

DIPARTIMENTO DI
ECONOMIA E FINANZA

**METODI E ANALISI
STATISTICHE
2022**



UNIVERSITÀ
DEGLI STUDI DI BARI
ALDO MORO

**DIPARTIMENTO DI
ECONOMIA E FINANZA**

**METODI E ANALISI
STATISTICHE
2022**



Tutti i diritti di traduzione, riproduzione e adattamento, totale o parziale, con qualsiasi mezzo (comprese le copie fotostatiche e i microfilm) sono riservati

Toma E., (a cura di). *Metodi e Analisi Statistiche 2022*, Dipartimento di Economia e Finanza, Università degli studi di Bari Aldo Moro.

© Copyright 2022 by Università degli Studi di Bari Aldo Moro
www.uniba.it

Prima edizione: dicembre 2022

ISBN 978-88-6629-078-0

Gli articoli qui presentati sono stati tutti oggetto di valutazione interna.

Editing finale: F.D. d'Ovidio, E. Toma

Sommario

Ernesto Toma <i>Presentazione</i>	pag. 5
Giovanni Armillotta, Crescenza Calculli, Alessio Pollice, Letizia Sion, Porzia Maiorano, Francesca Capezzuto, Angela Carluccio, Gianfranco D’Onghia <i>Studio della distribuzione spazio-temporale dei pesci cartilaginei nel Mar Ionio nord-occidentale</i> «	7
Paola Perchinunno, Antonella Massari, Samuela L’Abbate, Lucia Mongelli <i>Sustainable Development Goals per l’analisi statistica della povertà</i> «	27
Samuela L’Abbate, Paola Perchinunno <i>Un nuovo approccio al DBSCAN per raggruppare aree dense</i> «	41
Angela Maria D’Uggento, Massimo Armenise, Monica Carbonara <i>A synthetic index for the analysis of the eno-gastronomic tourist attractiveness of Italian territories</i> «	55
Sabrina Diomede, Giovanni Taglialatela <i>Disuguaglianze bilaterali di tipo Pearson per alcune classi di variabili aleatorie</i> «	65
Francesco Domenico d’Ovidio, Angela Maria D’Uggento, Monica Cazzolle, Ernesto Toma <i>Dall’università al mercato del lavoro: un focus sulle aspettative dei laureati</i> «	83
Angela Maria D’Uggento, Riccardo Roselli, Barbara Cafarelli <i>Modelli probabilistici per la gestione dell’incertezza in ambito finanziario</i> .. «	95
Angela Maria D’Uggento, Luana Cecca, Barbara Cafarelli <i>Tecniche di text mining per l’analisi delle dichiarazioni non finanziarie delle aziende</i>	« 115

Dominga Capozzo, Nunziata Ribecco, Clelia Punzo, Rita Laforgia, Giustina Angellotti, Angela Pezzolla <i>Analisi statistica sulla patologia nodulare tiroidea</i> «	131
Marta Schito, Nunziata Ribecco, Giustina Angellotti, Clelia Punzo <i>Analisi statistica su macro metastasi linfonodali dopo biopsia del linfonodo sentinella nel trattamento chirurgico del carcinoma mammario</i> .. «	143
Carmine Lamanna, Barbara Cafarelli, Angela Maria D’Uggento <i>Tecniche statistiche per la previsione del rischio di insolvenza</i> «	155
Laura Antonucci, Corrado Crocetta, Angelica Di Terlizzi <i>Analisi del gender gap nelle Università italiane fra il 2021 e il 2020</i> «	179
Laura Antonucci, Antonia Cofano, Corrado Crocetta, Massimo Russo <i>La customer satisfaction della Sanitaservice ASL Foggia s.r.l.</i> «	187
Monica Carbonara, Agata Maria Madia Carucci, Giovanni Vannella <i>Una valutazione del benessere “a misura di comune” attraverso una proposta di indicatori sintetici per il reddito e per il dominio “istruzione”</i> ... «	203
Laura Antonucci, Giulio Biscardi, Najada Firza, Leonardo Grilli <i>Un modello ad effetti misti con dati longitudinali per il progress test dei corsi di laurea italiani in odontoiatria</i> «	219
Angela Maria D’Uggento, Barbara Cafarelli, Massimo Iaquinta <i>Il fenomeno dell’abbandono universitario in Italia</i> «	235

Tecniche di text mining per l'analisi delle dichiarazioni non finanziarie delle aziende

Angela Maria D'Uggento^{1*}, Luana Cecca¹, Barbara Cafarelli²

¹Dipartimento Economia e Finanza, Università degli Studi di Bari Aldo Moro

²Dipartimento di Economia, Management e Territorio, Università di Foggia

Riassunto: Il Decreto Legislativo 30 dicembre 2016, n.254 prevede, per determinate tipologie di aziende, la redazione della Dichiarazione Non Finanziaria con cadenza annuale al fine rendicontare sui principali aspetti sociali e ambientali: sostenibilità aziendale; gestione del personale, in particolare le politiche sulla parità di genere; impegno nella lotta alla corruzione e rispetto dei diritti umani. Lo scopo del presente lavoro è quello di analizzare l'evoluzione del contenuto dei documenti dal 2017, anno di prima attuazione del decreto, al 2020, per individuare possibili miglioramenti nel contenuto dei documenti al fine di soddisfare la crescente domanda di disclosure in materia non finanziaria da parte degli stakeholder delle aziende. In particolare, è stato analizzato il settore tessile per esplorare l'evoluzione delle politiche di sostenibilità delle aziende italiane. I risultati hanno confermato un progressivo miglioramento dei contenuti dei documenti sia in termini informativi che di conformità al dettato normativo.

Keywords: sostenibilità; text mining; dichiarazione non finanziaria DNF; informazione non finanziaria.

1. Introduzione

Il Decreto Legislativo 30 dicembre 2016 n.254 prevede, per determinate tipologie di aziende, la redazione della Dichiarazione Non Finanziaria con cadenza annuale (Marcello et al, 2018; Deloitte, 2021, 2018). L'introduzione di questo documento

* Autore corrispondente: angelamaria.duggento@uniba.it

Il lavoro qui descritto è frutto di un progetto comune, ma D'Uggento ha provveduto alla redazione dei paragrafi 1 e 2, mentre Cecca. ha redatto il paragrafo 3 e Cafarelli i paragrafi 2.1 e 4.

ha fatto sì che le imprese cominciassero, per obbligo di legge o volontariamente, a confrontarsi con un nuovo modo di comunicare le proprie performance e ha, quindi, da un lato, evidenziato le best practice che alcune aziende avevano da tempo adottato, e, dall'altro, indotto le restanti a rivedere i propri sistemi di rendicontazione (Helfaya, A. and Moussa, T., 2017).

L'importanza di questo documento e della sua chiarezza, trasparenza e completezza, deriva dal fatto che le aziende che rendicontano le proprie strategie ambientali, sociali e di governance (ESG) sono molto apprezzate dalla comunità finanziaria, e, in primis, dagli investitori, i quali si avvalgono delle informazioni non finanziarie, tra le altre, per orientare le proprie scelte di investimento (Cupertino e Vitale, 2021; Salvatore et al, 2022).

Lo scopo del presente lavoro è quello di analizzare l'evoluzione del contenuto, sia in termini quantitativi sia qualitativi, delle Dichiarazioni Non Finanziarie pubblicate dal 2017, anno di prima attuazione del decreto, al 2020 per trarre possibili indicazioni al fine di soddisfare la crescente domanda di disclosure in materia non finanziaria da parte degli stakeholder delle aziende (Santamaria et al, 2021; Spallini et al, 2021; Phan et al, 2020). In particolare, è stato analizzato il settore tessile selezionato tra quelli presenti nella banca dati dell'Osservatorio DNF dell'Università di Siena (Osservatorio, 2022) perché ritenuto uno dei più adatti ai fini della completezza dell'analisi testuale in relazione agli obiettivi del presente lavoro (Li e Leonas, 2022), che è strutturato come segue: dopo una breve introduzione, il paragrafo 2 illustra i dati e le tecniche di Text mining utilizzate; il paragrafo 3 presenta i principali risultati e il paragrafo 4 delle sintetiche conclusioni.

2. Materiali e metodi

La quantità di documenti che possiamo trovare nel web cresce esponenzialmente giorno dopo giorno con la conseguenza che i dati e le informazioni disponibili diventano troppi per essere processati manualmente.

Le tecniche di Text Mining sono state sviluppate con lo scopo di individuare pattern nei dati ed estrarre conoscenza partendo da testi non strutturati. Si tratta di strumenti che sintetizzano l'eccessiva quantità di informazioni usando, tra le altre, tecniche di data mining, machine learning e Natural Language Processing. Sono evidenti, quindi, le enormi potenzialità, ma anche la varietà degli ambiti di applicazioni, dall'individuazione di spam all'analisi del sentiment fino alla misurazione delle preferenze del cliente.

In questo lavoro, l'analisi è condotta ipotizzando che la frequenza dell'utilizzo di termini legati alla sostenibilità in una DNF rifletta l'importanza e l'enfasi che l'impresa che ha redatto il report pone riguardo tali argomenti. Maggiore la frequenza o l'occorrenza di un certo termine, maggiore l'enfasi che l'azienda attribuisce ad esso e maggiore l'importanza attribuita (Te Liew et al., 2014). Attraverso il Text Mining e l'analisi Term-Frequency si è cercato, quindi, di identificare le più importanti tendenze relativamente alla sostenibilità.

Il Text Mining si può realizzare essenzialmente attraverso due fasi: Text Refining, o pulizia del testo, e Knowledge Distillation, o estrazione di conoscenza. La prima fase si riferisce al pre-processing dei testi di input per preparare i dati per la fase successiva, trasformando i testi originali non strutturati in forma standardizzata per le analisi. La seconda fase, invece, serve a individuare eventuali pattern e identificare le informazioni, per far emergere conoscenza che non sarebbe percepibile dai singoli testi, ma che è possibile evincere dalla lettura contemporanea di una serie di documenti.

Per l'estrazione delle DNF si è scelto di utilizzare il linguaggio di programmazione Python, il quale, tramite la modifica di poche righe di codice, renderebbe possibile anche l'estrazione futura di ulteriori documenti pubblicati negli anni a venire, ampliando così il corpus in analisi.

Il sito dal quale sono stati acquisiti i file è quello dell'Osservatorio DNF (<https://www.osservatoriodnf.it/>) che raccoglie e consente di accedere direttamente alle rendicontazioni non finanziarie delle aziende italiane suddivise per settore. Nello specifico, le aziende del settore tessile oggetto di analisi sono elencate nella Tabella 1 secondo la classificazione Consob. La scelta di analizzare le aziende del settore tessile è dettata dal fatto che, negli ultimi anni, questo è stato oggetto di pesanti critiche in merito agli impatti ambientali e sulla salute umana derivanti dai residui di sostanze nocive sui prodotti ed ai processi produttivi ad esso associati, nonché perché considerato, soprattutto relativamente al fast-fashion, uno dei settori più inquinanti (Li e Leonas, 2022).

Le DNF scaricate sono in formato PDF e contengono immagini di testo che non possono essere elaborate dal software di elaborazione di testi. Per questo, e per garantire l'automaticità del processo (senza dover riscrivere il testo contenuto nei documenti), sono stati sottoposti a trattamento con OCR (Optical Character Recognition) attraverso il pacchetto Python "pdfreader".

Tabella 1. *Elenco delle aziende analizzate settore tessile*

Denominazione	Classificazione
AEFFE SPA	Società con azioni quotate in Italia
BENETTON GROUP SPA	Società con azioni quotate in Italia
BRUNELLO CUCINELLI SPA	Società con azioni quotate in Italia
CSP INTERNATIONAL FASHION GROUP SPA	Società con azioni quotate in Italia
RATTI SPA	Società con azioni quotate in Italia
MONCLER SPA	Società con azioni quotate in Italia
SALVATORE FERRAGAMO SPA	Società con azioni quotate in Italia
VINCENZO ZUCCHI SPA	Società con azioni quotate in Italia

L'OCR ha lo scopo di leggere stringhe di testo da convertire in codice telegrafico. Una volta ottenuti i testi, questi sono stati salvati in formato TXT, pronti per la successiva fase.

Successivamente all'estrazione dei testi, è necessario effettuarne la pulizia attraverso la rimozione di caratteri numerici, di punteggiatura, di spazi multipli e la trasformazione di tutti i caratteri in minuscolo. Questa fase non ha una procedura standardizzabile poichè varia a seconda dell'oggetto di studio e delle finalità dell'analisi. Inoltre, non è una fase considerata obbligatoria ma indubbiamente utile per migliorare la qualità del testo ed eliminare caratteri non significativi.

In seguito, i testi processati sono stati depurati dalle cosiddette Stopword, parole che, poiché molto frequenti in una lingua, sono poco significative ai fini di una ricerca testuale, nel senso che non aggiungono alcun valore al processo di identificazione del concetto. Queste parole sono solitamente articoli, preposizioni, congiunzioni ed alcuni avverbi, ma anche verbi ausiliari, verbi servili, ecc.

I testi così manipolati costituiscono il corpus testuale su cui effettuare le analisi. Ad essi sono stati, comunque, associati i metadati più importanti quali nome azienda, settore e anno di riferimento della DNF.

Una volta terminate le attività previste dalla fase di pre-processing, l'analisi è stata sviluppata avvalendosi del software RStudio.

2.1 Term-Document Matrix, algoritmo TF-IDF e N-Grams

Alcuni degli strumenti più utilizzati per l'analisi dei dati di forma testuale sono la Term-Document Matrix, l'algoritmo TF-IDF e gli N-Grams.

La Term-Document Matrix è una matrice che rappresenta la frequenza con cui un termine ricorre all'interno di un testo. L'importanza di tale strumento è motivata dal fatto che il documento, a livello concettuale, è un insieme di n termini; a livello tecnico, invece, è un'unica stringa. In questa analisi, la Term-Document Matrix verrà costruita in modo che vi siano n righe quanti sono i termini apparsi almeno

una volta nell'intero corpus e un numero di colonne pari al numero delle DNF considerate. Invertendo righe e colonne verrebbe generata la Document-Term Matrix che funziona in modo analogo. Ciascuna cella individua il numero di volte in cui quel termine appare nel relativo documento.

La probabilità che una TDM (o una DTM) sia costituita da un numero elevato di celle con valori pari a zero è molto alta. Questo fenomeno, tipico dei dati non strutturati (per i quali è molto frequente incontrare termini "rari"), è noto col nome di sparsità. Più standardizzati sono i documenti, più il valore di sparsità risulta basso. La sparsità di una matrice, inoltre, rende necessaria la rimozione di termini molto rari prima di procedere con le analisi, in modo da limitarne la complessità computazionale.

L'algoritmo TF-IDF è composto da due parti: TF, o Term Frequency e IDF, o Inverse Document Frequency. Il suo valore aumenta proporzionalmente al numero di volte che il termine è contenuto nel documento, tuttavia, cresce in maniera inversamente proporzionale con la frequenza del termine nel totale dei documenti (Te Liew et al., 2014). Lo scopo è quello di dare più importanza ai termini che compaiono spesso nella singola DNF, ma che, in generale, appaiono in pochi documenti. In poche parole, con questo indicatore si va a pesare ogni termine di un determinato documento in misura tanto maggiore quanto è meno frequente negli altri documenti esaminati. Se un termine è presente in tutti i documenti, allora il suo peso sarà basso, al contrario, se il termine è contenuto solo in alcuni documenti, avrà un peso maggiore.

Come accennato, la funzione può essere scomposta in due fattori. Il primo fattore rappresenta il numero delle volte che il termine è presente nel documento e, di solito, viene rapportato alla lunghezza del documento al fine di evitare di sovradimensionare i documenti più lunghi. Il secondo fattore è il logaritmo del rapporto tra il numero di documenti analizzati in totale ed il numero di documenti che contengono il termine.

Nello specifico, la formula per calcolare l'indice è la seguente:

$$TF\text{-}IDF_{td} = TF_{td} * IDF_t$$

dove

$$IDF_t = \log (|D| / DF_t).$$

Risulta chiaro come, qualora il termine apparisse in tutti i documenti, l'IDF sarebbe 0.

Per completare la rassegna degli strumenti più usati nel Text mining, gli N-Grams sono sequenze di n parole, solitamente due (Bi-Grams) o tre (Tri-Grams),

che appaiono vicine in un determinato testo. In particolare, i termini devono essere contigui perché lo scopo è proprio quello di evidenziare l'uso congiunto dei termini.

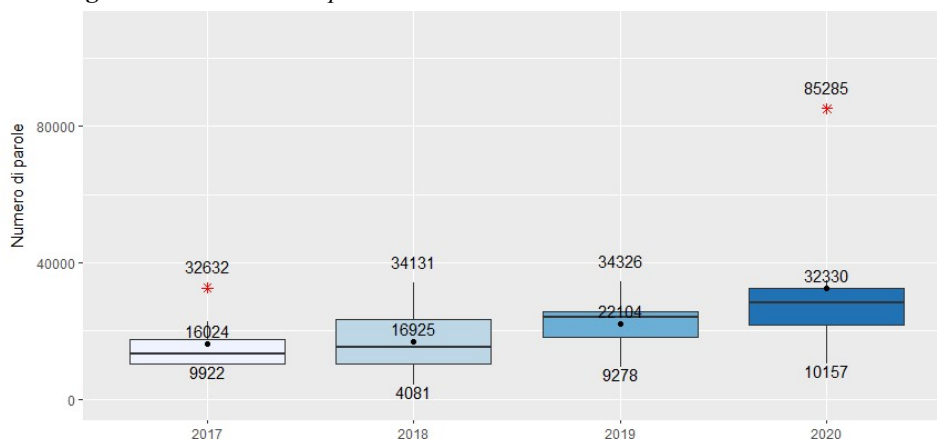
Per studiare la relazione tra parole adiacenti occorre in primis “tokenizzare” per sequenze consecutive di n parole, chiamate appunto N-Grams. Questo primo step di analisi può essere condotto con la libreria “igraph” e la libreria “ggraph” di RStudio.

La rappresentazione grafica, chiamata Network Graph, di rapida e comprensibile visualizzazione, ha lo scopo di mostrare le relazioni tra punti in modo più immediato, permettendo di visualizzare come le parole siano interconnesse tra loro. I termini vengono visualizzati come nodi e la relazione tra gli stessi viene visualizzata con linee o frecce, tanto più marcate quante più volte co-occorrono contigualmente. Per effettuare tali analisi, nel presente contributo si è scelto di utilizzare i Bi-Grams.

3. Risultati

Ipotizzando che le DNF abbiano subito un'evoluzione in termini di contenuti dal 2017, i box plot in Figura 1 consentono di verificare se, dall'anno di prima implementazione del decreto agli anni successivi, si sia verificato un miglioramento nella struttura e nei contenuti, ad esempio un aumento medio di parole per documento.

La Figura 1 evidenzia come le aziende del settore tessile nel primo anno abbiano prodotto dei documenti dai contenuti “essenziali”. Difatti, il box plot relativo all'anno 2017 mostra una media di 16 024 parole, che corrisponde ad un numero complessivo di termini del corpus testuale non depurato dalle stopword nella fase di precedente cleaning. Inoltre, relativamente allo stesso anno, si notano un minimo di 9922 parole per l'azienda Zucchi e un massimo di 32 632 per l'azienda Aeffe. Successivamente a questo primo anno di recepimento della normativa, le aziende hanno avuto la possibilità di approfondire la conoscenza del tema e gli obiettivi del decreto e quindi hanno potuto discostarsi dal contenuto minimale da quest'ultimo imposto, integrando nelle DNF la narrazione delle attività svolte, gli impatti e i principali risultati conseguiti. I box plot in Figura 1 mostrano, nel settore tessile, una tendenza costante all'incremento del volume totale dei termini contenuti nei documenti fino a registrare un raddoppio del dato nel 2020 rispetto al 2017.

Figura 1. Box Plot delle parole utilizzate nelle Dichiarazioni Non Finanziarie

La Tabella 2 mostra i principali indici di statistica descrittiva calcolati per sintetizzare e confrontare i testi delle DNF delle 8 aziende del settore tessile (Harymawan et al., 2020). Il numero massimo di parole significative è 85285, presenti nel documento redatto dall'azienda Salvatore Ferragamo nell'anno 2020. Notevole è la differenza con Zucchi Group che, nello stesso anno, utilizza 10157 parole nella propria DNF. La deviazione standard massima, per azienda, è rilevabile in Salvatore Ferragamo con 34733 parole di scarto.

Tabella 2. Statistiche descrittive sulle parole utilizzate all'interno delle Dichiarazioni Non Finanziarie.

Anno	2017	2018	2019	2020	STATISTICHE				
Azienda					Minimo	Massimo	Somma	Media	Dev. Standard
Aeffe	32632	34131	34326	34662	32632	34662	135751	33938	898
Benetton Group	12989	14693	17645	22981	12989	22981	68308	17077	4381
Brunello Cucinelli	10255	10949	24273	31547	10255	31547	77024	19256	10429
Csp	10152	4081	10650	17717	4081	17717	42600	10650	5578
Gruppo Ratti	15750	22740	23961	26496	15750	26496	88947	22237	4599
Moncler	22745	25403	26876	29793	22745	29793	104817	26204	2940
Salvatore Ferragamo	13750	15651	18366	85285	13750	85285	133052	33263	34733
Zucchi Group	9922	7754	9278	10157	7754	10157	37111	9278	1082
STATISTICHE									
Minimo	9922	4081	9278	10157					
Massimo	32632	34131	34326	85285					
Somma	128195	135402	165375	258638					
Media	16024	16925	20672	32330					
Dev. Standard	7932	9951	8396	22814					

3.1 Costruzione della Term-Document Matrix per le analisi qualitative

Per capire se e come cambia il contenuto delle DNF aziendali nel settore oggetto di analisi, è stato effettuato uno studio sulla frequenza dei termini significativi, considerando le varie aziende come componenti di un unico corpus “settore tessile”.

Per passare, quindi, all’analisi del contenuto, si parte dalla costruzione della Term-Document Matrix. La TDM generata ha una sparsità dell’84%. Tale risultato non sorprende poiché stiamo trattando testi redatti in base ad uno schema prestabilito dal decreto. In ogni caso, è indicativa di un’elevata attenzione al rispetto degli standard. A livello tecnico, quindi, nel caso di studio non è necessario ridurre la sparsità della matrice eliminando i termini con le frequenze più basse poiché la matrice risulta già adeguatamente “densa”.

Per mezzo di questa matrice possiamo capire quali sono i cinquanta termini che hanno la frequenza più alta. La frequenza è calcolata sommando il numero di volte in cui compare il termine preso in analisi nel settore. Tale analisi di frequenza potrà dare utili elementi per investigare con quale approccio le aziende affrontano le questioni legate alla sostenibilità, e se, in particolare, prevalga l’uso delle parole negative, che indicano preoccupazione e criticità, al contrario di parole positive che rappresentano l’attenzione e gli sforzi compiuti delle aziende. I risultati sono illustrati nella Tabella 3 e, data l’immediatezza dei confronti possibili, gli stakeholders, in primis investitori, clienti e fornitori, sono agevolati nel comprendere le strategie dell’azienda e il modo in cui questa affronta le tematiche economiche, sociali e ambientali.

Tabella 3. Rank delle parole (a carattere positivo/negativo) più utilizzate all’interno delle Dichiarazioni Non Finanziarie 2017-2020.

Rank del termine	Contenuto	
	Positivo	Negativo
Rank 1	valore	rischi
Rank 2	sostenibilità	emissioni
Rank 3	rispetto	rifiuti
Rank 4	sicurezza	debiti
Rank 5	sviluppo	perdite
Rank 6	formazione	corruzione
Rank 7	responsabilità	infortuni
Rank 8	attenzione	consumo
Rank 9	collaborazione	pericolosi
Rank 10	tutela	serra

Le parole: "sostenibilità", "valore", "sicurezza", "rispetto" e "sviluppo" sono le parole più frequenti legate alla strategia impiegata dall'azienda per trattare i temi economici, sociali e ambientali della sostenibilità (Garcia-Sanchez et al, 2020). Al contrario, i termini "emissione", "rischi", "rifiuti", "corruzione", "infortuni" e "pericolo" sono parole frequenti legate ai principali problemi affrontati nel bilancio di sostenibilità. Tra le parole positive più frequenti troviamo "sostenibilità", "valori" e "sicurezza". Questo risultato evidenzia che le aziende analizzate sono attente e impegnate nella strategia di implementazione di uno sviluppo sostenibile. I termini "rispetto", "attenzione" e "protezione" descrivono i principi dell'azienda riguardo le attività di responsabilità sociale d'impresa (Balluchi e Furlotti, 2019) svolte a favore del pianeta e dell'ambiente e l'importanza del principio di giustizia a cui l'azienda dichiara di conformarsi. Al contrario, tra le parole negative più frequenti ci sono "emissioni", "riscaldamento" e "rifiuti", indicando che uno dei principali problemi affrontati dalle aziende del settore tessile riguarda le emissioni collegate al riscaldamento globale, ma anche le questioni connesse ai rifiuti industriali. Successivamente, le parole "corruzione" e "frode" evidenziano come all'interno delle DNF siano trattate anche le tematiche sociali, in questo caso con accezione particolarmente negativa, unitamente a quelle positive, quali le parole "partecipazione", "collaborazione" e "formazione". I termini "pericolo", "criticità" ed "emergenza" si riferiscono all'impatto significativo e meno che mai trascurabile delle attività industriali sull'ambiente.

Una ulteriore modalità di presentazione dei risultati dell'analisi delle frequenze è di tipo grafico e prende il nome di Wordcloud, una nuvola di parole che fornisce una rappresentazione immediata delle parole più frequentemente utilizzate nell'arco dei quattro anni. L'immediatezza della tecnica di visualizzazione deriva dal fatto che la dimensione delle parole nella nuvola è proporzionale alla loro frequenza nel corpus oggetto d'analisi. Si è scelto di mostrare i 50 termini più utilizzati. La Figura 2 mostra come le parole più frequenti siano "gestione" e "attività", il che non sorprende essendo i documenti analizzati redatti proprio per rendicontare l'operato delle aziende, ma anche "stakeholders", cioè i destinatari dei report. Emergono anche i termini tipici delle tre dimensioni di sostenibilità: "sviluppo", "GRI", "emissioni", "rispetto", "sicurezza", "salute", "persone".

Figura 2. Wordcloud dei 50 termini più frequenti nei quattro anni settore textile.

3.2 L'Algoritmo TF-IDF

L'algoritmo TF-IDF consente di effettuare un'analisi più puntuale rispetto alla visualizzazione ottenuta con la wordcloud. In particolare, nel presente lavoro è stato impiegato per studiare l'evoluzione dei termini nel corso degli anni attraverso i quattro corpora testuali, considerando ogni anno come un documento a sé.

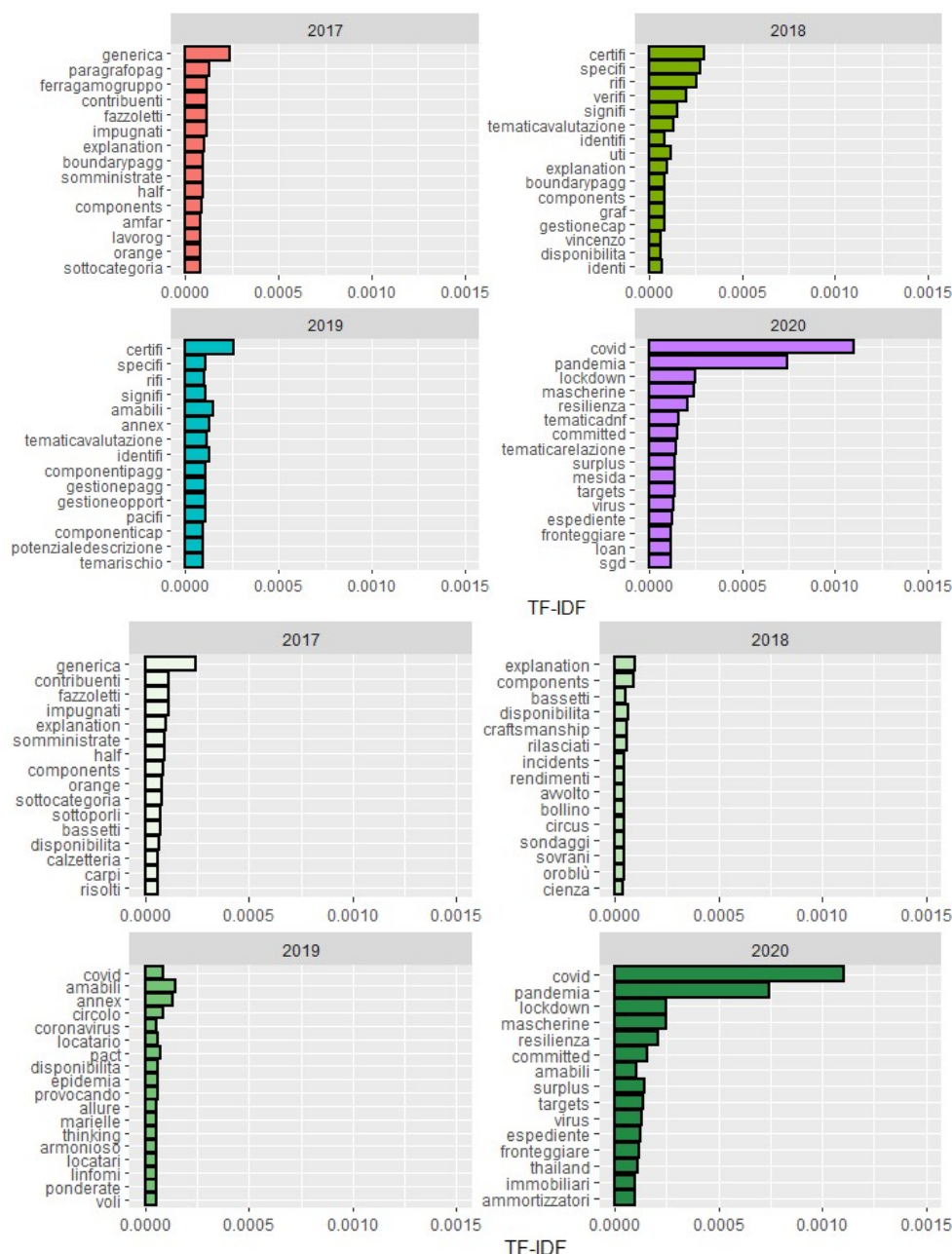
Dopo il primo step della tokenizzazione del corpus testuale, usando la libreria tidytext di R, sono stati calcolati, per ogni termine, la Term Frequency, cioè la frequenza relativa con cui questo appare e, successivamente, il logaritmo del rapporto tra il numero di anni totali (quattro) e il numero di anni in cui compare il termine.

Ottenuto il TF-IDF, la rappresentazione grafica dell'algoritmo, di più chiara interpretazione, ha consentito di individuare i pattern tematici presenti nel testo e di capire, altresì, se un determinato anno è stato condizionato da alcune tematiche specifiche non trattate con lo stesso livello di approfondimento negli altri anni.

Nella Figura 3 è evidente come, i documenti degli anni 2017, 2018 e 2019 non sembrano caratterizzati da tematiche particolari in quanto il TF-IDF è estremamente basso. Al contrario, le DNF del 2020 sono caratterizzate da termini inerenti alla pandemia da Covid-19, e tale risultato non è sorprendente, dato che la pandemia è scoppiata proprio agli inizi del 2020, protraendosi anche negli anni a venire, ma mostra il vantaggio di usare tale strumento nell'analisi testuale (Biafora et al., 2020).

Interessante da un punto di vista più tecnico è notare come nel grafico compaiano termini in inglese poichè, essendo le DNF in lingua italiana, alcuni termini stranieri sono molto rari.

Figura 3. Grafici TF-IDF per anno della Dichiarazione Non Finanziaria settore textile.



3.3 Analisi Bi-Grams

Dopo la fase della tokenizzazione è possibile studiare anche la contiguità di coppie di termini all'interno del corpus mediante costruzione dei Bi-Grams.

La libreria `igraph` di R consente di creare un oggetto `igraph` da un dataframe che contiene, oltre le coppie di termini contigui, il numero di volte che questi co-occorrono l'uno accanto all'altro. Un passaggio fondamentale, seppur non obbligatorio, prima della creazione dell'`igraph`, è la scelta della numerosità minima delle co-occorrenze per cui filtrare, poiché, quanto più bassa la numerosità, tanto più il grafico risulterà caotico poiché mostrerà tutte le coppie di termini adiacenti presenti nel dataset quel numero di volte. Per la presente analisi si è scelta una numerosità pari o superiore a 30.

La terza fase, sviluppata grazie alla libreria `ggraph`, permette di trasformare l'oggetto `igraph` in un grafico personalizzabile. Per rendere la visualizzazione di facile interpretazione si è scelto di mostrare solo le connessioni parola-parola più comuni (con almeno 10 co-occorrenze adiacenti) impostando lo spessore della freccia di collegamento tra termini in modo da essere proporzionale alla frequenza di associazione tra le due parole.

La Figura 4 mostra il Network Graph dei Bi-Grams più frequenti nelle DNF del settore tessile. Il tema centrale è certamente la “sostenibilità” ma l'attenzione è rivolta anche verso la “protezione”, anche dei dati, sempre più importanti per ogni tipo di azienda; verso il trattamento, la raccolta, la gestione e lo smaltimento dei “rifiuti”, considerati certamente pericolosi; verso i “diritti” umani, e la loro tutela e rispetto e verso la lotta alla “corruzione”, temi la cui trattazione è richiesta dalla norma di legge. Notiamo, inoltre, una particolare attenzione alle sostanze “chimiche” e “pericolose” che entrano a far parte del processo produttivo di questo settore considerato il più inquinante al mondo.

I rapporti di sostenibilità rappresentano lo strumento ideale per rendicontare lo sviluppo sostenibile delle aziende nei confronti degli stakeholders e generano al contempo una serie di benefici tangibili per le stesse. In primis, le imprese sono incentivate a adottare un approccio integrato, prendendo in considerazione sia aspetti meramente finanziari che ambientali, sociali e di buon governo, con la conseguenza di adattarsi e anticipare i fattori di cambiamento integrandoli nelle loro strategie (Khan et al, 2015).

Le DNF sono documenti che presentano molteplici formati e sono costituite da strutture di testo complesse, ricche di grafici, tabelle e immagini. Un'analisi qualitativa basata sulla lettura contemporanea di un numero elevato di documenti, se implementata da un essere umano, presenterebbe indubbie difficoltà nella comparazione dei contenuti e richiederebbe un'enorme quantità di risorse in termini di tempo. Le nuove tecnologie sono utili a superare tali limiti e supportare tale tipo di

analisi poiché possono essere di supporto con l'automatizzazione di ciascuno step del processo.

Il presente lavoro utilizza le principali tecniche di Text Mining per analizzare i trend ed individuare eventuali pattern all'interno delle Dichiarazioni Non Finanziarie di otto aziende del settore tessile nel quadriennio 2017-2020. I documenti in PDF sono stati scaricati tramite Web Scraping dal sito www.osservatoriodnf.it, la cui banca dati raccoglie numerosi report di sostenibilità in un unico repository. Questa fonte è preziosa per i ricercatori perché evita di dover scaricare le DNF singolarmente dai vari siti aziendali. I documenti in PDF sono stati poi trasformati in file testuali attraverso il processo di OCR (Optical Character Recognition) utilizzando il linguaggio di programmazione Python, in modo da renderli leggibili dalla macchina e analizzabili dal software R. I testi sono stati poi sottoposti ad un processo preparatorio di pulizia del contenuto e, infine, analizzati attraverso diverse tecniche di Text Analytics, implementate in R.

I principali risultati del presente lavoro dimostrano come sia possibile effettuare un'analisi approfondita su dati testuali non strutturati che, perfino in contesti molto standardizzati, consenta di estrarre informazioni sui contenuti specifici dei documenti programmatici. In particolare, i risultati hanno evidenziato un progressivo arricchimento dei contenuti delle DNF sia in termini quantitativi che qualitativi.

Inoltre, le tecniche introdotte ed utilizzate nel presente lavoro sono estremamente flessibili e possono risultare strategiche in diversi contesti, consentendo al ricercatore di individuare pattern latenti all'interno di grosse moli di dati testuali (Bolasco, 2013) con tempi e costi nettamente ridotti rispetto alle analisi qualitative tradizionali.

Riferimenti bibliografici

- Balluchi F., Furlotti K. (2019). *La responsabilità sociale delle imprese. Un percorso verso lo sviluppo sostenibile. Profili di governance e di accountability*. Giappichelli editore, Torino. 9788892110618.
- Biafora, A., D'Uggento, A.M., Toma, E., Grassia, M. G. Mining the sentiment during Covid 19 pandemic. In Book of Abstracts of Third International Conference on Data Science & Social Research, Bari, 10-11 December 2020. ISBN: 978-886629-051-3.
- Bolasco, S. (2013). *L'analisi automatica dei testi*, Carocci editore, Roma

- Cupertino S. and Vitale G. (2021). *Do companies integrate sustainability into management systems? Evidence from Italian non-financial reports*. Wolters Kluwer Italia, Milano. 9788813381578.
- Decreto legislativo 27 gennaio 2010, n. 39.
- Decreto legislativo 30 dicembre 2016, n. 254.
- García-Sánchez, I.M., Aibar-Guzmán, B., Aibar-Guzmán, C., Rodríguez-Ariza, L. 2020. “Sell” recommendations by analysts in response to business communication strategies concerning the Sustainable Development Goals and the SDG compass. *J. Clean. Prod.* 2020, 255, 120–194.
- Harymawan, I., Ardianto, R., Soeprajitno, R., Cahyaning Ratri, M., Paramitasari, Y. 2020. Text Mining on Sustainability Reporting: A Case Study. *Journal Of Security and Sustainability Issues* ISSN 2029-7017/ISSN 2029-7025 (online) 2020 Volume 9 Number May [http://doi.org/10.9770/jssi.2020.9.M\(4\)](http://doi.org/10.9770/jssi.2020.9.M(4)).
- Helfaya, A., Moussa, T. (2017) Do board’s corporate social responsibility strategy and orientation influence environmental sustainability disclosure? UK evidence. *Bus. Strategy Environ.* 26, 1061–1077.
- Informativa extra finanziaria: da compliance a governance strategica dei rischi e delle opportunità, KPMG, ottobre 2020.
- Khan, M., Serafeim, G., Yoon, A. (2015) Corporate Sustainability: First Evidence on Materiality. *The Accounting Review*, Vol. 91, No. 6, pp. 1697-1724.
- Li, J., Leonas, K.K. (2022) Sustainability topic trends in the textile and apparel industry: a text mining-based magazine article analysis, *Journal of Fashion Marketing and Management*, Vol. 26 No. 1, pp. 67-87. <https://doi.org/10.1108/JFMM-07-2020-0139>
- Marcello R., Magrassi L., Raoli E. (2018) Rilevanza (e materialità) nella disclosure non finanziaria Definizioni e criticità anche ai fini della compliance alle disposizioni del decreto legislativo 30 dicembre 2016, n. 254, sulla rendicontazione non finanziaria. Consiglio Nazionale dei Dottori Commercialisti e degli Esperti Contabili.
- Osservatorio DFN, Report 2019.
- Osservatorio Nazionale sulla Rendicontazione non Finanziaria ex D.Lgs. 254/2016 - 1° Report | ottobre 2018. Osservatorio DNF_I Report_Ottobre 2018_Deloitte Italia.pdf
- Osservatorio sulla Dichiarazione Non finanziaria. Deloitte Italia, marzo 2021.
- Phan H.T.P., De Luca F., Iaia (2020) The “Walk” towards the UN Sustainable Development Goals: Does Mandated “Talk” through Nonfinancial Disclosure Affect Companies’ Financial Performance? *Sustainability*, 12(6), 2324.
- Regolamento (UE) 2019/2088 del Parlamento europeo e del Consiglio del 27 novembre 2019 relativo all’informativa sulla sostenibilità nel settore dei servizi finanziari.

- Regolamento (UE) 2020/852 del parlamento europeo e del consiglio del 18 giugno 2020 relativo all'istituzione di un quadro che favorisce gli investimenti sostenibili e recante modifica del regolamento (UE) 2019/2088.
- Salvatore, C., Biffignandi, S., Bianchi, A. (2022) Corporate Social Responsibility Activities Through Twitter: From Topic Model Analysis to Indexes Measuring Corporate Social Responsibility Activities Through Twitter: From Topic Model Analysis to Indexes Measuring Communication Characteristics. *Social Indicators Research*. DOI: 10.1007/s11205-022-02993-8.
- Santamaria R., Paolone F., Cucari N., Dezi L. (2021) Non-financial strategy disclosure and environmental, social and governance score: Insight from a configurational approach. *Business Strategy and the Environment*, 30(4), 1993-2007.
- Spallini S., Milone V., Nisio A., Romanazzi P. (2021) The dimension of sustainability: A comparative analysis of broadness of information in Italian companies. *Sustainability*, 13(3), 1457.
- UN. Trasformare il nostro mondo: l'Agenda 2030 per lo Sviluppo Sostenibile, Organizzazione delle Nazioni Unite A/RES/70/1.
- Wan Te Liew, Adhitya, A., Srinivasan, R. (2014) Sustainability trends in the process industries: A text mining-based analysis. *Computers in Industry*. Vol. 65, Issue 3, April 2014, 393-400
- World Economic Forum, Global Risks Report 2020.

Sitografia

1. <http://www.cran.r-project.org/>
2. <http://www.consob.it/web/area-pubblica/soggetti-che-hanno-pubblicato-la-dnf>
3. <http://www.esg-rating.it/informativa-non-finanziaria-di-sostenibilita-dnf/>
4. <http://www.gruppoiren.it>
5. <http://www.nltk.org/>
6. <http://www.osservatoriodnf.it/it>
7. <http://www.toptal.com/python/twitter-data-mining-using-python>