web, sia in Italia che in Germania, e quali siano i temi a cui, oggigiorno, essa è strettamente correlata.
Il testo è organizzato nel modo seguente: nel primo paragrafo, di carattere metodologico, si descriveranno le principali fasi operative seguite nello svolgimento dell'analisi testuale, i successivi due paragrafi saranno dedicati alla presentazione dei risultati, dapprima mediante un'analisi lessicale-paradigmatica e, in seguito, una testuale-sintagmatica.

1 L'ESTRAZIONE DEI DATI DALLA RETE E IL LORO TRATTAMENTO AUTOMATIZZATO

Come argomentato nell'introduzione, le opinioni espresse in rete possono essere considerate una sorta di sintesi sociologica, vale a dire una conoscenza sociale digitalmente diffusa, in questo caso degli esperti (opinionisti e giornalisti), mediante la quale è possibile fornire una visione esploratoria e riassuntiva (sicuramente sommaria e perciò computazionalmente potente) di quello che è oggetto di discussione, hic et nunc, e che la rete necessariamente riverbera, in maniera più o meno fedele.
Per giungere a questa sintesi il percorso metodologico complessivo che è stato seguito si è basato sostanzialmente su tre fasi (figura 1): nella prima, si è approfondita la riflessione sui dati grezzi disponibili in rete; nella seconda, si è svolta l'attività di *web scraping*; nella terza, si sono predisposte le informazioni, estratte in rete, per la fase di *text mining*.

[7] Il *web scraping* è, in generale, una tecnica di raccolta dati on-line mediante appositi strumenti. «This is most commonly accomplished by writing an automated program that queries a web server, requests data (usually in the form of the HTML and other files that comprise web pages), and then parses that data to extract needed information» (Mitchell, 2015:VIII). A tutt'oggi, sono disponibili diversi strumenti per ottenere questi dati, almeno tre modalità principali, a seconda delle finalità conoscitive: 1) mettere a punto un *crawler* ad hoc; 2) avvalersi di *scraping tools* disponibili gratuitamente; 3) far ricorso ad un servizio commerciale di dati (data-as-a-service, DaaS). In questo caso è stata scelta la terza opzione come si vedrà meglio più avanti.
[8] «Nowadays the web is becoming a space for more than the study of on-line culture. Rather it has become a site to study a range of cultural and social issues» (Rogers, 2015:2).

Figura 1 Le fasi metodologiche seguite per l'estrazione dei web data fino alla web news coverage



Fonte: elaborazione Inapp

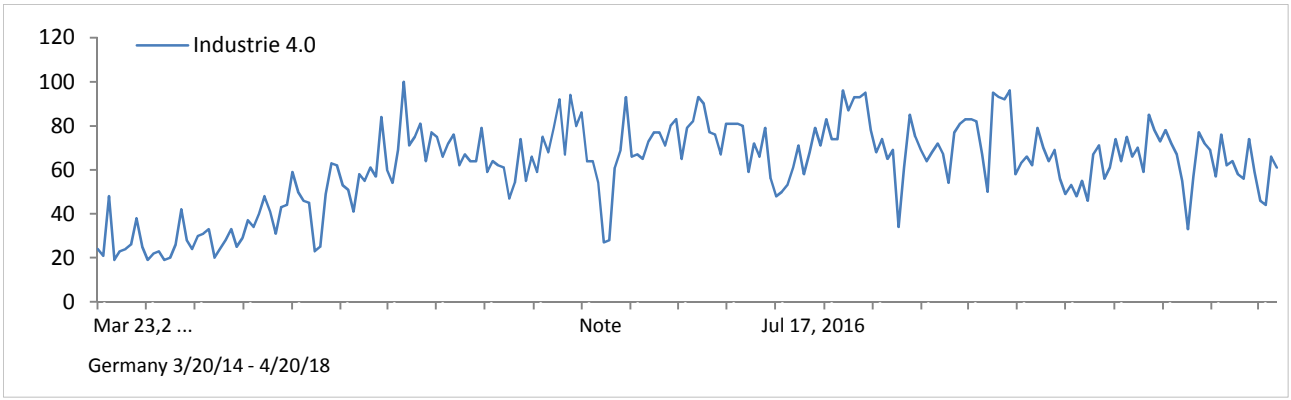
1.1 L'individuazione delle parole chiave

Durante la I fase si sono svolte delle indagini preliminari al fine di testare la diffusione on-line di alcune parole chiave relative al fenomeno oggetto di studio facendo uso delle applicazioni Google Trends e Google Search.

L'unità d'analisi investigata è il singolo messaggio testuale (post), con funzione di opinione o commento, il quale contiene nel titolo la parola chiave (industria 4.0; impresa 4.0; industrie 4.0), pubblicato nei blog e nei siti di news. A questo riguardo, si rileva che vi siano stretti rapporti di interdipendenza tra codice testuale e canale (internet) per cui le caratteristiche delle piattaforme tecnologiche, su cui vengono allocati i post, condizionano l'elaborazione degli stessi enunciati discorsivi

Google Trends (trends.google.com) permette di monitorare la dinamica della diffusione delle keywords che nel caso della Germania, data dall'aprile 2011 (figura 2), anno in cui fu coniato il termine in una comunicazione tenuta alla Fiera di Hannover e in Italia, dalla presentazione del cosiddetto "Piano Calenda", dal nome del ministro dello Sviluppo economico, Carlo Calenda, il 21 settembre 2016[9] (figura 3).

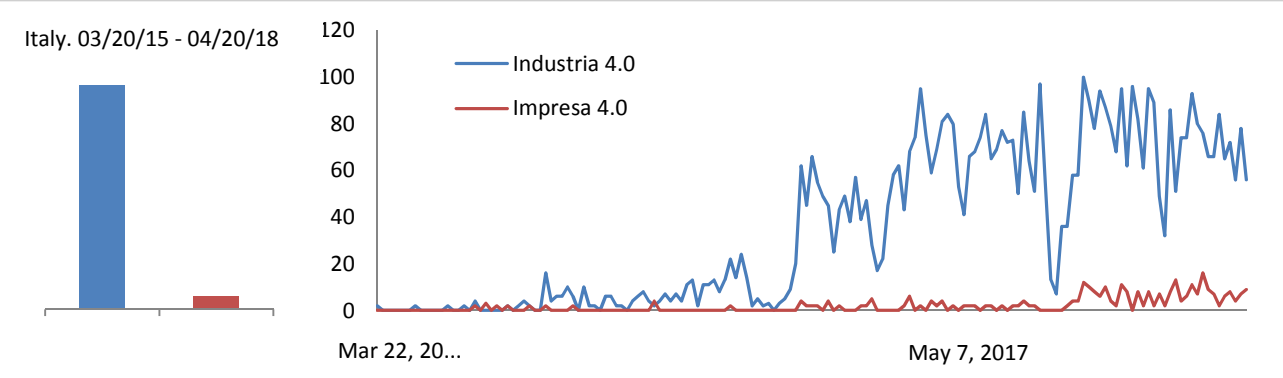
Figura 2 - La diffusione in rete del termine "industrie 4.0" sui siti tedeschi



Fonte: elaborazione Inapp su dati estratti dalle rete (https://trends.google.com)

[9] goo.gl/ydaNRX.

Figura 3 - La diffusione in rete dei termini "industria 4.0" e "impresa 4.0" sui siti italiani



Fonte: elaborazione Inapp su dati estratti dalla rete (https://trends.google.com)

Si tratta di ben cinque anni di differenza tra le due situazioni a raffronto e ciò trova ampie conferme anche mediante Google Search (google.com). Si è effettuata, difatti, una query ristretta, tra apici, in data 20 aprile 2018, la quale ha fornito i seguenti risultati:

- industrie 4.0: 4.180.000;
- industria 4.0: 162.000;
- impresa 4.0: 63.000.

1.2 L'attività di estrazione dei dati dalla rete

Per quel che riguarda la II fase, le tecniche di estrazione dei dati dal web, si è utilizzata la piattaforma di una società israeliana, Webhose.io, la quale fornisce un servizio avanzato di *data crawling API service*. Mediante tale strumento si può avere accesso a dati provenienti da centinaia di migliaia di fonti a livello globale. La scelta è parsa quanto mai appropriata, ai fini di questo studio, poiché permette di distinguere i dati riferiti ai siti di news e blog[10]. Le semplici funzionalità di query permettono di effettuare delle ricerche in base alle proprie esigenze ottenendo i risultati nei formati più comuni quali JSON, RSS, XML. In questo caso, il file in JSON è stato poi parserizzato in un file TXT mediante un piccolo programma scritto in C[11].

Sulla piattaforma Webhose.io si è effettuato un *data crawling* facendo uso di alcune parole chiave. Oltre alle preventivabili industria 4.0 e a impresa 4.0 (quest'ultima rappresenta, del resto, la fase due del piano nazionale)[12] si è ipotizzato che potessero fornire un contributo all'illustrazione del fenomeno, anche altre parole chiavi quali: competence center, innovazione digitale, hub digital innovation in quanto in grado di catturare altri aspetti della stessa tematica. L'estrazione dei dati è avvenuta il 17

[10] Webhose.io per cercare di soddisfare, il più possibile, i criteri di rappresentatività dei dati aggiunge continuamente nuovi siti (a seconda dei dati di traffico generati), evita di estrarre degli articoli duplicati (in base a una unique URL) nonché cerca di assicurare la consistenza dei dati mantenendo le fonti costanti all'inizio e alla fine di un determinato periodo di tempo e archiviando i post anche qualora questi venissero successivamente eliminati dai siti (*survivorship bias*).

[11] Il programma è stato messo a punto da Marco Claudio Andreozzi che qui si ringrazia per la disponibilità nonché per la competenza professionale dimostrata.

[12] Gaia Vendettuoli, Ecco i risultati del piano Industria 4.0 (che cambia nome). Calenda: "Adesso fase due", 20 settembre 2017, goo.gl/2bPKpu.

aprile 2018, mediante la ricerca della parola chiave tra apici, contenuta solo nel titolo dell'articolo, al fine di ottenere una maggiore focalizzazione sui contenuti di interesse. Sono stati selezionati gli articoli pubblicati durante l'arco di un mese, con il primo post del 19 marzo e l'ultimo del 16 aprile.
La query ha fornito i seguenti risultati:

- industrie 4.0: 364 post (siti in lingua tedesca, localizzati in Germania);
- industria 4.0: 165 post (siti in lingua italiana, localizzati in Italia);
- impresa 4.0: 28 post (siti in lingua italiana, localizzati in Italia).

La keyword "competence center", la quale ha fornito 16 risultati, non è stata, in seguito, analizzata a causa dell'esiguità dei post collazionati dalla rete mentre "hub digital innovation", "hub digitale" e "innovazione digitale" sono stati oggetti di ricerca ma, in seguito, esclusi in quanto rimandavano a universi discorsivi non strettamente inerenti al fenomeno qui indagato.

1.3 L'attività di analisi testuale

Per quel che concerne la III fase, giacché i dati estratti dalla rete sono di natura testuale diviene, di conseguenza, una scelta obbligata quella di far uso delle tecniche di *text mining,* una metodologia già ben consolidata in dottrina, ma in via di costante perfezionamento, dal punto di vista del software. Questa tecnica fa riferimento alla statistica testuale, nata intorno agli anni Settanta, e sviluppatasi, in maniera significativa alla fine degli anni Ottanta, quando i ricercatori francesi Ludovic Lebart e André Salem (1988), tra altri, definirono i confini della statistica testuale basata sull'analisi per forme grafiche e, in parallelo, svilupparono i primi programmi per l'analisi dei dati testuali.
L'oggetto di analisi complessivo, chiamato corpus, formato da tanti frammenti (post), viene descritto grazie a diverse rappresentazioni, ovvero una di tipo lessicale-paradigmatica, basata sul linguaggio utilizzato, e una di tipo testuale-sintagmatica, basata sul senso del corpus. Vi è un continuo dialogo tra il livello paradigmatico del lessico e quello sintagmatico di analisi del discorso e in questo processo si arriva ad una rappresentazione sia del contesto complessivo sia del senso, anche latente, espresso dalle informazioni contenute nel corpus.
In questo modo, il documento si conosce mediante la rappresentazione del lessico, vale a dire il vocabolario[13] utilizzato, e mediante la rappresentazione del testo, inteso come discorso di un locutore[14].
All'interno di un determinato corpus l'unità minima è chiamata occorrenza, ovvero ogni parola che appare in un corpus e viene chiamata anche *token* (Lebart e Salem 1994). Il numero totale delle occorrenze (*tokens*) determina la lunghezza o dimensione del corpus.

[13] Il vocabolario di un corpus può essere espresso in forme grafiche (*types*) le quali sono le parole tali e quali come sono scritte nel testo: un *type* rappresenta una "entrata" nel vocabolario del corpus.
[14] In questo senso, il corpus è «un qualsiasi insieme di informazioni (ovvero una base-dati) composto da uno o più testi, ciascuno dei quali è suddiviso in vari frammenti da considerarsi come unità di analisi. Tali frammenti sono costituiti in genere da proposizioni, ma anche da qualsiasi altra segmentazione del corpus – utile a coglierne il senso in esso contenuto – che possa considerarsi alla stregua di un insieme di "enunciati", caratterizzati cioè da un senso compiuto» (Bolasco 1995, 88).

Per la fase di *text mining* è stato utilizzato il software francese IRaMuTeQ (*Interface de R pour les analyses multidimensionnelles de textes et de questionnaires*)[15] il quale considera un testo come l'espressione di un punto di vista, generato da un locutore, cosicché quando il corpus è stato prodotto da diversi soggetti si possono comprendere questi diversi punti di vista. L'assunzione sottostante è che differenti modi di pensare producono narrazioni dissimili, con la generazione di specifici vocabolari, analizzando le co-occorrenze delle parole: il senso delle diverse forme di discorso può essere catturato dall'analisi delle parole che si trovano insieme nelle stesse sequenze.
Il programma non permette di testare ipotesi quanto piuttosto di esplorare e descrivere il documento in analisi; il suo punto di forza è, del resto, proprio questo: in un breve lasso di tempo si può avere un'iniziale investigazione di un pur voluminoso corpus di dati. Ciò detto il ruolo del ricercatore non viene per questo meno in quanto vi sono da effettuare molteplici interventi sul corpus testuale così come vanno, purtuttavia, evidenziati i limiti oggettivi di tali approcci che sono quelli di far uso di dati non strutturati, con informazione sparsa, il cui riscatto dall'ambiguità è da inserire sempre all'interno di una stretta interdipendenza tra testo e contesto.

2 L'ANALISI LESSICALE-PARADIGMATICA DEL LINGUAGGIO UTILIZZATO NEI DUE CORPORA

In questa parte del testo verrà svolta sostanzialmente un'analisi lessicale-paradigmatica la quale tende a fornire una rappresentazione del vocabolario, ossia del linguaggio utilizzato nel corpus. È un'analisi di tipo verticale in cui la rappresentazione del testo è fatta senza tener conto dello sviluppo del discorso ma solo enumerando le parole più comuni[16].
Il primo corpus estratto dalla rete, denominato Industria 4.0, consta di più di 87.000 occorrenze per cui può essere classificato come un corpus medio-grande mentre le dimensioni del file sono di 587 KB. Si tratta di 165 articoli in lingua italiana (di cui 90 post estratti da siti di news e 75 dai blog).
Il secondo corpus, Impresa 4.0, è composto da quasi 10.000 occorrenze e può essere ritenuto di piccole dimensioni, con un peso di 65 KB. In questo senso la ridotta dimensione numerica del corpus può fornire delle statistiche testuali troppo dipendenti da alcuni messaggi A livello di genere discorsivo esso è composto da 21 articoli di news e 7 estratti dai blog.
Il terzo corpus, denominato Industrie 4.0, è in lingua tedesca e consta di 364 post, di cui 275 news e 89 blog della dimensione di 1.344 KB; si tratta, dunque, di un corpus medio-grande.

[15] IRaMuTeQ è di fatto la versione open-source, con adattamenti, di Alceste (*Analyse des lexèmes co-occurents dans les énnoncés simples d'un texte*). Il programma è basato su molte rappresentazioni grafiche e ciò trova un riscontro nel fatto che «nella statistica testuale, le analisi basate sulle forme grafiche hanno il vantaggio di essere indipendenti dalla lingua. Si tratta di un approccio puramente formale che privilegia i segni (significanti) per arrivare al senso (in quanto insieme di significati) come rappresentazione del contenuto o del discorso» (Bolasco 2005, 21).
[16] Un ulteriore livello di analisi verticale, che qui non viene svolto, potrebbe riguardare lo studio delle parole vuote (connettivi, preposizioni, congiunzioni, determinanti, interiezioni), degli incipit di frase, della punteggiatura, della lunghezza e struttura della frase o altre analisi d'interesse più strettamente linguistico.

I corpora, estratti a metà aprile 2018 e risalenti fino a un mese prima, sono stati preparati per l'analisi successiva optando di distinguere i messaggi derivanti dai blog da quelli delle news in modo da verificare eventuali differenze tra generi discorsivi on-line.
Una volta terminate le fasi di preparazione dei tre corpora essi sono stati importati in IRaMuTeQ il quale ha fornito le seguenti statistiche generali (tabella 1): il numero di testi; le occorrenze (il numero totale delle parole); le forme grafiche presenti (*types*), ovvero le parole uniche; gli *hapax* (parole con una sola occorrenza, che ricorrono una sola volta nel vocabolario di un corpus)[17]; la media di occorrenza per testo: occorrenze/testi.

Tabella 1 - Statistiche riassuntive dei tre corpora on-line

	Industrie 4.0	Industria 4.0	Impresa 4.0
Numero di testi	**364**	**165**	**28**
Numero di occorrenze	176.966	87.586	9.797
Numero di forme	14.516	7.026	2.023
% di forme/occorrenze	8,20	8,2	20,64
Numero di hapax	**6.477**	**2.405**	**984**
% di hapax/occorrenze	3,66	2,75	10,04
% di hapax/forme	44,62	34,23	48,64
Media di occorrenze per testo	**486,17**	**530,82**	**349,89**

Fonte: elaborazione Inapp

La media di occorrenze per testo, vale a dire la lunghezza media dei singoli post, è maggiore nel caso di Industria 4.0 (530,82) e Industrie 4.0 (486,17) rispetto a Impresa 4.0 (349,89).
La percentuale delle forme (*types*) rispetto alle occorrenze è di 8,20% (Industrie 4.0), 8,02% (Industria 4.0) e di 20,64% (Impresa 4.0). Ciò vuol dire che nell'ultimo caso vengono utilizzate, in media, più forme grafiche uniche, un numero che potrebbe esprimere, con tutte le precauzioni del caso, una sorta di complessità del testo (Manning e Schütze 1999, 22). Nel caso di Impresa 4.0, però, questi valori potrebbero non essere del tutto significativi a causa della modesta dimensione numerica del corpus.
Da un veloce confronto tra generi discorsivi (blog vs. news) si vede (tabelle 2 e 3) come il numero dei post delle news è sempre maggiore rispetto a quelli estratti dai blog in tutti e tre i corpora e ciò è ancora più vero per Industrie 4.0.
La media di occorrenze per testo fa registrare una costante lunghezza media maggiore nel caso dei blog, per ciò che riguarda Industria 4.0 mentre nel caso tedesco non sembra esserci una sostanziale differenza tra news e blog, se non una maggiore "complessità" dei secondi (15,25%) rispetto ai primi (9,16%).

[17] In linea generale, poco più del 49% delle forme sono presenti con un solo valore ed esse vengono chiamate *hapax legomena*, dal greco "detto una volta sola". Questo perché le parole occorrono in maniera molto infrequente, più del 90% di esse sono presenti nei testi con 10 occorrenze o anche meno, così come circa il 12% del testo è formato da parole che occorrono 3 volte o meno (Manning e Schütze 1999, 23).

Tabella 2 - Statistiche riassuntive dei due generi discorsivi in Industrie 4.0

	Genere News	Genere Blog
Numero di testi	**275**	**89**
Numero di occorrenze	133.227	43.739
Numero di forme	12.205	6.673
% di forme/occorrenze	9,16	15,25
Numero di hapax	5.413	**2.942**
% di hapax/occorrenze	**4,06**	6,73
% di hapax/forme	44,35	44,09
Media di occorrenze per testo	**484,46**	**491,45**

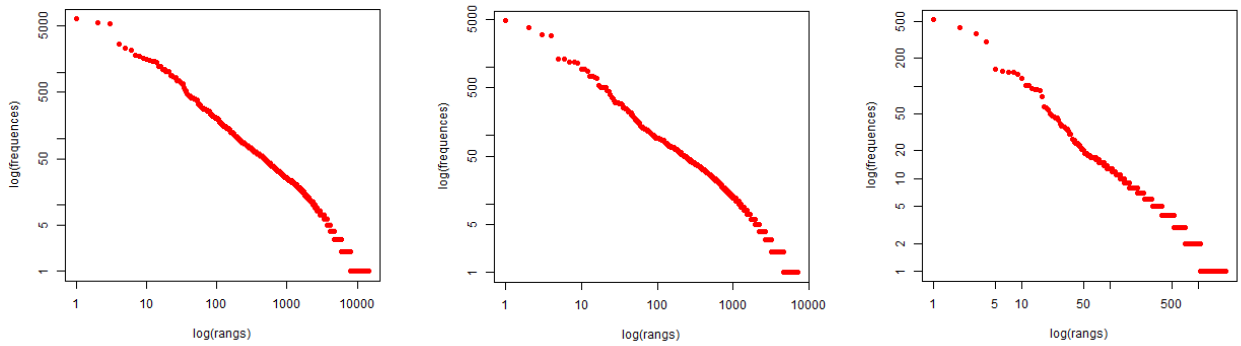
Fonte: elaborazione Inapp

Tabella 3 - Statistiche riassuntive dei due generi discorsivi in Industria 4.0 e Impresa 4.0

	Industria 4.0		Impresa 4.0	
	Genere News	Genere Blog	Genere News	Genere Blog
Numero di testi	**90**	**75**	**21**	**7**
Numero di occorrenze	38.723	48,863	8.166	1.631
Numero di forme	4.807	5.010	1.735	716
% di forme/occorrenze	12,41	10,25	21,24	43,89
Numero di hapax	**2.125**	**1.585**	**848**	**500**
% di hapax/occorrenze	5,49	3,24	10,34	30,66
% di hapax/forme	44,21	31,64	48,88	69,83
Media di occorrenze per testo	**430,26**	**651,51**	**388,86**	**233,00**

Fonte: elaborazione Inapp

Figura 4 - I grafici di Zipf dei tre corpora (Industrie 4.0, Industria 4.0 e Impresa 4.0, rispettivamente)



Fonte: elaborazione Inapp

Dai grafici di Zipf[18] (figura 4) si vede come la testa della distribuzione di Industrie 4.0 sia connotata da tre termini mentre negli altri due corpora vi sono quattro parole chiave; vi è, inoltre, una distanza marcata tra il primo termine e il secondo e tra il gruppo di testa e tutti i termini successivi. La coda della distribuzione è formata dagli *hapax* e si vede come essi siano maggiormente diffusi in Impresa 4.0 e in Industrie 4.0 e in, misura minore, in Industria 4.0.

[18] La legge di Zipf (Zipf 1949), dal nome del linguista e filologo statunitense George Kingsley (1902-1950), afferma che dato un corpus di enunciati in linguaggio naturale, la frequenza di ogni parola è inversamente proporzionale al suo rango in una tabella di frequenza. In buona sostanza, ci sono veramente poche parole comuni, un medio numero di parole con frequenza media e moltissime parole con una bassissima frequenza. Viene rappresentata mediante un grafico *log-log*, vale a dire su scala logaritmica.

3 L'ANALISI TESTUALE SINTAGMATICA DEL DISCORSO UTILIZZATO IN DUE CORPORA

La fase di analisi testuale riguarda tutte le operazioni svolte direttamente sul corpus, quindi in grado di fornire una rappresentazione sintagmatica del testo, sia puntualmente attraverso analisi di concordanze, più o meno sofisticate a seconda del tipo di query, sia globalmente mediante analisi di co-occorrenze. In generale, il fine dell'analisi testuale-sintagmatica è quella di individuare il senso del discorso. Vale qui sottolineare che il contenuto di un discorso è sempre espressione di un contesto, inteso come l'universo concettuale di riferimento.

I primi trattamenti dei corpora, in IRaMuTeQ, hanno seguito le seguenti fasi: segmentazione, riconoscimento e lemmatizzazione delle forme grafiche, scomposizione in unità di contesto elementari. Dopo le statistiche generali sono state prodotte le «forme attive», ovvero quell'insieme di parole atte a rappresentare un «mondo lessicale stabilizzato» (Reinert 2008, 983). Le «forme attive» più frequenti, di solito, stanno ad indicare quanto significative siano le stesse, in un documento testuale, e come possano essere considerate un buon descrittore dello stesso[19].

Basandosi sulle «forme attive», IRaMuTeQ permette l'utilizzo di sofisticate tecniche le quali cercano di rispondere tutte su come semplificare una moltitudine di informazioni, per mettere ordine nei dati che si vogliono trattare, e ciò avviene, generalmente, mediante il ricorso a delle forme grafiche. Tra tutte quelle disponibili si è scelto di presentare qui l'analisi delle similitudini (ADS) in quanto è stata la rappresentazione grafica che ha permesso la migliore sintesi delle informazioni derivanti dai tre corpora. Per questo tipo di analisi l'utilizzo di un grafo (una struttura di dati consistenti in una serie di vertici e nodi) è quanto di più appropriato perché permette di descrivere i profili di specificità o di similarità dell'intero corpus. Essa si basa su una matrice testuale, variabile (X, in colonna) per casi (Y, in riga). Il punto di forza dell'ADS è che essa, a differenza di un'analisi delle corrispondenze e di un'analisi fattoriale, basate su una matrice delle distanze del Chi quadrato, nel primo caso, e di una di correlazione, nel secondo caso, considera i valori più forti tra due nodi, a livello locale e non globale, senza rappresentare graficamente i legami più deboli (Vergès e Bouriche 2001, 36)[20]. Tra le varie opzioni permesse dal programma si è optato per un algoritmo *force-directed layout* (Fruchterman e Reingold 1991)[21], in cui l'albero delle relazioni lessicali del corpus è stato calcolato sulle 100 occorrenze più diffuse, vale a dire le prime 100 «forme attive» (si rammenta che le parole sono state lemmatizzate).

Dai corpora on-line si sono ottenute le seguenti rappresentazioni grafiche le quali vengono commentate qui di seguito.

---

[19] Esse costituiscono «des traces possibles des contenus de nos activités. Ils ne sont pas les signifiants mais bien des traces possibles de ce contenu en acte. (…) Les mots pleins ont cependant la faculté de susciter des contenus en tant qu'ils stabilisent nos visions du monde, mais le contenu n'est pas dans le mot ; il est dans l'acte, dont le mot est une trace» (Reinert 2008, 983).

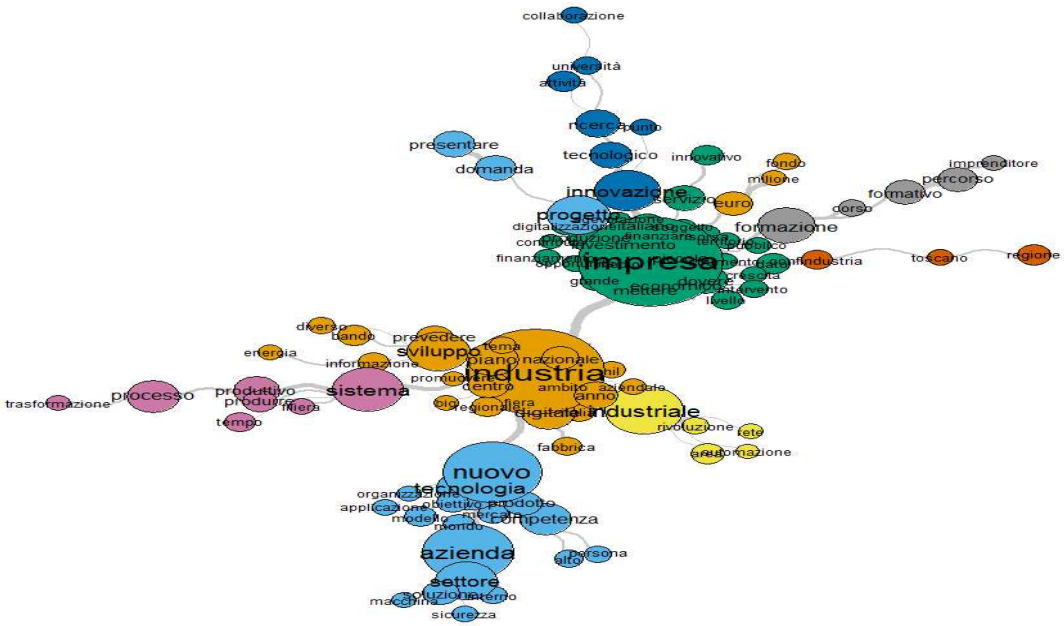
[20] «Autour d'un sommet (d'une variable) on prend en considération les valeurs les plus fortes (en particulier dans l'arbre maximum et dans les graphes seuil) sans se préoccuper de représenter graphiquement les valeurs les plus faibles» (Vergès e Bouriche 2001, 36).

[21] «The idea of a force directed layout algorithm is to consider a force between any two nodes. In this algorithm, the nodes are represented by steel rings and the edges are springs between them. The attractive force is analogous to the spring force and the repulsive force is analogous to the electrical force. The basic idea is to minimize the energy of the system by moving the nodes and changing the forces between them», goo.gl/S11LEF.

I corpus, Industria 4.0.

Nel caso italiano, dall'esame del grafico delle similitudini (figura 5), si può innanzitutto rilevare che la prima «forma attiva» è "impresa" seguita da "industria". Il termine "nuovo" ha una frequenza elevata, e fa riferimento sia all'ambito concettuale tecnologico (nuove tecnologie) ma anche in generale a svariati fenomeni (nuovo regolamento, nuova Sabatini, nuovo piano). Il concetto di nuovo e di novità è, del resto, strettamente associato a un fenomeno che si vuole e si racconta come rivoluzionario (la quarta rivoluzione) così come sono inediti (o così vengono raccontati) gli strumenti che la politica e le imprese si accingono a mettere in campo. Il quarto termine più frequente è quello di "azienda" ed esso può essere ricompreso in quello già visto di "impresa" rafforzando l'ipotesi che essa rappresenti il punto focale di Industria 4.0.

Figura 5 - Analisi delle similitudini delle prime 100 occorrenze del corpus Industria 4.0



Fonte: elaborazione Inapp

Altri nuclei tematici, messi in evidenza dal grafico, sono i seguenti. Il termine "progetto" è legato ad occorrenze quali "domanda" e "presentare" ed è chiaro che si sta parlando delle modalità con cui presentare la domanda per accedere ai finanziamenti. L'importanza dei bandi, nell'attuale piano nazionale Impresa 4.0, viene confermata anche da un piccolo nucleo tematico, legato al termine "impresa" e composto da "euro", "milioni", "fondo". La keyword "formazione" è connessa con le occorrenze "corso", "formativo", "percorso", "imprenditore" e segnala un aspetto, quello legato alle

attività di qualificazione e di riqualificazione che un fenomeno innovativo inevitabilmente porta con sé e che il Governo italiano si appresta a promuovere mediante l'avvio dei Competence center[22].
Un altro nucleo tematico significativo è legato a "innovazione" e comprende anche "tecnologia", "ricerca", "attività", "università", "collaborazione" e rimanda alla necessità di fare sistema tra diversi attori istituzionali. Il sistema produttivo ha bisogno di collaborare con Università ed Enti pubblici di ricerca per sviluppare ed implementare le attività innovative richieste dalla IoT.
In conclusione, sulla base di una valutazione complessiva di tutto il corpus, nel quale prevalgono le news ma i blog non sono pochi, si può inferire che nella situazione nazionale si sia passati, assai presto, a connotare il concetto di Industria 4.0 in senso più generalista e a ricomprendere nel fenomeno l'impresa in senso lato: da un campo più specifico, quello iniziale dell'ambito produttivo ed industriale, riferito sostanzialmente alla IoT, si è passati a uno più genericamente sociale e culturale. In questo senso, si parla in maniera enfatica, anche di quarta rivoluzione industriale.

II corpus, Impresa 4.0.

In questa narrazione on-line (figura 6) prevalgono due punti focali: "impresa" e "innovazione" e come già sottolineato in precedenza le ridotte dimensioni del corpus hanno determinato una sorta di sfilacciamento maggiore e, dunque, un addensamento minore rispetto ad alcuni nuclei tematici. Sembrano prevalere in misura maggiore, rispetto ad Industria 4.0, i verbi piuttosto che i sostantivi quali

creare	presentare	realizzare	produrre	prevedere	mettere	volere	crescere

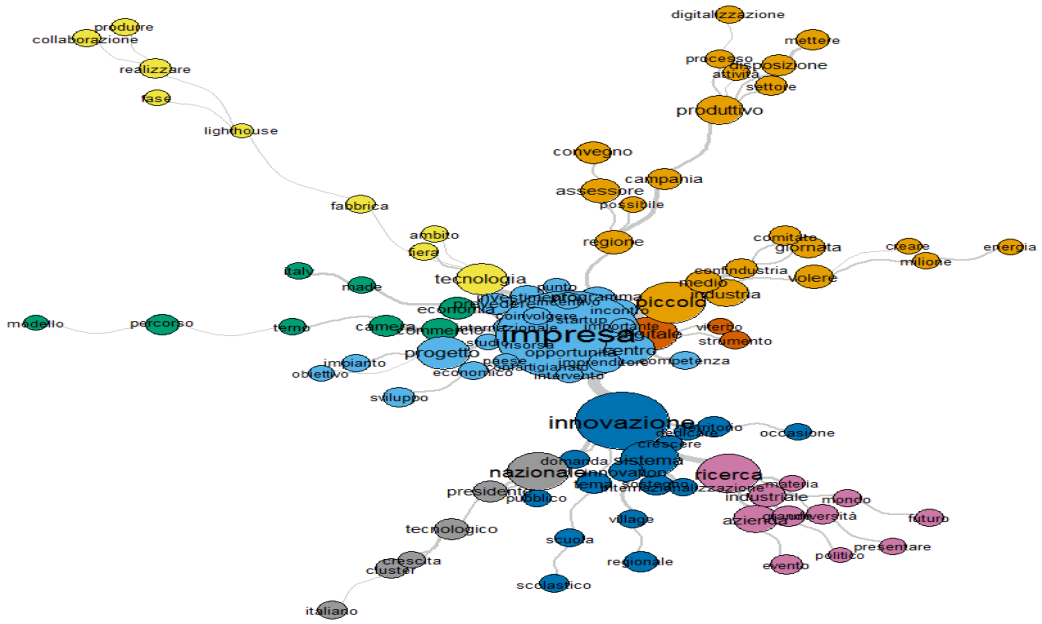
i quali rinviano, in maniera forte, a una spinta ad agire in direzione di un traguardo che forse si considera ancora in via di raggiungimento.
In questo corpus prevalgono le news e ciò sembra trovare un riscontro empirico nei termini

regione	campania	assessore	convegno	confindustria	comitato	giornata

i quali rimandano ad eventi, presentazione di progetti a carattere locale o associativo, convegni e manifestazioni.

[22] Luca Orlando, La corsa delle imprese per i competence center, 24 aprile 2018, goo.gl/cefH1D.

Figura 6 - Analisi delle similitudini delle prime 100 occorrenze del corpus Impresa 4.0



Fonte: elaborazione Inapp

III corpus, Industrie 4.0[23].

In questa rappresentazione grafica (figura 7) la parola chiave è senz'altro "industrie" al centro di un sistema che vede come satelliti "unternehmen" (impresa) e "können" (poter fare). È un corpus in cui prevalgono le news e, dall'esame complessivo del corpus, si può ipotizzare che esse siano funzionali a presentare ai lettori, e alla potenziale clientela determinati prodotti e tecnologie. Ad esempio, l'occorrenza "erp" (*enterprise resource planning*) è collegata a una delle principali società di software gestionale, la "sap" mentre il "cad" è legato ad un'altra società, la "schroer" specializzata nel disegno degli impianti e delle fabbriche.

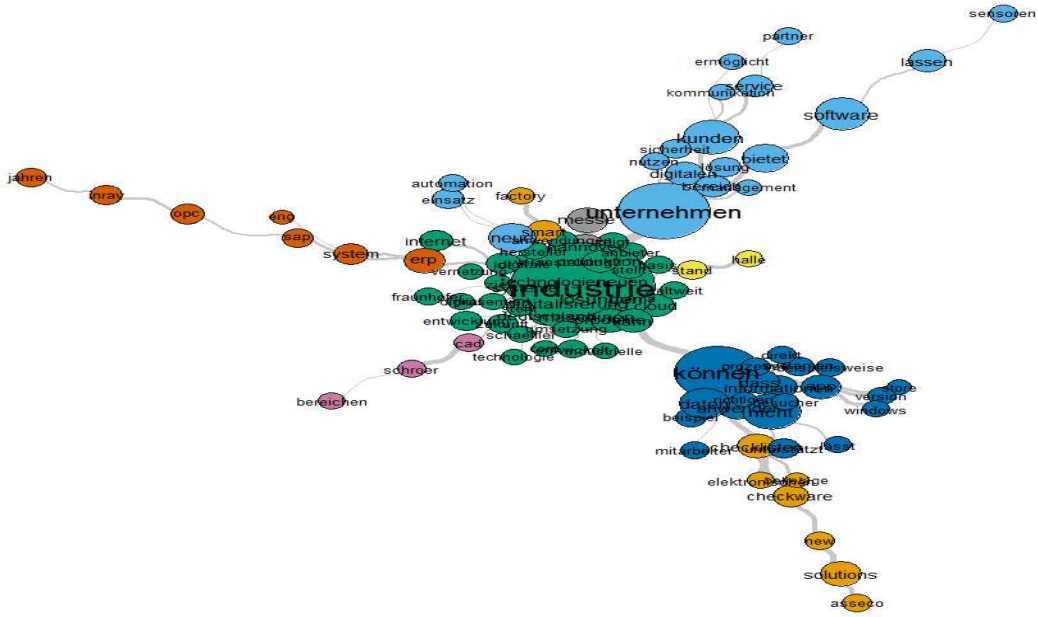
Vengono citate molte imprese quali "basf", "ericsson", "lufthansa", "salamander". Così come sono molto diffusi le parole legate all'ambito della produzione quali "software", "sensoren" (sensori), "app", "cloud", "maschinen" (macchine), "lte" (*long term evolution*) la quale indica la più recente evoluzione degli standard di telefonia mobile.

Il ricorso alla terminologia inglese è maggiore rispetto al caso italiano e oltre ai termini già citati si possono menzionare "checklisten", "software", "checkware", "solutions", "ray", "opc" (*open platform communications*), un protocollo per lo scambio di dati, da macchina a macchina, utilizzato nell'automazione industriale. L'aspetto di novità di Industrie 4.0 e delle tecnologie da implementare nella *smart factory* lo si evince, invece, dall'utilizzo dei termini "neue" e "new".

[23] Per la traduzione dei testi dal tedesco ci si è avvalsi della preziosa collaborazione del collega Inapp Ludovico Fioravanti.

La presenza di due termini quali "lösungen" (soluzioni) e "solutions" testimoniano che in rete vengono prospettati ed illustrati aspetti concreti di alcune tecnologie e software. Da rilevare anche la presenza della "fraunhofer", la più grande organizzazione di ricerca applicata in Europa.
In conclusione, la narrazione in rete sembra connotata dal frequente utilizzo di tecnicismi e da una maggiore apertura all'internazionalizzazione. La clientela è rappresentata dai termini "kunden" e "anwender" i quali contribuiscono a rafforzare l'ipotesi che si è in presenza di una narrazione essenziale, tecnica, senza fronzoli, tesa a illustrare ai clienti globali, e da qui l'utilizzo frequente della lingua inglese, le nuove tecnologie e i sistemi cyber-fisici già implementati nelle fabbriche tedesche.

Figura 7 - Analisi delle similitudini delle prime 100 occorrenze del corpus Industrie 4.0



Fonte: elaborazione Inapp

CONCLUSIONI

Lo sfruttamento dei big data, a fini di investigazione sociale, richiede nuovi strumenti per compiere attività di *web scraping* e nell'integrare, strutturare e visualizzare dati da molteplici fonti. Le opinioni diffuse on-line possono essere di interesse sociologico in quanto si avvicinano a una sorta di mentalità collettiva capace di "sintetizzare" un fenomeno sociale. Tale tipo di conoscenza sociale digitalmente diffusa evolve, infine, nel tempo poiché essa è tipicamente dinamica e può essere costantemente monitorata.

Il web, i cui limiti spazio-temporali sono soggetti a una sorta di sfumatura continua dei loro confini predeterminati, si presta, dunque, assai bene a collazionare, prima, e ad indagare, poi, in tempo reale e a prescindere dai diversi contesti nazionali, una determinata tematica. In questo senso, la rete può essere considerata un luogo privilegiato dove si produce, si scambia e si consuma conoscenza sociale prodotta da agenti esperti quali giornalisti e blogger.

Dall'esame dei risultati svolti in questo studio comparativo della pubblica opinione qualificata, tra la situazione nazionale e quella tedesca, nel periodo marzo-aprile 2018, la *web news coverage* di Industria 4.0 ha fornito i seguenti risultati.

Sui siti italiani si è scritto soprattutto di tematiche inerenti le imprese, in quanto la narrazione on-line si è spostata progressivamente dal concetto di Industria 4.0 a quello di Impresa 4.0, favorita in ciò dal Governo e dal ministro Calenda, vale a dire con una traslazione da un piano strettamente produttivo e tecnico a uno più genericamente culturale e sociale. In questo senso, in un prossimo futuro, i contenuti tematici potrebbero interessare ancor più l'ambito della formazione e delle competenze dei lavoratori (grazie all'avvio dei Competence center). In sintesi, si tratta di una fotografia iniziale scattata in un processo ancora in fieri.

A livello di disamina dei generi discorsivi in rete è stata evidenziata una differenza, di non poco conto, tra la conoscenza digitale diffusa dai blog e quella dai siti di news, dovuta alle diverse costrizioni, non solo tecnologiche, delle piattaforme informatiche e, di conseguenza, al genere testuale ivi allocato.

Per quel che concerne le news, queste hanno messo l'accento più su aspetti legati all'agenda-setting (McCombs e Shaw 1972) con notizie riportanti dichiarazioni di politici e addetti ai lavori, di eventi di presentazione quali mostre o convegni, di iniziative a livello nazionale o locale. La dimensione territoriale è, difatti, molto presente con riferimenti a specifiche iniziative, oppure ad associazioni di categoria ed enti istituzionali. Vi sono, infine, continui riferimenti alle possibilità di accesso ai fondi per la sostituzione dei macchinari oppure alle opportunità di mercato da cogliere, da parte di aziende e Pmi, grazie all'adozione delle nuove tecnologie.

Per quel che riguarda i blog, date le caratteristiche strutturali di tali piattaforme, sono state trattate in rete tematiche di maggior respiro con analisi più approfondite riguardanti gli effetti delle innovazioni, riflessioni sulle competenze da sviluppare, sulle dinamiche competitive.

Nel caso del corpus Impresa 4.0 si deve rilevare che l'esiguità dei messaggi impone una certa cautela nell'analisi dei risultati e qui si può solo sottolineare che il focus della narrazione in rete ha visto l'impresa al centro con alcuni satelliti discorsivi quali innovazione e ricerca.

Sui siti tedeschi la narrazione in rete ha evidenziato dei nuclei concettuali riferiti soprattutto all'ambito specifico di Industrie 4.0, un ambito molto più tecnico e specialistico, vale a dire fortemente interrelato con la produzione industriale, gli impianti, l'automazione e, assai meno, con quello culturale, politico e sociale con l'accento su prodotti, sensori, specifici marchi industriali. Altro dato da evidenziare è che, a differenza della situazione italiana, a livello di sub corpora, non sembra emergere una sostanziale differenza tra news e blog. In conclusione, dunque, la comunicazione relativa al caso tedesco è di carattere più tecnico e maggiormente focalizzata su tematiche IoT; anche gli articoli di news approfondiscono le tematiche proprie di Industrie 4.0.

La comparazione ha messo in evidenza che la situazione italiana appare troppo compressa su notizie legate all'attualità e ad eventi contingenti. Se gli opinionisti qualificati e i giornalisti hanno divulgato in rete tale conoscenza si potrebbe supporre che Industria/Impresa 4.0 abbia ancora un labile sostrato nella blogosfera nazionale. La sedimentazione comunicativa non è, allora, ancora avvenuta appieno e dagli articoli pubblicati (soprattutto news) non sempre si riescono a cogliere gli aspetti di profonda evoluzione di oggetti e strumenti della produzione contemporanea se non per alcuni aspetti superficiali e contingenti. Le implicazioni profonde di tali sviluppi tecnologici sembrano rimanere ancora sullo sfondo avvalorando le tesi di coloro che teorizzano che a uno sviluppo tecnologico si accompagna spesso un ritardo culturale nell'adozione e comprensione profonda dello stesso. Tale analisi può ricomprendere, però, in senso più ampio, non solo il livello comunicativo, oggetto di analisi in questo studio, quanto piuttosto anche il sistema produttivo nazionale il quale risulta connotato da carenza di competenze tecnico specialistiche e skills.

Stando così le cose si potrebbe ipotizzare un'azione governativa che abbia l'obiettivo di sensibilizzare il sistema dei media attraverso la disseminazione di materiale informativo/comunicativo al fine di sviluppare una maggiore alfabetizzazione tecnologica a livello generale. Si dovrebbe cercare, inoltre, di indirizzare in maniera più incisiva il sistema dell'istruzione e della formazione professionale, anche in tema di alternanza scuola/lavoro, per rendere maggiormente consapevoli i giovani delle opportunità offerte dai diversi percorsi professionali nell'ambito delle nuove tecnologie e di Industria 4.0. Un altro aspetto che occorrerebbe sviluppare è quello della formazione on-line, in primis quella disponibile sulle piattaforme MOOC (*massive on-line open courses*) la quale viene svolta in Germania, ad esempio, con corsi specifici promossi dalla Akademie der Technikwissenschaften (Acatech)[24], oltre che nel Regno Unito, Francia e Stati Uniti.

Per quel che riguarda l'indagine, infine, sul piano strettamente metodologico, potrebbero essere svolti ulteriori approfondimenti ad hoc, in un prossimo futuro, così come la messa a regime di un monitoraggio continuo relativo alla *web news coverage* di Industria 4.0: una materia che dice relazione non solo a un ambito peculiare quale può essere considerata la tematica produttiva e industriale ma che riguarda la società intera con aspetti culturali e politici di non poco conto.

[24] Cfr. ad esempio, Smart Service Welt – Data and Platform-Based Business Models, ospitato sulla piattaforma openSAP, goo.gl/ixHxkq, oppure Imagine IoT, sempre sulla stessa piattaforma, goo.gl/DYvu4z.

BIBLIOGRAFIA

AKADEMIE DER TECHNIKWISSENSCHAFTEN (ACATECH) (2013), Recommendations for implementing the strategic initiative Industrie 4.0. Securing the future of German manufacturing industry, *Final report of the Industrie 4.0 Working Group*, April, pp. 82

BALLARINO G., CHECCHI D. (2013), La Germania può essere un termine di paragone per l'Italia? Istruzione e formazione in un'economia di mercato co-ordinata, *Rivista di Politica Economica*, vol. 1, pp. 39-74

BOLASCO S. (1995), *Criteri di lemmatizzazione per l'individuazione di coordinate semantiche*, in Cipriani R., Bolasco S., *Ricerca qualitativa e computer. Teorie, metodi e applicazioni*, Milano, FrancoAngeli, pp. 87-111

BOLASCO S. (2005), Statistica testuale e text mining: alcuni paradigmi applicativi, *Quaderni di Statistica*, vol. 7, pp. 17-53

BRYNJOLFSSON E., MCAFEE A. (2011), *Race Against the Machine. How the Digital Revolution is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy*, Lexington (MA), Digital Frontier Press

CHEN Y. (2017), Integrated and Intelligent Manufacturing. Perspectives and Enablers, *Engineering*, vol. 3, pp. 588-595

DEGENNE A., VERGES P. (1973), Introduction à l'analyse de similitude, *Revue française de sociologie,* vol. 14, n. 4, pp. 471-511

FLAMENT C. (1962), L'analyse de similitude, *Cahiers du centre de recherche opérationnelle*, vol. 4, pp. 63-97

FRUCHTERMAN T. M. J., REINGOLD E. M. (1991), Graph Drawing by Force-Directed Placement*, Software. Practice and Experience*, vol. 21, n. 11, November, p. 1.129-1.164

HERRING S. C., KOUPER I., PAOLILLO J. C., SCHEIDT L. A, TYWORTH M., WELSCH P., WRIGHT E., YU N. (2005), Conversations in the Blogosphere. An Analysis From the Bottom Up, *Proceedings of the 38th Hawaii International Conference on System Sciences*, vol. 9, January 3-6, working paper, pp. 11

KITCHIN R. (2014), *The Data Revolution. Big Data, Open Data, Data Infrastructures & Their Consequences*, London (UK), Sage

KOLBERG D., ZÜHLKE D. (2015), Lean Automation enabled by Industry 4.0 Technologies, *International Federation of Automatic Control (IFAC)-PapersOnLine*, vol. 48, n. 3, pp. 1870-1875

LEBART L., SALEM A. (1994), *Statistique textuelle*, avec une *Préface* de Christian Baudelot, Paris (FR), Dunod

LEE J., BAGHERI B., KAO H.-A. (2015), A Cyber-Physical Systems architecture for Industry 4.0-based manufacturing systems, *Manufacturing Letters*, vol. 3, pp. 18-23

LEE J., KAO H.-A., YANG S. (2014), Service innovation and smart analytics for Industry 4.0 and big data environment, *Procedia CIRP*, vol. 16, pp. 3-8

MANNING C. D., SCHÜTZE H. (1999), *Foundations of Statistical Natural Language Processing*, Cambridge (MA), MIT Press

MCCOMBS M. E., SHAW D. L. (1972), The Agenda-Setting Function of Mass Media, *The Public Opinion Quarterly*, vol. 36, n. 2, Summer, pp. 176-187

MITCHELL R. (2015), *Web Scraping with Python. Collecting Data from the Modern Web*, Sebastopol (CA), O'Reilly

NORRIS P. (2009), *The Globalization of Comparative Public Opinion Research*, in Robinson N. and Landman T. (eds.), *Handbook of Comparative Politics*, London, Sage, pp. 522-539

PALIOTTA A. P. (2018a), Nuevas profesiones y técnicas de web data mining en Argentina. El caso del Data scientist, *Revista del Centro de Estudios de Sociología del Trabajo*, n. 10, pp. 97-113

PALIOTTA A. P. (2018b), La ricezione on-line dell'Amoris Laetitia di Papa Francesco. Generi discorsivi in rete e tecniche di text mining, *Religione e Società* (in corso di pubblicazione)

PALIOTTA A. P. (2016), Ricerca di lavoro e metodologie di web data mining. Il profilo del Data scientist nelle inserzioni on-line, *Osservatorio Isfol*, n.s. a. VI, n. 3, pp. 131-149

OECD (2017), OECD Skills Strategy Diagnostic Report. Italy, *Report*, Paris (FR)

REINERT M. (2008), Mondes lexicaux stabilisés et analyse statistique de discours, *9es Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, pp. 981-993

ROGERS R. (2015), *Digital Methods for Web Research* in Scott R., Kosslyn S. (eds.), *Emerging Trends in the Social and Behavioral Sciences*, New York (NY), John Wiley & Sons, pp. 1-22

SCHUH G., POTENTE T., WESCH-POTENTE C., WEBER A. R., PROTE J.-P. (2014), Collaboration Mechanisms to increase Productivity in the Context of Industrie 4.0, *Procedia CIRP*, vol. 19, pp. 51-56

VERGES P., BOURICHE B. (2001), L'analyse des données par les graphes de similitude*, Sciences Humaines,* pp. 90

ZIPF G. K. (1949), *Human Behaviour and the Principle of Least-Effort,* Cambridge (MA), Addison-Wesley

ZEZULKA F., MARCON P., VESELY I., SAJDL O. (2016), Industry 4.0. An Introduction in the phenomenon, *International Federation of Automatic Control (IFAC)-PapersOnLine*, vol. 49, n. 25, pp. 8-12

ZHONG R. Y., XU X., KLOTZ E., NEWMAN S. T. (2017), Intelligent Manufacturing in the Context of Industry 4.0. A Review, *Engineering*, vol. 3, pp. 616-630

Sitografia: nel corso del testo sono riportati molti link a pagine web i quali non vengono qui riportati per non appesantire troppo la bibliografia: tutti questi, quando non espressamente indicato, sono stati visitati ad aprile 2018.

ISSN 2533-2996

ISBN 978-88-543-0169-6