Università
Ca'Foscari
Venezia

Corso di Laurea Magistrale
in Economia e Gestione delle Aziende

Tesi di Laurea

La Business Intelligence nell'era dei Big Data

Relatore
Raffaele Pesenti

Laureando
Giovanni Lattanzi
Matricola 828275

Anno Accademico
2017 / 2018

Indice

INTRODUZIONE 1

1 L'INFORMAZIONE ED IL SISTEMA INFORMATIVO 4

1.1 Il bisogno di conoscenza ed il valore dell'informazione *4*

1.2 Il tema della conoscenza nella teoria d'impresa *6*

1.3 Incertezza, razionalità limitata *8*

1.4 Sistema Informativo e Sistema Informatico *10*

1.5 Business Intelligence *16*

1.5.1 Architettura di Business Intelligence 17

1.5.2 Il valore della Business Intelligence 29

2 BIG DATA 31

2.1 Dati: un patrimonio importante *31*

2.2 Introduzione ai Big Data *36*

2.3 Creare valore con i Big Data *40*

2.4 Osservatorio Big Data *43*

3 PROGETTARE LA BUSINESS INTELLIGENCE 50

3.1 L'importanza della metodologia *50*

3.2 Le figure professionali *52*

3.3 Le fasi del progetto *54*

3.3.1 Analisi preliminare e giustificazione del progetto 54

3.3.2 Pianificazione delle attività 57

3.4 Regole per il successo ed errori da evitare *69*

4 SENTIMENT ANALYSIS 74

4.1 Data Mining *74*

4.1.1 Data Mining è un Business Process 76
4.1.2 Modelli e regole Significativi 77
4.1.3 Knowledge Discovery in Database (KDD) 77
4.1.4 Uno sguardo ravvicinato al data mining 79
4.2 Introduzione al Text Mining *82*
4.2.1 Il Processo di Text Mining 85
4.3 La Sentiment Analysis: Concetti e Metodologie *92*
4.3.1 Metodi di Classificazione dei testi 96
4.4 Twitter Sentiment Analysis *103*
4.4.1 R: strumento di Analisi 105
4.4.2 Sentiment Analysis in Twitter: Implementazione 105
4.4.3 Il sistema ed il Mondo Reale: Sentiment in Azienda 121
CONCLUSIONI 141
BIBLIOGRAFIA E SITOGRAFIA 143

Introduzione

Il lavoro svolto durante questa tesi di laurea mira alla progettazione e allo sviluppo di un sistema di Business Intelligence (BI) per l'analisi dei Big Data. Nello specifico, il progetto di tesi prevede la creazione di un sistema automatizzato di Sentiment Analysis su dati provenienti da Twitter. In sostanza si andrà a realizzare un sistema che giornalmente estrae, filtra, pulisce e analizza tweet su una serie di aziende preselezionate. L'idea generale è partita da una considerazione: negli ultimi anni il sistema informativo aziendale si è evoluto costantemente, andando a considerare dati e informazioni che sono aldilà dei confini aziendali (si pensi a tutte le informazioni che l'azienda produce o ricerca nei social media). L'avvento dei Big Data, risorse informative ad alto volume, alta velocità e ad alta varietà, ha accelerato questo processo, spingendo le aziende a cercare valori in informazioni al di fuori dei propri confini, integrando i dati interni con questi nuovi dati esterni. Questo processo ha spinto le società a rivedere il ruolo svolto dalla Business Intelligence, arricchendolo di strumenti e procedure nuove e creando ulteriori figure professionali.

Con quest'elaborato si cerca di rispondere ad una serie di domande:

- come si struttura un sistema di BI per l'analisi dei Big Data?
- Per le aziende investire in questi processi è solo un rischio o un vantaggio competitivo?
- I Big Data possono realmente creare valore all'azienda?

Per rispondere a queste domande, il progetto di tesi prevede la creazione di un sistema automatizzato di Sentiment Analysis su dati provenienti da Twitter. Oggi gran parte delle aziende è entrata nel modo dei Social Network, strumento che viene da loro utilizzato principalmente per fini di marketing e branding. Con Sentiment Analysis, si intende l'analisi di brevi testi, tweet, con lo scopo di determinare il sentimento, l'opinione espressa dall'autore del testo. In altre, parole si cerca di capire il l'orientamento, il giudizio dell'individuo che ha creato il messaggio. La nostra Sentiment Analysis cercherà di determinare il sentimento, l'opinione degli utenti di Twitter verso un insieme di aziende, brand.

Si è scelta questo tipo di analisi per due ragioni semplici: la prima riguarda la tipologia di dati che verranno trattati, ossia dati testuali. Nei vecchi sistemi di BI era molto raro la trattazione di questa tipologia di dati, perché non si disponeva

della tecnologia e della conoscenza adatte a trattarli. La seconda ragione invece si riaggancia al discorso all'inizio dell'introduzione; i dati espressi in Twitter sono informazioni che stanno fuori dai confini aziendali, alle quali viene associato poco valore. L'obiettivo di questa tesi è quello di mostrare come anche questa tipologia di dati, possa creare valore per l'azienda.
Presentiamo di seguito una sintesi del contenuto dei capitoli che seguiranno.

Capitolo 1: Nel primo capitolo viene brevemente descritto il Sistema Informativo, soffermandoci nel dettaglio sul Sistema Informatico. Nella seconda parte del capitolo viene descritta nel dettaglio l'architettura della Business Intelligence, e mostrando in che modo la BI produce valore per l'azienda.

Capitolo 2: Nel secondo capitolo viene fatto un focus sui Big Data, descrivendo gli elementi caratteristici.

Capitolo 3: Proseguendo viene introdotta una metodologia per implementare un sistema di Business Intelligence. La metodologia esposta è frutto di una ricerca in letteratura e da best pratice apprese sul "campo", nel mio lavoro di consulente.

Capitolo 4 : Nel quarto e ultimo capitolo viene descritto brevemente il mondo del Data Mining, dando maggior risalto alla tecniche di Sentiment Analysis (in particolare al metodo bayesiano). Nella seconda parte del capitolo vengono descritti i vari step per la creazione del modello dati per l'analisi, sia a livello teorico sia nella pratica del caso reale. Infine, vengono esposti i risultati dell'analisi attraverso una serie di report.

1 L'informazione ed il Sistema Informativo

1.1 Il bisogno di conoscenza ed il valore dell'informazione

A Delfo, ai piedi del Monte Parnaso, si ergeva un famoso tempio di Apollo, dove le sacerdotesse davano responsi tanto ambigui ed oscuri da essere spesso pericolosamente sviati. Ne fu vittima anche Filippo il Macedone, che, al tempo della sua spedizione persiana, ricevette questo oracolo: "La vittima designata, incoronata a morte, sta davanti all'altare". Egli ritenne si trattasse del re di Persia, ed invece si parlava di lui. I nostri "oracoli" moderni talvolta sanno essere altrettanto ambigui e svianti.

T.S. Eliot scrisse nel dramma La rocca: "Dov'è la conoscenza che abbiamo smarrito nell'informazione?"[1], a questa domanda un analista di dati potrebbe aggiungere: "Dov'è l'informazione che abbiamo perduto nei dati?". Vale l'assioma che: l'informazione non sempre si trasforma in conoscenza[2]. Ci sono molte ragioni per essere cauti circa la validità e l'interpretazione dei dati, ma il desiderio di conoscenza e quindi la necessità di informazione è antico quanto l'uomo e molti problemi, che dobbiamo affrontare oggi, hanno origine remota. Non siamo in grado di rispondere in modo esplicito ai quesiti enunciati, ma qualunque domanda ci si ponga sulla tecnologia dell'informazione si dovrebbe tener presente che la funzione che essa svolge non è una novità, bensì un fatto antico nella vita e nella società umana. Nel passato parecchie tecnologie hanno favorito varie modalità di

[1] TS. Eliot - *Choruses from «The Rock»* – Faber & Faber, London, 1936: "*Where is the knowledge we have lost in information?".*

[2] "Nell'epoca di Internet e dell'e-business, i dati, l'informazione e la conoscenza sono considerati elementi fondamentali di successo per le imprese. Ma non bisogna confondere i tre livelli. Raccogliere ed organizzare i dati non significa automaticamente generare delle informazioni, e anche una volta recepite le informazioni non si trasformano automaticamente in conoscenza." Shinynews.it, *Business Intelligence: conosciamola meglio*, in Business Online, maggio 2006.

informazione: la stampa ha favorito l'informazione scritta; il telefono e la radio hanno privilegiato l'informazione parlata; la fotografia e la televisione hanno agevolato l'informazione visiva. Oggi la tecnologia digitale permette di reperire, trasmettere, archiviare ed elaborare grosse moli di dati, che diventano così informazione. Pur essendo molto potenti le tecnologie dell'informazione non sono altro che strumenti nelle mani dell'uomo: se usati in modo corretto possono essere molto utili a soddisfare il bisogno di conoscenza, un loro uso scorretto può per contro produrre effetti dannosi. Per poterle utilizzare nel migliore dei modi bisogna comprenderne il valore e conoscerne il giusto uso. Oggi il fabbisogno di informazioni per dirigere un'azienda è cresciuto rapidamente, i profondi mutamenti avvenuti nei modi di produzione e nelle relazioni economiche, hanno conferito una crescente rilevanza agli scambi di beni immateriali, riconducibili in larga misura a trasferimenti di informazioni. A partire dall'ultimo decennio del secolo scorso il contesto socioeconomico entro il quale si svolgono le nostre attività viene indicato come *società dell'informazione e della conoscenza*. Sono due i fattori principali che hanno spinto e accelerato i processi di trasformazione in atto: da un lato la globalizzazione, intesa come crescente interdipendenza fra le economie dei diversi paesi, che ha condotto allo sviluppo di un'unica *economia globale* segnata da un alto livello di integrazione. Dall'altro le nuove tecnologie dell'informazione, caratterizzate dalla massiccia diffusione di *internet* e di dispositivi *wireless*, che hanno consentito di trasferire grandi quantità di dati ad elevata velocità e di sviluppare forme evolute di comunicazione[3]. Avere un accesso immediato ed efficace alle informazioni e alla conoscenza presenta vantaggi per i diversi attori dell'ambiente socioeconomico. Da un lato gli individui possono reperire più liberamente e velocemente notizie, accedere a servizi con maggior facilità, effettuare operazioni commerciali e bancarie on-line. Dall'altro le imprese possono sviluppare prodotti e servizi innovativi più vicini alle esigenze dei clienti e ottenere vantaggi competitivi utilizzando in maniera efficace le conoscenze maturate[4]. I processi direzionali, che sono chiamati a regolare il comportamento del sistema di impresa, sono sempre più qualificati da un'impostazione scientifica. Dall'uso di raffinate tecniche e dall'impiego di sofisticati strumenti di calcolo, di elaborazione

[3] Vercellis C., *Business Intelligence. Modelli matematici per le decisioni*, Mc Graw Hill, Milano, 2006, p.1
[4] Ivi, p.1

e di trasmissione delle informazioni[5]. Vi è quindi la necessità di un flusso continuo e coordinato di informazioni, che devono essere raccolte, elaborate ed inviate a tutti coloro che operano all'interno della struttura organizzativa. Il problema di raccogliere, elaborare e fare circolare le informazioni riveste, dunque, un ruolo critico soprattutto nelle grandi organizzazioni e aziende. In questi casi, infatti, non solo si amplia il volume di informazioni necessarie ma si allungano anche i circuiti di comunicazione, aumentano i punti di interconnessione nell'integrazione dei vari flussi di dati, si complicano le modalità di elaborazione e diffusione della documentazione. Da queste premesse si comprende come acquisire dati e informazioni per creare conoscenza sia un elemento centrale all'interno di un'impresa. Ma in che modo la creazione di conoscenza è un vantaggio competitivo per l'impresa? Per rispondere a tale domanda è necessario comprendere in che modo il tema della conoscenza si è inserita all'interno della teoria economica e nello specifico nella teoria d'impresa.

1.2 Il tema della conoscenza nella teoria d'impresa

Il tema della conoscenza è sempre stato trattato in modo indiretto nella questione riguardante la natura dell'impresa. Solo nelle teorie più recenti l'impresa viene vista come struttura di gestione e di valorizzazione della conoscenza. Ma affrontiamo la questione, del ruolo della conoscenza nella natura dell'impresa, in linea con l'evoluzione delle teorie sull'impresa. Nella teoria neoclassica l'impresa è vista come modalità tecnica per trasformare i prodotti. La conoscenza è pertanto qui riconosciuta come tecnologia a disposizione dell'impresa, ma non si è in grado di dire nulla sul perché questa conoscenza esista e sui processi di sua acquisizione e sviluppo. Tutte le imprese dispongono della medesima funzione di produzione, possedendo le medesime conoscenze tecnologiche. Nella teoria dei costi di transazione, si evidenzia il ruolo dell'organizzazione come modalità di governo più efficiente rispetto alle transazioni di mercato. La capacità di governare i processi attraverso la gerarchia porta ad una maggiore efficienza all'interno dell'impresa, efficienza che non può che derivare da quella che oggi chiameremo la conoscenza manageriale[6]. Nel modello della specializzazione flessibile, si evidenzia come la

[5] C. Caramiello, *L'indagine prospettiva nel campo aziendale*, Cursi, Pisa, 1965
[6] S. Vicari. *Conoscenza e impresa*, Sinergie Italian Journal of Management, 2011, p.49

possibilità che gruppi di lavoratori hanno di specializzarsi in singole fasi del ciclo produttivo comporti un aumento dell'efficienza degli impianti e del rendimento dei fattori produttivi impiegati[7]. Diretta conseguenza di tale approccio è il distretto industriale: basato per l'appunto su una divisione del lavoro fra imprese, in cui tutte danno il loro contributo al dinamismo collettivo, mantenendo un alto tasso di innovazione, soprattutto del processo produttivo, e un'accentuata capacità di reazione alle esigenze del mercato.[8] Nel modello distrettuale la conoscenza ricopre un ruolo molto importante, lo sviluppo di conoscenze specializzate e la localizzazione di tali conoscenze in una specifica area geografica è uno dei fattori determinanti per la nascita dei distretti. In tale modello produttivo la conoscenza è vista come il patrimonio culturale e imprenditoriale di imprese, imprenditori, e istituzione di uno specifico territorio. Come abbiamo brevemente visto nei vari modelli d'impresa, il tema della conoscenza è emerso lentamente, sino a divenire veicolo per acquisire un vantaggio competitivo. Il vantaggio competitivo dell'impresa si configura come la capacità di cercare ed elaborare informazioni (sulla clientela, sui concorrenti, sulle tecnologie, sulle risorse) e di tradurle in forme organizzate di produzione. Al massimo livello di astrazione, la competitività dell'impresa, è il risultato della capacità di incorporare nelle proprie attività: *scienza* (informazioni sulle regole di funzionamento dei sistemi fisici e sociali) e *conoscenza* (informazioni sugli stati dei sistemi fisici e sociali, sulle esigenze dei soggetti, sulle modalità tecniche con le quali produrre i beni richiesti). Il passaggio di un modello d'impresa in cui la conoscenza diviene il più importante fattore di sviluppo e di competitività pone però una serie di problemi concettuali che sono tuttora in fase di esplorazione. Sia nella sua dimensione scientifica (know-why) sia nella sua dimensione tecnica (know-how, conoscenza relativa al modo di operare per ottenere un certo risultato) la conoscenza è una risorsa del tutto singolare in quanto è qualcosa di distribuito (una volta "prodotto" la conoscenza può essere riprodotta a costi molto bassi se non nulli, ossia condivisa) e non concentrabile (la nuova conoscenza in grado di creare valore è sempre più complessa e articolata e pertanto difficilmente può essere acquisita da un solo soggetto o da una sola organizzazione). Date queste caratteristiche la conoscenza si diffonde maggiormente in quelle imprese che hanno la capacità di costituire delle "filiere

[7] G. Corò e S.Micelli, *I nuovi distretti produttivi*, Marsilio Editore 2010, p.71
[8] G. Volpato, *Economia e Gestione delle Imprese*, Carocci Editore, 2006, p.42

cognitive", ossia delle reti di imprese in grado di dividersi il lavoro cognitivo e di comunicarsi reciprocamente i risultati conseguiti, attraverso l'uso di linguaggi condivisi[9]. Siamo entrati in un'era nuova caratterizzata da una possibilità di sviluppo data dal riconoscimento pieno del ruolo che la conoscenza gioca nell'economia. Grazie alle possibilità legate allo sviluppo tecnologico, una risorsa fondamentale per lo sviluppo dell'economia, la conoscenza, è disponibile a basso costo e senza vincoli di scarsità. Pertanto l'economia si trasforma: da scienza che tratta dell'allocazione ottimale di risorse scarse diviene scienza della riproduzione e dell'utilizzo di risorse cognitive. La questione più critica oggi è come accedere alla quantità di conoscenza praticamente illimitata che sta fuori dall'impresa, inserendola dentro i circuiti di riproduzione della conoscenza all'interno delle reti di relazioni, interne ed esterne, che costituiscono la moderna impresa. In futuro le imprese capaci del maggiore sviluppo non saranno quelle dotate della migliore conoscenza, ma quelle capaci di accedere alla conoscenza inserendole nei propri meccanismi riproduttivi nel modo più efficace ed efficiente[10].

1.3 Incertezza, razionalità limitata

L'uomo, per poter assumere delle decisioni, ha bisogno di informazioni. Sebbene si possa pensare che aumentando il numero di queste aumenti la probabilità di una scelta migliore, gran parte delle decisioni sono la risultante di un processo di acquisizione di pochissime informazioni. Infatti, a causa della nostra natura non si potranno assumere tutte le informazioni rilevanti ai fini decisionali, di conseguenza la percezione delle informazioni, sarà selettiva e la loro elaborazione sarà sequenziale. L'uomo, infatti, ha una capacità ridotta di gestire il flusso informativo, derivante da quattro vincoli fondamentali, quali[11]:

- Problemi di attenzione: non si possono eseguire e/ controllare più cose contemporaneamente. Inoltre, il funzionamento dell'attenzione è soggettivo e legato alle esperienze, ai gusti e alla motivazione di ciascun individuo.
- Problemi di memoria: oltre ad essere limitata e fallace, è difficile trasmettere la propria esperienza.

[9] Ivi, p.49
[10] S. Vicari. *Conoscenza e impresa*, Sinergie Italian Journal of Management, 2011, p.64
[11] M. Martinez, *Organizzazione, informazioni, tecnologie*. Il mulino, 2004.

- Problemi derivanti dalle limitate capacità di comprensione: non si riesce sempre a gestire ed utilizzare le informazioni in modo da creare relazioni tra eventi e fenomeni, a capire il loro reale valore, ad interpretarle.
- Problemi di distorsione della comunicazione tale problema è relativo alle difficoltà che emergono nel trasmettere le informazioni tra categorie differenti (per esempio tra gli utenti finali di un software aziendale e i suoi progettisti).

Di conseguenza la qualità delle decisioni dipenderà dal modo con cui tali informazioni verranno elaborate e, più precisamente, dalla quantità e dalla completezza delle informazioni raccolte e dalla capacità di elaborazione razionale del soggetto. Ne consegue che, per poter scegliere, si acquisiranno un numero di informazioni ritenuto adeguato ma non ottimale. Secondo la teoria delle decisioni la scelta razionale dell'individuo sarà limitata, dovendo l'uomo far fronte ad una conoscenza incompleta delle alternative di scelta, un'incapacità di valutare gli effetti che ciascuna alternativa produce e, infine, a delle capacità di calcolo inadeguate per ciascun problema. Il risultato sarà pertanto il frutto di un compromesso tra le spinte, le influenze e i limiti che caratterizzano la scelta dell'individuo. Tali limiti sono di diversa natura: *informativa* (in quanto non si possono reperire e comprendere tutte le informazioni utili ai fini di una decisione), *inconscia* (nelle scelte influiranno elementi soggettivi come la personalità, le esperienze, i desideri) e *sociale* (per esempio: i bisogni di appartenenza, di status, ma anche la cultura e i valori propri della società in cui si è nati)[12]. Pertanto gli individui agiscono in condizioni di incertezza[13]. Uno dei maccanismi con cui l'impresa può ridurre i limiti alla razionalità, è costituito dai sistemi informativi in quanto questi consento di ridurre i flussi di informazioni rilevanti per i processi decisionali aumentando le conoscenze del decisore sulle alternative a disposizione e sulle conseguenze della sua scelta.

[12]A. Spada, *La razionalità in Herbert Simon e le ambiguità tra behavioral economics ed economica cognitiva*, Department of Economics and Statistics "S. Cognetti De Martiis" and CESMEP, Working Paper, 2010

[13] "Vi è Rischio quando il decisore potrà attribuire una distribuzione di probabilità oggettive al verificarsi di eventi ignoti, perché associata a fenomeni probabilistici controllabili come, ad esempio, l'interruzione del ciclo di produzione per incidenti agli impianti. Si parla, invece di Incertezza quando tale distribuzione non può essere attribuita e dunque la scelta diviene un problema, in quanto l'individuo sarà incapace di distinguere, tra le alternative che ha a disposizione, quella più conveniente". M. Grillo, F. Silva, *Impresa Concorrenza ed Organizzazione*, Carocci, 2007.

1.4 Sistema Informativo e Sistema Informatico

Il sistema informativo aziendale può essere definito come l'insieme delle attività, dei processi, delle risorse tecnologiche (hardware e software) e delle persone preposte alla gestione dell'informazione. Possiamo dunque chiederci cosa significhi gestire l'informazione, presentando, come risposta, il seguente elenco di attività[14]:

- Acquisizione dei dati;
- Elaborazione dei dati al fine di produrre informazioni[15];
- Archiviazione dei dati grezzi e delle informazioni;
- Trasmissione dei dati e delle informazioni;
- Presentazione dei dati e delle informazioni.

Il sistema informativo aziendale è composto da due elementi principali[16]:

- da una parte informatizzata chiamata sistema informatico aziendale;
- da una parte non automatizzata come le conversazioni frontali e telefoniche, i documenti cartacei strutturati e non strutturati, l'organizzazione aziendale, la prassi operativa, la prassi decisionale ecc.

Il sistema informatico è la porzione automatizzata del sistema informativo, ed è composto dall'insieme degli strumenti informatici (calcolatori, software di base, programmi applicativi, etc.) impiegati per il trattamento automatico delle informazioni allo scopo di migliorarne l'efficienza. Possiamo dunque comparare il sistema informativo aziendale ad un vero e proprio sistema nervoso dell'azienda, che permette alle informazioni di arrivare ad ogni "organo" aziendale. Il sistema informatico, invece, è il cervello il quale consente all'azienda di controllare, pianificare e gestire in modo integrato tutte le attività, nonché di elaborare velocemente una maggiore quantità di dati ed informazioni.

---

[14] A. Rezzani, "*Data Skills Understandig the World*", dal sito: www.dataskills.it

[15] "Il concetto di dato è profondamente diverso da quello di informazione: il dato rappresenta la misurazione di un evento, un fatto oggettivo espresso attraverso numeri o parole. L'informazione è ottenuta attraverso l'elaborazione di un dato (o di un insieme di dati) e ha la caratteristica di assumere un significato, uno scopo e/o un contesto per il destinatario". A. Clerici, M. De Pra, *Informatica e web*, EGEA, 2012.

[16] Da appunti corso universitario: "Sistemi di supporto alle decisioni aziendali"

In un sistema informatico le informazioni vengono rappresentate per mezzo di dati, che diventano informazioni dopo un processo di interpretazione. Spesso non viene fatta distinzione tra i due termini, anche se in realtà denotano due diversi livelli di conoscenza[17]:

- I dati sono immediatamente presenti alla conoscenza prima di ogni elaborazione;
- Le informazioni sono i dati opportunamente interpretati.

Il sistema informativo svolge differenti operazioni che possono essere classificate in base alle esigenze che scaturiscono dalle attività dei diversi attori dell'organizzazione aziendale, secondo lo schema della piramide di Anthony[18]. Tale schema individua tre categorie di attività aziendali:

Fonte: A. Figura 1: Rezzani, "*Data Skills Understandig the World*", dal sito: www.dataskills.it

- Attività Strategiche: definizione degli obiettivi aziendali e delle politiche aziendali volte al loro raggiungimento.
- Attività Tattiche: allocazione efficace ed efficiente delle risorse aziendali al fine di conseguire gli obiettivi posti dal management strategico. Rientrano anche in questo gruppo le attività di controllo del raggiungimento di tali obiettivo.

[17] Da appunti corso universitario: "Reti e Sistemi Informativi"
[18] A. Rezzani, "*Data Skills Understandig the World*", dal sito: www.dataskills.it

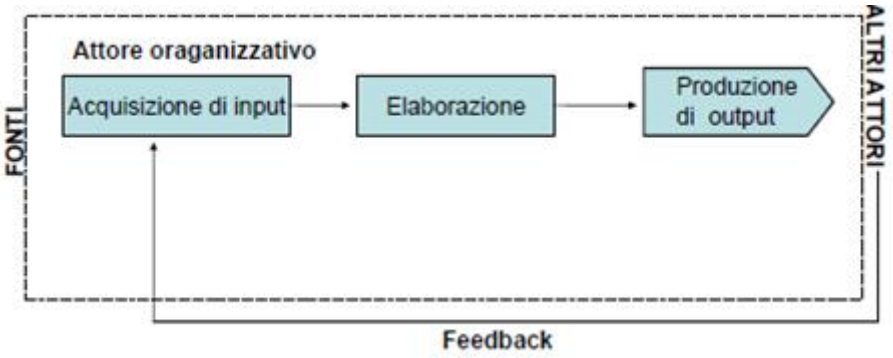
- Attività Operative: rappresentano l'operatività corrente (gestione ordini, magazzino, fatturazione, etc.).

La tipologia di informazione necessarie, al personale aziendale per prendere decisioni, sarà di conseguenza differente in base al livello di attività che viene ricoperto[19]:

1. Il personale dirigenziale, richiede principalmente informazioni analitiche di sintesi, trasmesse attraverso un supporto grafico, che gli permette di avere una visione dell'andamento aziendale o dei singoli settori, il più velocemente possibile. In tale livello le esigenze informative variano nel tempo, e sono poco prevedibili, spesso riguardano previsioni future.
2. Le direzioni funzionali necessitano anche loro di dati analitici sintetici, tuttavia le richieste sono relative a dati consuntivi e le esigenze rimangono stabili nel tempo, permettendo la costruzione di elaborazioni ripetitive.
3. Il personale esecutivo, invece, data la natura operativa delle proprie attività esige dati di dettaglio, forniti con tempestività (possibilmente in tempo reale).

Da questo si evince, che all'interno di un'azienda circolano differenti tipologie di informazioni, necessarie per svolgere le varie attività aziendali. Il sistema informativo crea, quindi informazioni su tutti gli aspetti della gestione.

Figura 2: Rappresentazione del flusso informativo o funzioni del sistema informativo[20]

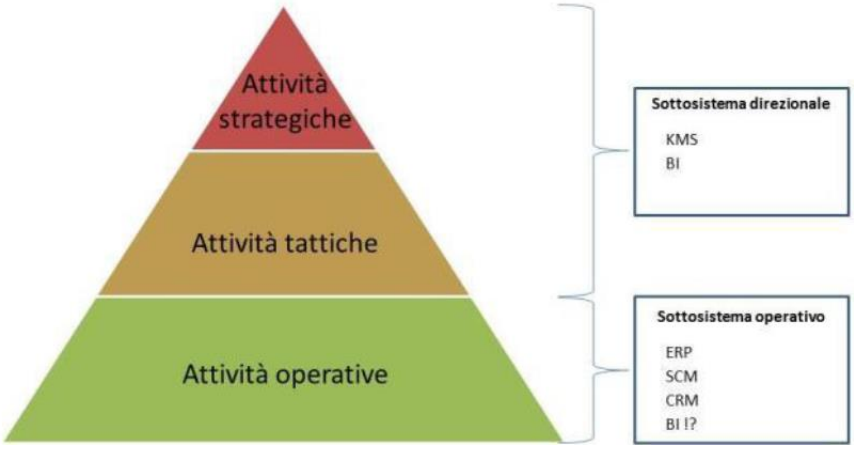


[19] A. Rezzani, "*Data Skills Understandig the World*", dal sito: www.dataskills.it

[20] Fonte "*M. Martinez, Organizzazione, Informazioni e Tecnologie*." Il Mulino, 2004

L'immagine qui sopra rappresenta le tre attività principali del flusso informativo, che producono le informazioni necessarie all'azienda per prendere decisioni, controllare le operazioni, analizzare i problemi e creare nuovi prodotti o servizi. L'attività di "input" consiste in acquisire i dati grezzi dall'interno dell'azienda o dall'ambiente esterno che grazie all'attività di "elaborazione", verranno convertiti in una forma utile. Attraverso l'attività di "output", invece, si trasferiscono le informazioni elaborate precedentemente alle persone che le utilizzeranno, ovvero alle attività per le quali verranno impiegate. Infine il "feedback" è il responso prodotto dai membri dell'azienda per valutare e/o raffinare l'input. Come si vede nello schema, il flusso informativo è circolare: l'informazione che viene elaborata servirà a supportare le decisioni che verranno eseguite, successivamente, grazie alla trasmissione dei risultati, vi sarà un controllo che porterà, infine, all'assunzione di altre decisioni[21]. I dati che entrano nel flusso informativo possono nascere direttamente all'interno dell'azienda durante lo svolgimento dei vari processi aziendali o essere acquisiti come risultato delle relazioni con soggetti esterni. Analogamente tali dati possono essere destinati al "consumo" interno" o essere destinati a terzi. Come abbiamo visto precedentemente, il sistema informativo svolge differenti tipologie di attività, di conseguenza sono stati ideati diverse categorie di software applicative ognuna preposta a svolgere una o più attività. Se associamo ciascun livello di attività nella piramide di Anthony un elenco di categorie software si ottiene lo schema seguente:

Figura 3 Fonte: A. Rezzani, "*Data Skills Understandig the World*", dal sito: www.dataskills.it



[21] M. Martienz, "*Organizzazione, Informazione e Tecnologie*", Il Mulino, 2004.

Le categorie di applicazioni, che nello schema appaiono con i loro acronimi, sono descritte brevemente di seguito. Innanzitutto vengono utilizzati nelle attività strategiche e tattiche i seguenti gruppi di applicazioni software[22]:

- **Knowledge Management System (KMS):** è il sistema di gestione della conoscenza, ossia l'insieme degli strumenti software per la ricerca, l'identificazione, la strutturazione di tutte le informazioni riguardanti le attività svolte in azienda. Un esempio è individuabile nei software di gestione documentale, che consentono di creare, organizzare e consultare vere e proprie biblioteche elettroniche costituite dai documenti aziendali.
- **Business Intelligence (BI):** finalmente incontriamo per la prima volta l'argomento cardine di questo primo capitolo. Al momento ne daremo una definizione generale, che pur se incompleta, rende bene l'idea su cosa sia la business intelligence: "è l'insieme di strumenti software che forniscono un supporto alle decisioni, tramite la trasformazione dei dati aziendali in informazioni".

Le attività operative invece si servono delle seguenti categorie di applicativi:

- **Enterprise Resource Planning (ERP):** tali software sono strumenti gestionali che interessano praticamente tutte le aree aziendali, dalla gestione degli ordini alla fatturazione, dalle paghe al bilancio, dai pagamenti alla contabilità.
- **Supply Chain Management (SCM):** i software di questa categoria gestiscono la catena di fornitura (sia relativa agli acquisti, sia relativa alle vendite), velocizzando la trasmissione di ordini di acquisto, ottimizzando le scorte di magazzino, definendo piani di produzione in base alla domanda, monitorando i processi di consegna.
- **Customer Relationship Management (CRM):** tali software supportano l'attività di contatto con il cliente e forniscono strumenti per la gestione delle vendite, delle campagne di marketing e del customer service[23].

[22] A. Rezzani, "*Data Skills Understandig the World*", dal sito: www.dataskills.it

[23] *Gartner Group* definisce il CRM come segue: "Il CRM è una strategia di business il cui risultato è l'ottimizzazione della profittabilità, del fatturato e della soddisfazione dei clienti perché attraverso la profilatura dei clienti, lo sviluppo di comportamenti che portano alla soddisfazione dei clienti e l'implementazione di processi che hanno al centro il cliente. La tecnologia deve favorire un maggiore interazione con lo stesso e una maggiore integrazione attraverso i canali di comunicazione e tutte le funzioni aziendali di backoffice". Dal sito: www.gartner.com

- **Business Intelligence (BI):** la Business Intelligence può offrire un valido supporto anche alle attività operative: nel prossimo paragrafo vedremo come.

Le esigenze del management stanno vivendo un'evoluzione che ha portato le applicazioni analitiche ad avere sempre più un ruolo strategico all'interno dell'azienda, diventando così uno strumento prioritario per la definizione degli orientamenti aziendali. Nello stesso tempo aumenta il fabbisogno di una reportistica sempre più complessa e cresce la necessità di condividere le informazioni all'interno dell'azienda creando collaborazioni fra diversi ruoli funzionali. È pertanto fondamentale, per l'azienda, dotarsi di un sistema informativo ben progettato che riesca a valorizzare ciascuno degli elementi di cui è composto (persone, procedure, dati, infrastrutture, principi e valori ispiranti)[24].
Le caratteristiche principali di un buon sistema informativo sono chiarezza, correttezza e completezza:

- Chiarezza: i dati raccolti devono essere precisi;
- Correttezza: nelle procedure con le quali i dati vengono raccolti e trasformati in informazioni;
- Completezza: il sistema informativo deve rispecchiare al meglio il sistema aziendale.

Per quanto riguarda il patrimonio dei dati, tuttavia, il sistema informativo deve necessariamente risolvere alcuni problemi che affliggono molte aziende[25]:

- I dati provengono da fonti diverse ed eterogenee (per tecnologia, formato, ubicazione);
- A volte i dati sono ridondanti, imprecisi e incompleti;
- Sono di difficile fruizione;
- Sono difficilmente integrabili.

Di conseguenza, il sistema informativo deve intraprendere operazioni di eliminazione delle ridondanze, "pulizia" dei dati, unificazione di strutture e formati. Inoltre occorre che i dati possano trasformarsi facilmente in informazioni e possa

[24] A. Rezzani, "*Data Skills Understandig the World*", dal sito: www.dataskills.it
[25] A. Rezzani, "*Data Skills Understandig the World*", dal sito: www.dataskills.it

essere facilmente fruibili da parte degli attori aziendali. L'implementazione di un sistema di Business Intelligence è il mezzo per raggiungere l'obiettivo di valorizzare il patrimonio dei dati aziendali.

1.5 Business Intelligence

La rapida panoramica del sistema informativo aziendale ci ha fornito una prima definizione di Business Intelligence (BI) e aiutato a collocarlo all'interno delle attività strategiche e tattiche. Il management trova così supporto analitico, storico e previsionale alle proprie decisioni. Tuttavia, la BI è stata collocata anche all'interno delle attività operative, poiché essa sta assumendo un ruolo sempre più importante anche nelle normali attività giornaliere delle aziende[26].

La prima definizione di BI è stata proposta da Howard Dresner, analista di Gartner Group nel 1989, per descrivere gli strumenti in grado di soddisfare le esigenze dei manager aziendali[27]. La definizione originale di Dresner è la seguente: "*Business Intelligence describes the enterprise's ability to access and explore information, often contained in a Data Warehouse, and to analyze that information to develop insights and understanding, which leads to improved and informed decision making. BI tools includes: ad hoc query, report writing, decision support systems (DDSs), executive information systems (EISs) and, often, techniques such as statistical analysis and on line analytical processing (OLAP)*". Nella definizione di Dresner compaiono diversi concetti che descriveremo successivamente; prima è importante sottolineare che, in base al punto di vista di chi definisce la BI, questa assume accezioni diverse. A volte con il termine BI si indicano i processi aziendali di creazione delle informazioni, oppure gli elementi tecnologici (hardware e software) che sono utilizzati per tali processi, oppure ancora l'insieme stesso delle informazioni che il management utilizza per intraprendere le proprie azioni. Altri, invece, definiscono la BI come un'architettura di applicazioni di supporto alle decisioni. Vi sono inoltre numerose definizione fornite dai *vendor* di strumenti di

[26] Ivi

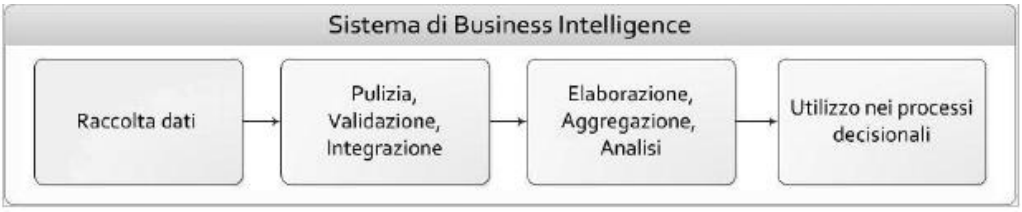
[27] Bisogna dire che il termine Business Intelligence è comparso per la prima volta addirittura nel 1958 nell'articolo "*A business intelligence system"* di Hans Peter Luhn, ricercatore di IBM. Nell'articolo Luhn parla di "*automatic method to provide current awareness services to scientists and engineers*".

BI, esse sono quasi sempre influenzate dalla volontà dei produttori di evidenziare gli aspetti positivi dei propri software[28].

Cerchiamo di dare una definizione di BI, la più possibile completa, che sia indipendente dalla tecnologia e dagli strumenti applicativi usati per realizzare e che possa ricomprendere in essa i diversi punti di vista:

La Business Intelligence è un sistema[29] di modelli, metodi, processi, persone e strumenti che rendono possibile la raccolta regolare e organizzata del patrimonio dati generato da un'azienda. Inoltre, attraverso elaborazioni, analisi o aggregazioni, ne permettono la trasformazione in informazioni, la loro conservazione, reperibilità e presentazione in una forma semplice, flessibile ed efficace, tale da costituire un supporto alle decisioni strategiche, tattiche e operative[30].

Figura 4 Fonte: A. Rezzani, *Business intelligence: processi, metodi, utilizzo in azienda*, Apogeo, 2012



Le attività di raccolta, integrazione, pulizia e validazione dei dati, oltre che le procedure di aggregazione e parte dei calcoli analitici, sono svolte in maniera automatica; è infatti impensabile, processare manualmente la mole enorme di dati che, per alcune aziende descrivono l'attività giornaliera[31].

Verrà ora brevemente descritta l'architettura di un sistema di Business Intelligence, e a conclusione del capitola si delineerà il valore della BI.

1.5.1 Architettura di Business Intelligence

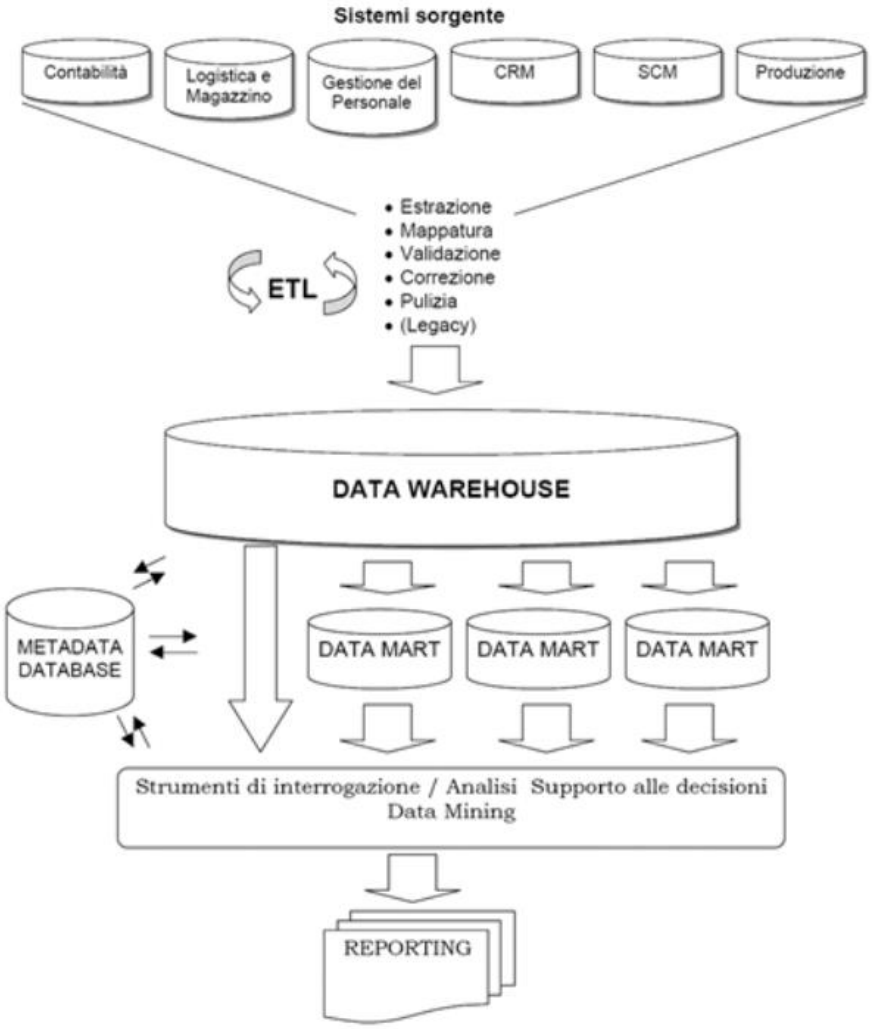
L'architettura di riferimento dei sistemi di BI è costituita da tre livelli:

[28] A. Rezzani, *Business intelligence: processi, metodi, utilizzo in azienda*, Apogeo, 2012, p. 7
[29] "Con sistema si intende un insieme di entità interconnesse. Le componenti di un sistema sono le entità di cui è costituito, le regole che lo governano e gli elementi strutturali, organizzativi e funzionali che legano le entità". Vercellis C., *Business Intelligence. Modelli matematici per le decisioni*, Mc Graw Hill, Milano, 2006, p.25
[30] A. Rezzani, *Business intelligence: processi, metodi, utilizzo in azienda*, Apogeo, 2012, p. 7
[31] Ivi

- Al primo livello vi sono i sistemi che producono e contengono dati elementari (sistemi alimentanti);
- Il livello intermedio è composto dai sistemi per l'integrazione e l'archiviazione dei dati semilavorati (sistemi di Data Warehouse);
- All'ultimo livello vi sono i sistemi per l'accesso ai dati e la produzione di informazioni "finite" (sistemi di Business Intelligence).

La figura sottostante rappresenta i tre livelli di un sistema di BI:

Figura 5: Architettura di Business Intelligence



1.5.1.1 Primo livello:

Nel primo livello, occorre raccogliere e integrare i dati presenti presso le diverse sorgenti primarie e secondarie, eterogenee per provenienza e tipologia. Esse si basano prevalentemente, sui dati dei sistemi operazionali (ERP, CRM, SCM, etc.), ma possono includere anche documenti non strutturati, quali le email e dati acquisiti

esternamente da agenzie e istituti economici. In generale fanno parte del primo livello tutte le informazioni provenienti sia dall'interno che dall'esterno dell'impresa[32].

1.5.1.1 Secondo livello:

I dati acquisiti nel primo livello, devono poi sottoporsi ad un processo di estrazione, pulizia e trasformazione prima di essere caricati nel Data Warehouse (DW), tale processo e denominato ETL (extract, transform, load). Le procedure di popolamento del data warehouse possono raggiungere elevati livelli di complessità, in relazione alle discrepanze esistenti tra le sorgenti, al loro livello di correttezza e al livello di precisione rappresentativa nel tempo che si desidera mantenere nel sistema informazionale. Le attività di estrazione, trasformazione e caricamento, ETL, vengono svolte in modo automatico da strumenti software specifici[33]:

- Nella fase di estrazione i dati vengono estratti dalle diverse fonti interne ed esterne disponibili. È possibile distinguere sotto il profilo logico tra una prima estrazione iniziale, nel corso della quale un data warehouse vuoto viene alimentato con i dati disponibili riferiti a tutti i periodi passati, e le successive estrazioni di natura incrementale rivolte all'aggiornamento del data warehouse mediante i dati che si rendono disponibili con il trascorrere del tempo.
- Nella fase di pulitura e trasformazione si cerca di migliorare la qualità dei dati estratti dalle diverse fonti, mediante la correzione di eventuali inconsistenze, inesattezze e carenze. Tra gli errori e le inconsistenze tipiche che rendono sporchi" i dati segnaliamo:
    - le inconsistenze tra valori presenti in campi diversi aventi il medesimo significato;
    - la duplicazione di dati;
    - la mancanza di dati;
    - la presenza di valori inammissibili.

  Nella fase di pulitura vengono applicate regole automatiche predefinite per la correzione di eventuali errori. La fase di trasformazione, invece, prevede

[32] Vercellis C., *Business Intelligence. Modelli matematici per le decisioni*, Mc Graw Hill, Milano, 2006, p.14
[33] Vercellis C., *Business Intelligence. Modelli matematici per le decisioni*, Mc Graw Hill, Milano, 2006, p.54 - 55

ulteriori conversioni dei dati che ne garantiscono l'omogeneità, rispetto all'integrazione delle diverse fonti. Vengono anche eseguiti calcoli di aggregazione dei dati, al fine di ottenere le sintesi richieste per svolgere le analisi cui il data warehouse è rivolto.

- Nell'ultima fase avviene il vero e proprio caricamento dei dati sul data warehouse. Questa procedura può avvenire in due modi:

    - *Refresh:* i dati vengono completamente riscritti all'interno del DW. Questa tecnica di caricamento è utilizzata nella fase di inizializzazione del DW, quando si procede all'estrazione di tutti i dati storici.
    - *Update:* vengono aggiunti al DW solo i nuovi dati e gli aggiornamenti dei dati storici. Questa tecnica è solitamente utilizzata in relazione all'estrazione incrementale, al fine di ottenere un aggiornamento periodico del DW.

A seguito delle procedure ETL, i dati entrano nel data warehouse a tutti gli effetti. Il DW corrisponde al livello intermedio dell'architettura di un sistema di BI, in cui i dati vengono integrati e archiviati. Ma che cos'è un data warehouse?
William H. Inmon, colui che per primo ha parlato esplicitamente di data warehouse, lo definisce come "*una raccolta di dati subject oriented, integrato, time variant, non volatile costruito per supportare il processo decisionale*[34]". Pertanto secondo Inmon la raccolta di dati deve essere[35]:

- Integrata: in un DW confluiscono dati provenienti da più sistemi e fonti esterne, è pertanto fondamentale l'integrazione dei dati raccolti. Si può raggiungere l'obiettivo di integrazione in più modi: utilizzando metodi di codifica uniformi, mediante l'utilizzo delle stesse unità di misura o attraverso l'omogeneità semantica di tutte le variabili;
- Orientata al soggetto: il Data Warehouse è orientato a temi aziendali specifici piuttosto che alle applicazioni o alle funzioni. In un Data Warehouse i dati vengono archiviati in modo da essere facilmente letti o elaborati dagli utenti. L'obiettivo, quindi, non è più quello di minimizzare la

[34] W. H. Inmon, *Building the Data Warehouse,* Wiley, 2005
[35] Ivi

ridondanza mediante la normalizzazione, ma quello di fornire dati organizzati in modo tale da favorire la produzione di informazioni. Si passa dalla progettazione per funzioni ad una modellazione dei dati che consenta una visione multidimensionale degli stessi;

- Variabile nel tempo: i dati archiviati all'interno di un Data Warehouse coprono un orizzonte temporale molto più esteso rispetto a quelli archiviati in un sistema operativo. Nel Data Warehouse sono contenute una serie di informazioni relative alle aree di interesse che colgono la situazione relativa ad un determinato fenomeno in un determinato intervallo temporale piuttosto esteso. Ciò comporta che i dati contenuti in un Data Warehouse siano aggiornati fino ad una certa data che, nella maggior parte dei casi, è antecedente a quella in cui l'utente interroga il sistema. Ciò differisce da quanto si verifica in un sistema transazionale[36], nel quale i dati corrispondono sempre ad una situazione aggiornata, solitamente incapace di fornire un quadro storico del fenomeno analizzato;
- Non volatile: tale caratteristica indica la non modificabilità dei dati contenuti nel Data Warehouse che consente accessi in sola lettura. Ciò comporta una semplicità di progettazione del database rispetto a quella di un'applicazione transazionale. In tale contesto non si considerano le possibili anomalie dovute agli aggiornamenti, né tanto meno si ricorre a strumenti complessi per gestire l'integrità referenziale o per bloccare record a cui possono accedere altri utenti in fase di aggiornamento.

Un DW, pertanto, non è altro che una collezione di dati in qualche modo correlati. Un DW può contenere un notevole volume di dati, sia storici che attuali. Di conseguenza bisogna affidarsi ad una tecnologia che riesca a gestire un grande volume di dati, e allo stesso tempo possa offrire flessibilità nell'accesso e estrazione

[36] I sistemi transazionali non sono altro che i sistemi operazionali (la sorgente dei dati di un DW) che forniscono supporto alle attività di tipo operativo e gestionale. I sistemi On Line Transaction Processing (OLTP), o transazionali, registrano, modificano e mostrano dei record in tempo reale in base alle esigenze operative delle varie Business Unit aziendali. Nei sistemi OLTP i dati sono distribuiti su più tabelle per ragioni di velocità e semplicità di inserimento e aggiornamento. L'informazione viene inserita e modificata una sola volta, non ci sono ridondanze. Anche se le query sono complesse non ci sono particolari problemi, perché vengono modificate raramente e solo da programmatori esperti. Inoltre le operazioni di inserimento, aggiornamento e ricerca sono estremamente rapide. Fonte sito: www. databasemaster.it

di tali dati. Vi sono due tipi di basi di dati che sono in grado di contenere e gestire i dati in DW:

- **Basi di dati relazionali:** sono caratterizzate per il fatto che i dati elementari e le relazioni che esistono tra di essi, vengono registrati in forma tabellare. La base dati relazionale è utilizzata sia nei sistemi transazionali che per un DW, tuttavia i dati sono ottimizzati in modo diverso a causa dei requisiti diversi che caratterizzano i due tipi di sistemi. I sistemi transazionali sono ideati per gestire le operazioni giornaliere, di conseguenza la base dati deve essere ottimizzata per permettere un aggiornamento efficace di singoli records. I dati vengono pertanto *normalizzati*, ossia i dati di una tabella vengono smistati su tabelle più piccole per evitare una ridondanza che sarebbe pericolosa per l'integrità dei dati stessi e per la rapidità delle operazioni di aggiornamento. Un DW è invece costruito per avere un accesso istantaneo ai dati e garantire flessibilità di interrogazione e analisi. I dati di un DW devono pertanto essere de normalizzati cioè memorizzati in un minor numero di grandi tabelle, questo per migliorare le prestazioni degli strumenti di interrogazione.
- **Basi di dati multidimensionali:** ideate per andare incontro alle esigenze dei managers, direttori e analisti che desiderano vedere i dati in maniera particolare, fare molte interrogazioni specialistiche, ed analizzare i risultati attraverso tecniche speciali. Nella base dati multidimensionale, i dati vengono rappresentati come fossero dimensioni e non tabelle, in questo modo è più facile da usare e più veloce di una base di dati relazionale. Questo tipo di base dati permette di avere una visione dei dati in più dimensioni, supporta la registrazione dei dati in vari livelli di aggregazione, in quanto le dimensioni sono strutturate gerarchicamente, inoltre permette di effettuare il *drill-down* e il *roll up*[37] dei dati. La base di dati multi-dimensionale è utilizzata maggiormente quando esistono molte

[37] Funzioni che permettono di navigare nei dati fino ad un singolo elemento di una dimensione e di visualizzarlo sia a livello di dettaglio che di aggregazione: Drill-down: dal livello di aggregazione ai dettagli che lo compongono, Roll-up: il procedimento inverso. Vercellis C., *Business Intelligence. Modelli matematici per le decisioni*, Mc Graw Hill, Milano, 2006, p.61 - 62

interrelazioni tra le dimensioni della base di dati, in caso contrario è più efficiente utilizzare una base dati relazionale.

Come abbiamo visto i DW sono generalmente basati sul modello detto multidimensionale, o Star Schema, ottimizzato per rispondere velocemente a interrogazioni di vario tipo. Un DW non ha una struttura predefinita, si determina in base ai componenti che si decide di utilizzare. In base alle necessità di un'azienda, organizzazione, viene disegnato lo schema del DW. Tali necessità possono essere determinate in relazione alla tipologia di analisi da effettuare o dal tipo di sorgente dei dati (base dati) che si deve inserire all'interno del DW. In linea di massima, individuiamo due componenti principale di un DW: l'area di Stage e i Data Mart. L'area di staging di un DW è l'area in cui vengono memorizzati i dati direttamente dalle fonti dati, senza subire grosse trasformazioni o pulizie. Il caricamento dei dati dalla sorgente all'area di staging avviene attraverso le procedure ETL, viste precedentemente. Attraverso l'area di staging si desiderano immagazzinare tutti i dati provenienti dalle varie fonti, aggregandoli e cercando di renderli il più possibile omogenei tra di loro. È chiaro che quest'area non sarà accessibile agli utenti di business ma, solo dal personale IT. I dati all'interno dell'area di staging, (così come in tutto il DW) come abbiamo già visto, sono organizzati secondo la rappresentazione multidimensionale, che si basa sullo schema a stella, in cui i dati vanno a popolare differenti tabelle che si possono ordinare in due gruppi[38]:

- Le **tabelle delle dimensioni,** che contengono appunto le dimensioni di analisi (le entità del modello relazionale), che rappresentano le chiavi di lettura ritenute significative per l'analisi dei fatti (misure). Le dimensioni tipiche di un DW sono rappresentate dai clienti, prodotti, punti vendita, località e tempo. Ciascuna tabella delle dimensioni risulta spesso strutturata al proprio interno secondo relazioni di tipo gerarchico, tra gli attributi che la contraddistinguono (ad esempio, la dimensione temporale può presentare generalmente due gerarchie principali: [giorni, settimane, anni] e [giorni, mesi, trimestre, anni]). Le dimensioni prefigurano le modalità secondo cui

[38] Vercellis C., *Business Intelligence. Modelli matematici per le decisioni*, Mc Graw Hill, Milano, 2006, p.55 - 58

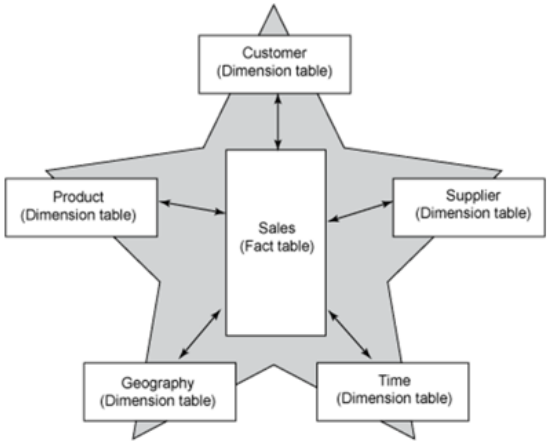
si potranno in seguito sviluppare le analisi di tipo OLAP (strumento attraverso cui organizzare l'analisi e la rappresentazione dei dati).

- Le **tabelle dei fatti**, si riferiscono, generalmente, a transazioni e contengono due tipi di dati:
    - i collegamenti alle tabelle delle dimensioni, che servono a collocare i riferimenti appropriati per le informazioni contenute in ciascuna tabella dei fatti;
    - i valori numerici degli attributi che caratterizzano le corrispondenti transazioni e che costituiscono l'effettivo oggetto delle successive analisi di tipo OLAP.

  Ad esempio, una tabella dei fatti può contenere le transazioni di vendita e fare riferimento a diverse tabelle delle dimensioni: clienti, punti vendita, prodotti, fornitori, tempo.

  In definita lo schema generale delle tabelle, nella sua distinzione tra fatti e dimensioni, rispecchierà il così detto diagramma a stella così raffigurato:

Figura 6: Star Schema



L'area di staging pertanto non è altro che l'insieme di tutte le tabelle dei fatti e delle dimensioni, le quali possono essere popolate da differenti sorgenti di dati. Di conseguenza è molto importante rendere i dati il più possibile omogenei tra di loro, e questo come già abbiamo visto avviene attraverso le procedure di ETL. Normalmente, il database di Stage viene aggiornato in base a una pianificazione fissa con dati grezzi dai sistemi di origine. Potrebbe sembrare un'attività ridondante spostare semplicemente i dati da un database a un altro database. Tuttavia, bisogna considerare che spesso i dati possono provenire da differenti fonti, in formati diversi

e da server fisici diversi. Di seguito vediamo alcuni dei motivi per i quali è importante avere un database di Stage:

- i dati potrebbero essere rimossi dai sistemi di origine il più rapidamente possibile, per ridurre il carico non necessario su questi sistemi;
- Il database di stage può risiedere su un server altamente ottimizzato per il crunch (spiegare cosa si intende) dei dati.
- L'ambiente di stage è usato come ambiente per modificare e aggregare i dati, successivamente i dati devono essere spostati, ed elaborati nel database multidimensionale finale (quello direttamente legato ai sistemi di reportistica). Questo è un modo per ottimizzare il server del DW per le richieste dati dai sistemi di business Intelligence, e per contenere la dimensione del DB multidimensionale preposto per l'interrogazione dei dati.

Una volta che i dati sono stati inseriti nell'area di staging, possono essere direttamente convogliati ai sistemi di Business Intelligence, o passare prima per un livello intermedio, ossia attraverso i Data Mart. I Data Mart, pur rientrando nell'architettura di un sistema di BI, non fanno parte effettivamente del DW (che si compone principalmente dell'ambiente di staging). Essi si posizionano tra il DW, che li alimenta, e i sistemi di Business Intelligence e permettono di ottenere prestazioni migliori essendo dimensionalmente inferiori al DW e particolarmente personalizzati rispetto alle analisi di uno specifico gruppo di utenti. Il Data Mart (DM) non è altro che un sottoinsieme o un'aggregazione dei dati presenti nel DW primario, contenente l'insieme delle informazioni rilevanti per una particolare area del business, una particolare divisione dell'azienda o una particolare categoria di soggetti. Il DM contiene solo le tabelle finali (dimensioni e fatti) che costituiscono il modello dimensionale (Star Schema) che verrà utilizzato dagli strumenti di analisi e/o costituirà il modello per l'implementazione del DB multidimensionale. Possiamo quindi considerare un DM come un DW funzionale o dipartimentale, di dimensioni più contenute e di natura più specifica al DW aziendale. I DW e i DM risultano, pertanto riconducibili a una medesima matrice tecnologica. Allo scopo di realizzare applicazioni di business intelligence, alcune aziende preferiscono progettare e realizzare in modo incrementale una serie di Data Mart tra loro integrati

piuttosto che un DW centrale, in modo da ridurre i tempi e le incertezze legate al progetto[39].

1.5.1.1 Terzo livello:

L'ultimo livello dell'architettura è costituito dai sistemi di BI che consentono la ricerca intelligente di dati e la produzione e analisi in "tempo reale" di informazioni per il supporto e attività di controllo e di decisione dei manager di qualunque livello aziendale. Attraverso questi strumenti gli utenti possono "crearsi" le informazioni di cui hanno bisogno nel rispetto dei propri tempi decisionali e quindi in tempo utile per osservare un fenomeno aziendale o per analizzare un problema emerso. I sistemi di analisi e previsione sono:

- Reporting;
- On-Line-Analytical Processing (OLAP);
- Simulazioni e previsioni;
- Data Mining.

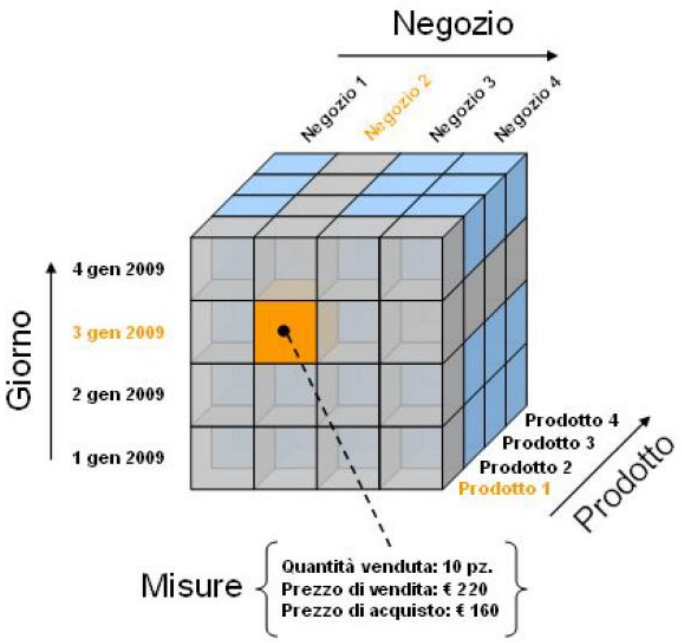
Questi strumenti forniscono un supporto all'ottenimento di dati da parte dell'utente e alla loro analisi. Tale supporto consiste nel fornire accesso diretto ai dati nel DW, visualizzare i dati attraverso viste multidimensionali e permettere interrogazioni ad hoc o predefinite.

Molto diffuso è il sistema denominato OLAP, che data la struttura multidimensionale dei dati, fornisce come strumento di sintesi per velocizzare l'analisi il Cubo Multidimensionale. Infatti una tabella dei fatti collegata a n tabelle delle dimensioni può essere rappresentata mediante un *Cubo di dati n-dimensionale* nel quale ciascuno degli assi corrisponde a una delle dimensioni. I cubi multidimensionali costituiscono una naturale estensione dei fogli elettronici, che possiamo vedere come cubi bidimensionali[40]. Per esempio, consideriamo la tabella dei fatti costituita dalle vendite collegata alle dimensioni prodotti, negozi, date, si otterrà il seguente cubo tridimensionale:

[39] Vercellis C., *Business Intelligence. Modelli matematici per le decisioni*, Mc Graw Hill, Milano, 2006, p.50 - 51

[40] Ivi, p. 57 - 58

Figura 7: Esempio di cubo tridimensionale



Ovviamente se le dimensioni di analisi sono più di tre, si tratta più propriamente di un ipercubo. Come abbiamo visto precedentemente, ad ogni dimensioni è associata una gerarchia di livello di aggregazione, che ne raggruppa i valori in diversi modi. I vari livelli che appaiono nella gerarchia sono chiamati attributi dimensionali: per esempio l'attributo data potrebbe avere la seguente gerarchia [Anno, Quadrimestre, Mese, Settimana, Giorno]. Le gerarchie delle dimensioni vengono utilizzare per svolgere diverse operazioni di visualizzazione dei dati nei cubi multidimensionali[41]:

- Roll-up: operazione che corrisponde a un'aggregazione nel cubo di dati, può essere ottenuta alternativamente in due modi. Nel primo caso, mediante il passaggio a un livello superiore lungo una gerarchia definita su una dimensione, ad esempio, nel caso della dimensione Località posso passare dal livello Comune al livello Provincia, e consolidare le misure di interesse mediante una somma condizionata di tipo *group-by* su tutti i record per i quali il comune appartiene a una stessa provincia. Nel secondo caso, mediante la riduzione di una dimensione, ad esempio, la rimozione della dimensione Tempo, comporta il consolidare le misure tramite la somma su tutti i periodi temporali presenti nel cubo di dati.

[41] Vercellis C., *Business Intelligence. Modelli matematici per le decisioni*, Mc Graw Hill, Milano, 2006, p.60 -61

- Drill-Down: operazione inversa del roll-up, consente di navigare attraverso un cubo di dati da informazioni aggregate e consolidate a informazioni con un maggiore grado di dettaglio. Si può quindi effettuare il drill-down in due modi, nel primo caso mediante il passaggio da un livello inferiore lungo una gerarchia definita su una dimensione, ad esempio data la dimensione Località si può passare dal livello Provincia al livello Comune, e disaggregare le misure di interesse lungo tutti i record per i quali il comune appartiene a una stessa provincia. Nel secondo caso, invece, si ottiene il drill-down mediante l'aggiunta di una dimensione, ad esempio l'introduzione della dimensione Tempo conduce a disaggregare le misure su tutti i periodi temporali presenti nel cubo di dati.

Le operazioni drill-down e roll-up sono entrambe operazioni di aggregazione il roll-up permette di ridurre la dimensionalità, il drill-down, essendo il suo duale, permette di aumentarla. Vi sono anche operazioni che permetto di restringere la grandezza del cubo, imponendo dei vincoli sugli attributi delle dimensioni, vi sono principalmente due operatori che permettono la restrizione di un cubo[42]:

- Slice: attraverso questo operatore il cubo viene "tagliato a fettine", ossia, la profondità delle sue dimensioni viene ridotta fissando un valore per una o più delle dimensioni originarie. Per esempio possiamo ridurre la dimensione tempo imponendo di voler vedere le misure con riferimento all'anno 2017, oppure ancora si potrebbe osservare le sole vendite del negozio x relativamente al prodotto y, in tal caso ad essere ridotte saranno le dimensioni negozio e prodotto;
- Dice: tramite questo operatore invece il cubo viene "tagliato a cubetti", ossia, l'insieme dei dati viene ridotto attraverso la formulazione di un criterio di selezione complesso. In questo caso il livello delle dimensioni rimane invariato. Un esempio di dice può essere voler visualizzare le sole vendite tra il 2010 e il 2016 per i soli prodotti che presentano un costo superiore ai 100 euro. Tale vincolo, va a delimitare enormemente il numero di misure, ma non il livello delle dimensioni.

[42] Vercellis C., *Business Intelligence. Modelli matematici per le decisioni*, Mc Graw Hill, Milano, 2006, p.60 -61

In conclusione un sistema OLAP permette di analizzare una grande quantità di dati, di visualizzare tali dati da prospettive differenti e di conseguenza offre un valido supporto ai processi decisionali.

1.5.2 Il valore della Business Intelligence

Nel paragrafo precedente, abbiamo visto come si struttura un sistema di Business Intelligence, a questo punto sorge naturale porsi la domanda, ma tale sistema aiuta effettivamente l'azienda e i suoi manager nel prendere le decisioni giuste?
Esiste sicuramente una correlazione tra l'uso efficace della BI e l'incremento delle performance aziendali, tuttavia, non basta garantire l'accesso ai dati per migliorare i margini. La BI evita alle persone che fanno parte di ogni livello aziendale di preoccuparsi del reperimento dei dati e della loro elaborazione e conservazione; consente loro di accedere ai dati, analizzarli, interagire con essi al fine di ricavare le informazioni per la gestione dell'intera azienda o parte di essa migliorandone l'efficienza (e di conseguenza le performance). È ovvio che senza persone, che interpretino in maniera corretta le informazioni, prendendo le giuste decisioni, la BI perde ogni sua valenza positiva. La creazione di un sistema di BI ed il suo utilizzo dipendono più dalla creatività, dalla cultura aziendale dell'informazione che dalla tecnologia; quest'ultima rende possibile la BI, ma, a volte, il fatto di porre al primo posto la tecnologia può mettere a rischio l'efficacia del sistema o può addirittura decretarne il fallimento. Spesso le aziende hanno la tendenza di farsi guidare dalla tecnologia nel disegno del sistema di BI. Niente di più sbagliato: sono le logiche del business che devono guidare l'ufficio IT nell'implementare il sistema di BI. La tecnologia deve essere al servizio della BI, non deve rappresentare un vincolo e non può imporre scelte di disegno del sistema[43]. Nel terzo capitolo parleremo ampiamente di quale metodologia seguire nell'implementazione di un sistema di BI. È tuttavia evidente che un sistema di Business Intelligence nel mondo odierno, è fondamentale per un'azienda (anche di piccole dimensioni) per poter essere competitiva sul mercato. Specie oggi, in cui si è assisto ad un aumento considerevole del volume di dati a cui un'azienda può accedere, non per nulla siamo

[43] A. Rezzani, *Business intelligence: processi, metodi, utilizzo in azienda*, Apogeo, 2012, p. 7

entrati nell’era dei Big Data che, possono essere una grande opportunità, ma anche un grande pericolo.

2 Big Data

2.1 Dati: un patrimonio importante

In azienda possiamo considerare i dati al pari di qualsiasi altro asset. I dati e le tecnologie di analisi, soprattutto nell'ultimo decennio, hanno assunto il carattere di risorsa indispensabile per la gestione aziendale. In questo paragrafo verrà presentata una panoramica utile a comprendere il patrimonio dei dati, considerandoli sotto diversi punti di vista: le fonti e il tipo di struttura[44].

Se guardiamo la provenienza dei dati possiamo operare una prima generale distinzione tra fonti interne e fonti esterne all'azienda. I **dati interni** ad una azienda provengo principalmente dalle fonti operazionali, cioè quelle che fanno riferimento all'attività operativa giornaliera dell'azienda e che, per questo motivo, variano a seconda della tipologia di business e settore economico. Per un'azienda industriale, alcuni esempi di fonti operazionali possono essere[45]:

- Applicativi di gestione della produzione: che registrano le quantità di materie prime utilizzate, i servizi e i beni accessori consumati;
- Applicativi di gestione degli acquisti: che si occupano di registrare ogni ordine di acquisto e i movimenti di magazzino;
- Applicativi di gestione dell'ordine e delle consegne: sono sistemi per la registrazione di ordini ricevuti dai clienti, attività di consegna e movimentazione del magazzino prodotti;
- Applicativi di contabilità: registrano i movimenti che scaturiscono dall'emissione di fatture di vendita o dal ricevimento di fatture d'acquisto, oltre a ogni altra movimentazione di cassa o banca;

[44] A. Rezzani, *Big Data Analytics*, Maggioli Editore 2017, p.3
[45] Ivi, p.4

- Applicativi di gestione del personale: tale sistema si occupa della gestione degli stipendi, in cui vi sono anche la gestione degli obiettivi, premi, malattie, infortuni;
- Applicativi di gestione del cliente: consentono la gestione completa del cliente, dalla registrazione dei dati anagrafici, fino alla gestione delle campagne di marketing.

Se consideriamo invece un'azienda bancaria, le fonti operazionali sono in parte simili a quelle dell'esempio precedente, ma a esse si aggiungono applicativi tipici dell'attività bancaria, quali: la gestione dello sportello, le applicazioni di gestione e valutazione degli strumenti finanziari, gli applicativi per la valutazione del rischio e l'erogazione di finanziamenti. Le aziende commerciali, come quelle della grande distribuzione, presentano oltre ai sistemi di contabilità, di gestione acquisti, magazzino e gestione del personale, anche sistemi di rilevazione delle vendite alle casse ed emissione degli scontrini fiscali oppure gli applicativi per la gestione delle promozioni e delle tessere clienti.

Per alcune aziende i dati operazionali possono assumere volumi rilevanti. Specie oggi che attraversi i sistemi computerizzati DCS (Distribution Control System) vengono generati enormi quantità di dati. Tali sistemi vengono utilizzati principalmente per il controllo di impianti industriali (ma anche nel settore bancario e commerciale, vi sono sistemi simili che monitorano l'entrata e uscita dei clienti da banche e centri commerciali). I componenti del sistema, connessi tramite una rete che consente il controllo, la comunicazione e il monitoraggio, generano dati relativi allo stato dell'impianto mediante sensori legati al componente stesso. Le rilevazioni dei dati possono avvenire ad intervalli temporali molto piccoli e ciò, assieme alla presenza anche di migliaia di sensori, porta a produrre una mole elevata di valori[46].

Come abbiamo visto nel capitolo precedente, i dati provenienti dalle fonti operazionali, possono arrivare da differenti applicativi, di conseguenza l'uniformità e la coerenza dei dati non sono per nulla garantite, poiché ci si trova in una situazione in cui dati fondamentali sono replicati e manipolati in ciascun software, presentando differenze di formati, di completezza o di aggiornamento dei dati. Di conseguenza, come visto in precedenza, al fine di garantire l'uniformità e la

[46] A. Rezzani, *Big Data Analytics*, Maggioli Editore 2017, p.5

coerenza dei dati, le aziende ricorrono all'utilizzo dei datawarehouse e di tutti gli strumenti di business intelligence, per filtrare, pulire e aggregare i dati, oltre che per creare report di sintesi da sottoporre ai manager.

Per molte aziende, tuttavia le fonti interne non bastano a coprire il fabbisogno informativo, di conseguenza non è raro che si renda necessario l'utilizzo di dati provenienti dall'esterno in aggiunta ai dati interni. In taluni casi, per **dati esterni** potrebbe trattarsi semplicemente di integrazioni alle anagrafiche (per esempio l'anagrafica ISTAT), mentre in altri casi i dati esterni sono il vero e proprio oggetto dell'analisi (analisi del sentiment sui dati provenienti dai principali Social Network). Il reperimento e l'utilizzo di dati esterni pongono alcuni problemi. Uno di essi consiste nella loro qualità, che potrebbe presentare difetti di accuratezza, completezza e coerenza. La qualità dei dati è determinata da un insieme di caratteristiche[47]:

- Completezza: presenza di tutti i dati necessari per descrivere un'entità, una transazione o un evento.
- Consistenza: assenza di contraddizione nei dati. La consistenza è legata al concetto di equivalenza dei dati. Per esempio, in una banca, la somma del saldo di fine mese precedente di un conto corrente con i movimenti attivi e passivi, deve essere uguale al saldo di fine mese corrente; ciò significa che la rilevazione dei saldi e dei movimenti devono essere tra loro consistenti. La consistenza è anche dettata dalle regole di business, che differiscono da azienda da settore a settore.
- Accuratezza: riguarda la conformità ai valori reali, cioè la correttezza dei valori stessi. Il punto di partenza per l'accuratezza dei dati risiede nella loro integrità.
- Assenza di duplicazione: campi, record o tabelle devono essere presenti soltanto una volta, evitando duplicazioni nello stesso sistema oppure in sistemi diversi. Oltre alla necessità di una doppia manutenzione, la duplicazione dei dati incide negativamente sulla qualità, poiché è possibile una mancata sincronizzazione tra le copie dei dati.

[47] A. Rezzani, *Big Data Analytics*, Maggioli Editore 2017, p.6

- Integrità: si utilizza il termine integrità con riferimento ai database relazionali. Essi infatti garantiscono, attraverso strumenti di tipo di dati, check constraint, le chiavi primarie e le chiavi esterne, che i dati rispettino alcuni vincoli: per esempio che in una colonna vi siano soltanto dati dello stesso tipo; oppure che non esistano due righe uguali all'interno di una tabella; oppure, ancora, che in caso di relazioni tra due tabelle, una colonna, su cui è definita tale relazione, dovrà contenere solo valori appartenenti ad una colonna di un'altra tabella.

Occorre precisare che la problematica relativa alla qualità riguarda anche i dati interni; tuttavia, sui dati esterni l'azienda non ha alcuna possibilità di manovra e spesso è difficile anche la semplice verifica del livello qualitativo.
Come precedentemente anticipato un altro punto di vista rispetto al quale possiamo classificare i dati riguarda la loro struttura o meglio, la presenza o assenza di una struttura che consente di identificare agevolmente ogni attributo. Distinguiamo quindi le seguenti tipologie[48]:

- **Dati Strutturati**: dati che sono rappresentabili in formato tabellare all'interno di un database relazionale, oppure tramite un formato quali l'XML o il JSON, che assieme ai dati contengono i metadati che definiscono i nomi dei campi e la loro struttura. "Rappresentabili in formato tabellare" significa che i dati non si devono per forza trovare in un database, ma potrebbero trovarsi, per esempio in un file di testo in formato CSV, nel quale i singoli campi sono ben separati e identificabili (quindi il file potrebbe essere importato senza problemi in una tabella).
- **Dati non Strutturati:** sono dati non strutturati i dati che non ha senso rappresentare in formato tabellare, anche se esse potrebbero essere inseriti in un database relazionale: non avendo una struttura, infatti, l'interno blocco di dati sarebbe contenuto in un unico campo all'interno di una tabella, rendendo tale struttura completamente inutile. Alcuni esempi di dati non strutturati sono: il contenuto testuale di un documento (email, pdf, tweet,

[48] A. Rezzani, *Big Data Analytics*, Maggioli Editore 2017, p.8

post di un blog), oppure i byte di un'immagine, di un video, di un file sonoro.

- **Dati semi-strutturati**: presentano una parte dotata di struttura e una parte non strutturata. Per esempio un documento Word, o PDF, possiede una serie di metadati che sono molto ben strutturati (titolo, autore e molto altro), mentre il corpo del documento è costituito da testo. Lo stesso vale per le immagini che all'interno del file presentano una serie di metadati che descrivono lo scatto, le impostazioni della fotocamera, la data e ora e addirittura le coordinate GPS.

Bisogna sottolineare che per lavorare con i dati non strutturati occorre effettuare su di essi trasformazioni in grado di fornire un risultato che possegga una struttura.
È possibile effettuare un'ulteriore classificazione e distinguere tra i dati generati dalle persone e dati generati dalle macchine. Alla prima categoria corrispondono sia dati esterni, come quelli provenienti dai social network o dai siti di e-commerce, sia alcuni dati interni imputati manualmente (operazioni di data entry). I dati generati dalle macchine comprendono numerose casistiche: dai dati provenienti da strumenti di misurazione, ai valori prodotti dai sensori, dai dati di log di web server e router ai sistemi di calcolo che da dati grezzi producono automaticamente nuovi valori e, ancora, dai device di lettura di codici a barre ai sistemi analoghi utilizzati dalle aziende di logistica. Il dato generato automaticamente è di solito meno soggetto a problemi di qualità, anche se malfunzionamenti del dispositivo che genera i dati possono creare numerose anomalie[49].

Nel capitolo precedente abbiamo visto come il dato sia l'elemento embrionale per creare conoscenza, in questa prima parte del secondo capitolo abbiamo visto l'eterogeneità dei dati in base alla loro fonte, o struttura o provenienza. È chiaro che tutte queste considerazione valgo anche quando si parla di Big Data, tuttavia in tale ambito ogni elemento è ingigantito dato che siamo in presenza di un enorme massa di dati.

[49] A. Rezzani, *Big Data Analytics*, Maggioli Editore 2017, p.9

2.2 Introduzione ai Big Data

Il fenomeno dei Big Data, nato attorno al 2011, ha ormai superato la fase in cui si trattava di una *buzzword*, cioè una parola in voga, alla moda, che è stata utilizzata più per promuovere prodotti o servizi informatici che per risolvere problematiche legate all'estrazione di valore da una nuova categoria di dati. Oggi, infatti, la disponibilità di tecnologie sia open source sia proprietarie e la disponibilità di piattaforme cloud sono fattori che concorrono a rendere più semplice e meno costosa l'attivazione di processi volti a sfruttare le opportunità che i big data offrono[50]. Ma che cosa sono i Big Data? Cosa rende tali dati così importanti da spingere collosi come Facebook, Google e tante altre aziende leader del mercato a investire milioni di dollari per gestirli e analizzarli.

Le definizioni proposte riguardo questo concetto sono molteplici, ma generalmente con il termine Big Data si intende una grande mole di dati, dai quali risulta difficile estrapolare informazioni utili. I Big Data, nascono grazie allo sviluppo tecnologico che ha portato un incremento delle risorse di gestione intensiva di dati: ci sono 4,6 miliardi di cellulari in tutto il mondo e circa 2 miliardi di persone che possono accedere ad internet. La capacità effettiva mondiale di scambiare informazioni attraversi le reti di telecomunicazione è quasi raddoppiata ogni 40 mesi dal 1989**:**

- 281 PB (Petabyte) nel 1986;
- 471 PB nel 1993;
- 2,2 EB nel 2000;
- 65 EB nel 2007;
- 2,5 EB al giorno nel 2012;
- 667 EB annui nel 2014;

e si prevede un ulteriore aumento del traffico causato dal massiccio uso di internet che nel 2020 porterà alla cifra di 59 ZB zettabyte (1021 Byte)[51]. Nel 2010 Eric Schmidt sosteneva: "Nel giro di due giorni produciamo la stessa quantità di dati generati dagli albori della civiltà sino al 2003". Secondo IBM, entro il 2025 il mondo avrà generato una massa di dati dell'ordine dei 160 zettabyte[52]. Alla luce di

[50] A. Rezzani, *Big Data Analytics*, Maggioli Editore 2017, p.13
[51] Dati presi dal sito: Wikipedia
[52] Big data Interim report nell'ambito dell'indagine conoscitiva di cui alla delibera n. 217/17/CONS, p.8.

questi dati, e queste dichiarazioni, è evidente come il mondo odierno sia sommerso da un enorme quantità di Dati.

I Big Data sono generalmente definiti come quei dati che presentano una o più delle seguenti tre caratteristiche, chiamate le tre V[53]:

- **Volume:** Il volume si riferisce alla capacità di acquisire, memorizzare ed accedere grandi volumi di dati, basti pensare che il 90 % dei dati di tutto il mondo è stato generato negli ultimi due anni. I dati generati automaticamente da macchine (sensori, DCS – *Distributed Control System*, strumenti scientifici) e quelli relativi a transazioni bancarie e movimenti sui mercati finanziari possono assumere volumi imponenti, soprattutto se considerati al loro massimo livello di granularità. Si può parlare di Big Data solo in presenza di una mole di dati nell'ordine di Terabyte(10^{12}) in su (Petabyte(10^{15}), Exabyte(10^{18}), Zettabyte(10^{21}).
- **Velocità:** Il termine velocità in questo caso non fa riferimento alla crescita, piuttosto alla necessità di ridurre al minimo e di comprimere dunque i tempi di gestione e analisi: questo aspetto è fondamentale in quanto alcune tipologie di dati potrebbero già risultare obsoleti in tempi molto ridotti. Diviene quindi strategico presidiare e gestire il ciclo di vita dei Big Data. L'analisi dei dati dovrebbe essere fatta in *real time* o quasi, del tipo che "a volte due minuti è già troppo tardi". Ottenere un vantaggio competitivo significa identificare una tendenza o un'opportunità in minuti o addirittura qualche secondo prima della concorrenza. Pensiamo al mondo dell'Internet Of Things e dei sensori, che sono in grado di generare dati con una velocita elevatissima.
- **Varietà:** La varietà di dati rappresenta oggi un cambiamento fondamentale nel modo in cui i dati sono memorizzati e analizzati. Essa è riferita alle diverse tipologie di dati con cui si ha a che fare le quali divengono difficilmente gestibili con sistemi "tradizionali" e richiedono pertanto tecnologie specifiche: appunto, le tecnologie Big Data. Quando si parla di varietà pertanto si intende la diversità di formati, delle fonti e delle strutture. Inoltre, come già abbiamo visto, alcuni dati possono anche non avere una struttura.

[53] A. Rezzani, *Big Data Analytics*, Maggioli Editore 2017, p.14

Vi sono altri due punti importanti che definiscono i Big Data: la *convenienza economica* e i *limiti tecnologici* di alcune soluzioni di gestione dei dati. Secondo tale approccio è possibile definire i Big Data come[54]:

dati che non è possibile analizzare con tecnologie tradizionali oppure dati per i quali l'adozione di tecnologie tradizionali risulti molto onerosa dal punto di vista economico.

Quando parliamo di tecnologie tradizionali intendiamo essenzialmente i database relazionali (RDBMS, Relational Database Management System), per quanto riguarda le basi dati e gli strumenti di analisi descrittiva che li utilizzano come sorgenti, ma anche quei tool di analisi predittiva e data mining che mostrano i loro limiti al crescere dei volumi di dati o al venir meno di strutture tabellari. La definizione basata non solo sui limiti tecnologici, ma anche sulla convenienza economica ci fa capire che non è strettamente necessario possedere moli di dati dell'ordine dei petabyte per pensare all'adozione di sistemi big data (Hadoop e Spark, per esempio). È ovvio che la componente del volume gioca un ruolo importante sia nella diminuzione della convenienza economica delle tecnologie tradizionali, sia nel raggiungimento dei loro limiti tecnici. Anche la diversità (o l'assenza) di formati può rendere poco adatto un database tradizionale all'analisi di certi dati. Quanto alla velocità, essa produce criticità soprattutto nelle fasi di acquisizione (*data ingestion*) e di salvataggio nel database. Il termine Big Data identifica dunque sia i dati con le caratteristiche sopra descritte, sia le tecnologie con cui si possono risolvere i problemi (tecnici o economici) di data ingestion, di conservazione e di analisi dei dati[55]. A titolo di esempio, citiamo alcune categorie di dati che possono rientrare nella definizione di big data, indicando per ciascuna le caratteristiche importanti[56]:

[54] A. Rezzani, *Big Data Analytics*, Maggioli Editore 2017, p.14
[55] A. Rezzani, *Big Data Analytics*, Maggioli Editore 2017, p.15
[56] Ivi, p. 16

Tabella 1: Esempi di Big Data[57]

Caso	Caratteristica	Esempi di utilizzo
Sensori e DCS	Velocità e volume	Analisi dei guasti e manutenzione predittiva.
Radio Frenquency Identification (RFID)[58]	Velocità e volume	Analisi del percorso d'acquisto in un negozio della grande distribuzione. Analisi e tracking delle merci, in combinazione con altri dati (ambientali, geografici, ecc.)
Quotazione e transazioni su mercati finanziari	Velocità e volume	Sistemi automatici di high frequency trading; analisi previsionale.
Dati da strumenti scientifici	Velocità e volume	Riconoscimento di pattern. Simulazioni.
Dati astronomici	Velocità e volume	Analisi del quantitativo enorme di dati raccolti da osservatori e radiotelescopi.
Dati metereologici	Volume	Previsioni meteo. Monitoraggio di eventi atmosferici estremi.
Informazioni sanitarie	Volume e varietà (diversità di formati)	Ministero della salute, enti di ricerca: identificazione e monitoraggio della diffusione di malattie.
Dati fiscali, bancari e patrimoniali	Volume	Il Ministero delle Finanze, Agenzia delle Entrate e Guardia di Finanza possono utilizzare le enormi banche dati a loro disposizione per l'identificazione di comportamenti anomali che indicherebbero casi di evasione fiscale.
Social Network	Varietà: diversità di formati, dati semi-struttati	Sentiment Analysis (come si sta parlando della nostra azienda? Come è stato accolto il nuovo prodotto?) CRM (Customer Relationship Management) Utilizzo da parte di servizi di intelligence.
Blog, Forum	Varietà: Dati semi-strutturati	Sentiment Analysis Utilizzo da parte di servizi di Intelligence.

[57] A. Rezzani, *Big Data Analytics*, Maggioli Editore 2017, p.16

[58] RFID è una tecnologia per l'identificazione automatica di oggetti, animali o persone basata su dispositivi elettronici (detti tag) capaci di rispondere all'interrogazione da parte di lettori fissi o portatili a radiofrequenza. Un tag identifica in modo univoco un oggetto. La tecnologia RFID può trovare impiego in: automatizzazione degli inventari, bigliettazione elettronica, logistica e spedizioni, passaporti, antitaccheggio, rilevazione di parametri ambientali, controllo della temperatura degli alimenti durante le spedizioni e molto altro. A. Rezzani, *Big Data Analytics*, Maggioli Editore 2017, p.25

Web server log	Volume	Analisi del traffico sui web server, identificazione dei comportamenti di navigazione degli utenti
Log del traffico di un router	Volume e velocità	Utilizzo da parte di provider.
Dati provenienti dai sistemi di sorveglianza	Volume, Velocità e diversità di formati	Utilizzo da parte di polizia, enti di vigilanza, servizi di intelligence.
Documenti	Volumi e assenza di struttura	Fraud detection: per esempio, l'analisi di richieste di risarcimento effettuate alle assicurazioni possono essere analizzate e associate a casi fraudolenti ed esaminate con più attenzione.
Dati geografici	Volume e velocità	I dati provenienti da sistemi GIS possono essere utilizzati assieme ad altri dati per scopi diversi.

Alle categorie esemplificate dalla Tabella 1 aggiungiamo i dati dei sistemi operazionali che in certe aziende raggiungono volumi ragguardevoli.

2.3 Creare valore con i Big Data

Tutti questi dati raccolti costituiscono un valore per un'azienda. È dall'analisi dei dati che si colgono le opportunità e si trae supporto per i processi decisionali in modo tale che possano avere un grande impatto sulla nostra attività, e più dati si hanno a disposizione più informazioni e valore si riescono ad estrarre. Tuttavia il solo volume dei dati non ne garantisce la loro "qualità". La veridicità e la qualità dei dati diventano pertanto dei requisiti fondamentali affinché i dati possano davvero "alimentare" nuove intuizioni, idee e costituire valore. La creazione di valore attraverso i dati può essere raggiunta attraverso aree principali d'intervento:

- *La creazione della trasparenza*: si può ottenere la trasparenza semplicemente rendendo accessibile e disponibili i dati e le informazioni a tutti gli "stakeholder" in modo tempestivo e in modo da creare un valore straordinario. Nel settore pubblico è in atto da tempo una tendenza (Open Data), in particolare nel mondo anglosassone, di apertura del patrimonio

informativo a tutte le organizzazioni private e/o ai cittadini. Per quanto le prime iniziative abbiamo messo in luce elementi di criticità, si tratta di una tendenza che prosegue e che, anche in Italia, ha messo in moto alcune interessanti iniziative sia a livello regionale che nazionale[59].

- *Miglioramento delle prestazioni*: tutte le organizzazioni, sia pubbliche che private, con una maggiore analisi ed elaborazione dei dati e delle informazioni disponibili possono migliorare le loro prestazioni ed i loro processi. In modo significativo a questo riguardo sono le esperienze di analisi dei dati di vendita e, in generale, di relazione con la clientela. Una più attenta elaborazione dei dati relativi alle vendite può incrementare in modo significativo la capacità di segmentare l'offerta e di personalizzarla sulle specifiche esigenze del cliente. Analizzare i dati sul venduto può dare origine ad una più precisa previsione dell'andamento futuro delle vendite con evidenti vantaggi anche nella logistica. Le politiche commerciali di supporto saranno più mirate, gli sprechi saranno ridotte e vi sarà un reenginering dei processi.
- *Segmentazione dei clienti e personalizzazione delle azioni*: i Big Data consentono alle organizzazioni di creare segmentazione specifiche e di adattare i prodotti e i servizi alle esigenze dei consumatori. Questo approccio è ben noto in marketing e nella gestione del rischio.
- *Supporto alle decisioni*: i processi decisionali, in tutte le organizzazioni, possono diventare sempre più data-driven e fact-based. Le decisioni prese con il supporto di strumenti analitici possono diventare la regola non l'eccezione. Fino ad ora queste tecniche erano diffuse solo nelle grandi organizzazioni, anche per il costo elevato dei sistemi e degli strumenti informatici per l'elaborazione e l'analisi dei dati. Le attuali evoluzioni tecnologiche, come il diffondersi del "cloud-computing" e degli strumenti

[59] In Italia il sito *DatiOpen.it* è un esempio di tali iniziative, il cui scopo è quello di rendere l'enorme patrimonio di dati aperti (i dati aperti, comunemente chiamati open data, sono dati accessibili a tutti, tipicamente via web, senza restrizioni di copyright, brevetti o altre forme di controllo che ne limitino la riproduzione) accessibili facilmente al grande pubblico: cittadini, professionisti, enti pubblici e/o privati. Per raggiungere questo obiettivo DatiOpen.it prevede il seguente approccio:

- Raccogliere e documentare la maggiore quantità possibile di open data italiani in un unico portale.
- Permettere la visualizzazione e consultazione diretta delle informazioni contenute in tali dati per mezzo di tabelle, grafici e mappe.
- Consentire agli enti pubblici e privati di pubblicare gratuitamente nuovi open data.

di analisi "open-source", rendono molto meno costoso e più flessibile l'adozione di queste tecniche da parte anche delle organizzazioni medio-piccole.

- *Innovare i prodotti, i servizi e i modelli di business*: eseguire l'analisi dettagliata, su grandi basi dati, consente alle organizzazioni di innovare i propri Business Model per cogliere le sfide di un ambiente in continuo e turbinoso cambiamento.

Poter quantificare l'impatto economico derivante dall'utilizzo dei big data è un'informazione importante e serve alle aziende per decidere se realizzare oppure accantonare il progetto stesso. Per far questo si ricorre al classico indice di bilancio ROI[60] (Return on Investment) adattato al contesto di riferimento:

Figura 8: Valutazione economica dell'uso dei Big Data[61]

Fattori di costo
- Hardware
- Licenze software
- Costi del personale, training, consulenze
- Costi di gestione e manutenzione dei sistemi

Benefici
- migliori campagne di marketing
- miglior sviluppo dei prodotti
- attività di cross selling
- riduzione del tasso di abbandono della clientela
- gestione più efficiente degli impianti
- prevenzione dei guasti in impianti industriali

$$ROI_{BD} = \frac{minori\ costi\ +\ maggiori\ ricavi\ -\ Costi\ di\ ricorrenti}{Investimento\ Iniziale}$$

Dietro a tali opportunità, tuttavia, si celano sfide a cui necessario rispondere efficacemente, al fine di ottenere un effettivo vantaggio competitivo. Tali sfide possono essere così sintetizzate:

- *Sfide tecniche*: il tema della qualità, dell'accessibilità e della fruibilità dei dati rimane un aspetto critico che molte organizzazioni non hanno ancora adeguatamente affrontato. Come citato in una ricerca, IBM stima che oltre il 50% dei manager, non consideri affidabile il dataset sul quale si basano i processi decisionali.
- *Mancanza di talenti*: la crescita esponenziale dei Big Data ha messo in evidenza la necessità, sempre maggiore, di persone con competenza

[60] Indice di bilancio che indica la redditività e l'efficienza economica della gestione caratteristica a prescindere dalle fonti utilizzate: esprime, cioè, quanto rende il capitale investito in quell'azienda. Fonte, sito: *Wikipidea.org*.

[61] A. Rezzani, "Data Skills Understandig the World", dal sito: www.dataskills.it

adeguate per elaborare, analizzare e visualizzare grandi dataset ed estrarre da essi le informazioni essenziali per il supporto delle decisioni. Emergeranno nuove figure professionali come quella del "Data Scientist", dotate di competenze multidisciplinari (Statistica, Informatica, Economia…etc), necessarie per creare valore attraverso i Big Data.

- *Sfide Legali*: il tema della sicurezza, della tutela della privacy e della protezione della proprietà dei dati e delle informazioni, in molti settori, costituisce un elemento critico e talora, un freno all'innovazione.
- *Sfide Culturali*: la cultura delle decisioni data-driven e fact-based non è ancora sufficientemente diffusa nelle organizzazioni, sia pubbliche che private (specie in Italia). È indispensabile, dunque, dedicare adeguata attenzione alle politiche formative ed alla gestione del cambiamento.

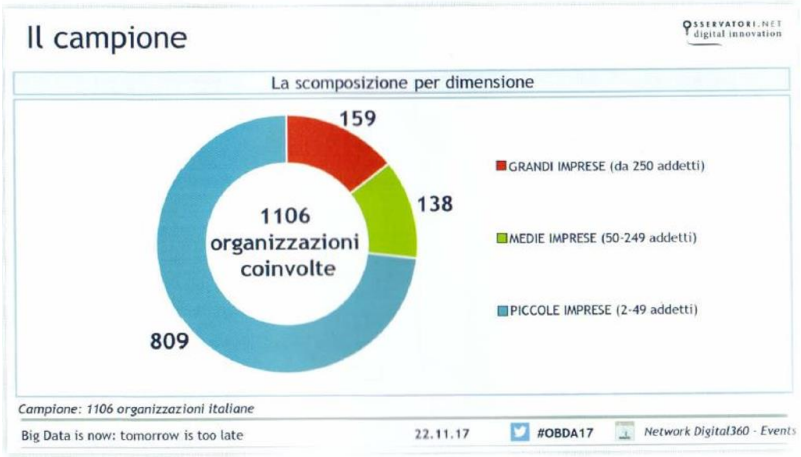
Questa grande complessità e quantità dei dati può essere gestita solo attraverso soluzioni tecnologiche nuove. È stata sviluppata e adattata un'ampia varietà di tecniche e tecnologie per aggregare, manipolare, analizzare e visualizzare i Big Data. L'insieme di tecniche e tecnologie attingono da settori diversi, quali la statistica, l'informatica, la matematica applicata e l'economia: questo significa che un'azienda, la quale intende trarre valore dai Big Data, dovrebbe adottare un approccio flessibile e multidisciplinare. Le tecniche e le tecnologie per aggregare, manipolare, analizzare e visualizzare i Big Data sono in continuo sviluppo per ottenere, sempre, un miglioramento delle prestazioni.

2.4 Osservatorio Big Data

Come abbiamo visto nei paragrafi precedenti lo sfruttamento dei Big Data da parte di aziende e organizzazioni può portare a vantaggi competitivi e nuove opportunità, se si ha il coraggio di affrontare le criticità insite in tale innovazione. In questo paragrafo mostriamo la crescita del mercato dei Big Data nel contesto italiano, come è composto tale mercato, in che modo sono stati utilizzati i Big Data nel mercato e quali sono gli ostacoli principali incontrati dalle aziende nell'usarli. Le osservazioni che andremo a fare si basano sui dati fornito dall'*Osservatorio Big*

Data Analytics & Business Intelligence[62],la cui ricerca è giunta alla quinta edizione, ed intende evidenziare il valore strategico che le metodologie di Big Data Analytics svolgono nelle grandi organizzazioni e nelle Piccole e Medie imprese, ponendo in luce i potenziali vantaggi relativi alla competitività, alla redditività, alla tempestività e all'aumento di efficacia nei processi decisionali. La Ricerca dell'Osservatorio si basa su un'analisi empirica che, attraverso survey e studi di caso, coinvolge ogni anno oltre 1000 attori, tra cui Responsabili dei Sistemi Informativi, Responsabili Digital e Innovation, Responsabili degli Analytics e Executive di linee di business, di organizzazioni e Pubbliche Amministrazioni, nonché i principali player del mondo dell'offerta in questo settore in Italia. Il campione preso in considerazione è così composto:

Figura 9: Composizione del Campione su cui è stata svolta l'indagine[63]



Come si può notare il campione è costituito al 73,12% dalle piccole imprese (ossia da imprese con al massimo 49 dipendenti), al 14,35% dalle grandi imprese e al 12,53% dalle medie imprese. Il campione delle grandi imprese è stato a sua volta scomposto per settore:

[62] Fa parte della School of Management del Politecnico di Milano. Costituita nel 2003, la School of Management accoglie le molteplici attività di ricerca, formazione e alta consulenza, nel campo dell'economia, del management e dell'industrial engineering, che il Politecnico porta avanti attraverso le sue diverse strutture interne e consortili. Vedi sito: https://www.osservatori.net/it_it/osservatori/chi-siamo.

[63] *Osservatorio Big Data Analytics & Business Intelligence*, https://www.osservatori.net/it_it/osservatori/chi-siamo

Figura 10: Scomposizione del campione per settori[64]

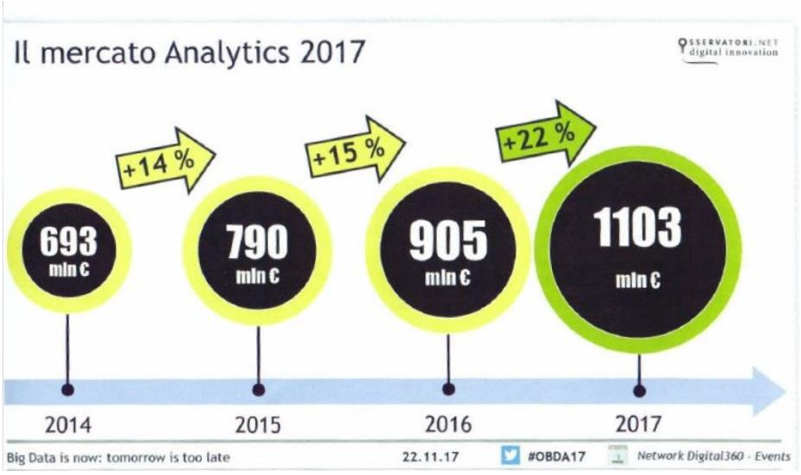
Data tale scomposizione per settore, si è valutata la quota di mercato degli Analytics (ossia quale settore ha fatto i maggiori investimenti in tecnologie legate allo sfruttamento dei Big Data).

Figura 11: Mercato degli Analytics per settore[65]

Come si può vedere i tre settori che hanno investito maggiormente nel mercato degli Analytics sono: le Banche, il settore manifatturiero e, con una percentuale considerevolmente minore, Telco e Media

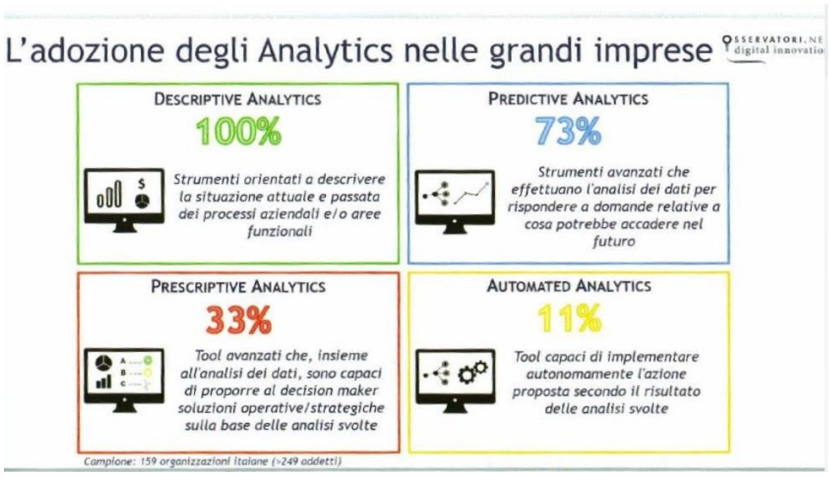
[64] *Osservatorio Big Data Analytics & Business Intelligence*, https://www.osservatori.net/it_it/osservatori/chi-siamo
[65] Ivi.

Figura 12: Trend del volume di fatturato nel mercato degli Analytics[66]



In questo grafico invece viene mostrata la crescita del valore del mercato degli Analytics a partire dal 2014 sino al 2017. Come si vede oggi il mercato italiano degli Analytics vale circa 1103 milioni di euro, sintomo che anche in Italia, molte aziende hanno compreso le opportunità che si celano nello sfruttamento dei Big Data.

Figura 13: Principali aree di adozione degli Analytics[67]

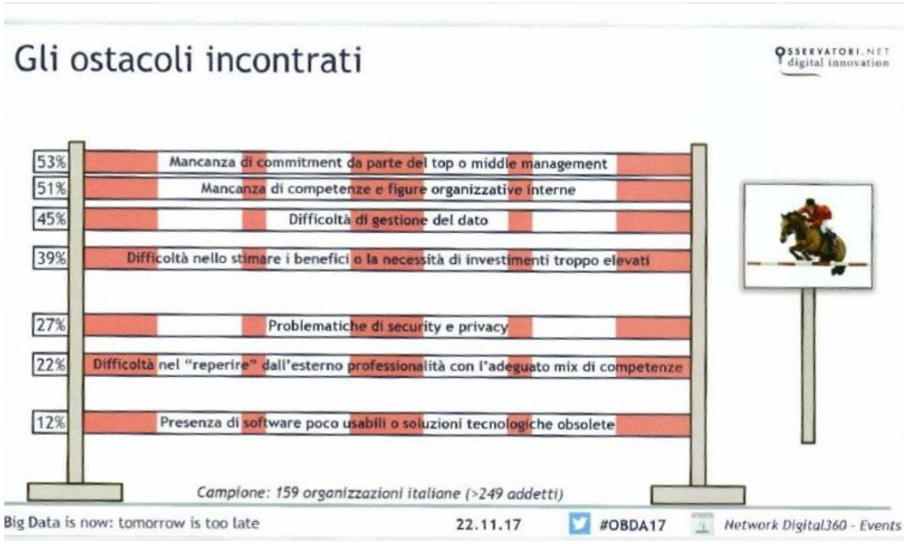


In questa immagine viene mostrato verso quale tipo di analisi le grandi imprese adottano i Big Data. È evidente che adottare una tipologia di analisi piuttosto che un'altra non pregiudica la possibilità di adottare parallelamente anche le altre tipologie. Come si vede le principali tipologie in cui vengono utilizzati i Big Data

[66] *Osservatorio Big Data Analytics & Business Intelligence*, https://www.osservatori.net/it_it/osservatori/chi-siamo
[67] Ivi.

nelle aziende sono la Descriptive e Predictive Analysis, analisi che descrivono la situazione in corso di un evento, e che cercano di dare un'idea su cosa possa accadere in futuro. Tali tipologie di analisi sono le più adottate, perché erano utilizzate nelle aziende ancor prima dell'avvento dei Big Data, ciò nonostante inserendo i Big Data in tali analisi le aziende possono ottenere vantaggi competitivi, nonché efficienza ed efficacia nelle loro decisioni, che prima erano fuori dalla loro portata. Un campo che necessità sicuramente tempo per crescere, ma che è il vero terreno su cui far fruttare i Big Data è l'Automated Analytics, ossia strumenti che permettono di automatizzare azioni dati i risultati delle analisi svolte. Tali sistemi sono ancora in fase embrionale di sperimentazione all'interno delle aziende, perché necessitano di tempo per essere sviluppate data la loro complessità, ma è già evidente il loro potenziale, si pensi alle potenziali legate all'Intelligenza Artificiale.

Figura 14: Principali ostacoli rivelati nel trattare i Big Data[68]



Come si vede dall'immagine gli ostacoli incontrati dalle grandi imprese nel trattare i Big Data, sono in linea con le criticità esposte nel paragrafo precedente. I problemi principali incontrati sono: la mancanza di impegno (o motivazione), da parte dei manager e dirigenti, nell'utilizzare i Big Data per prendere le decisioni o orientare i loro processi. Tale mancanza di impegno è dovuta principalmente, secondo chi scrive, ad un gap culturale nelle aziende italiane, sia nei confronti del tema dei Big Data sia verso un orientamento data-driven e fact-based. Il secondo problema

[68] *Osservatorio Big Data Analytics & Business Intelligence*, https://www.osservatori.net/it_it/osservatori/chi-siamo

principale è la mancanza di competenze e figure organizzative interne, è necessario che le aziende concentrino i loro investimento nel reperimento o formazioni dei così detti Data Scientist. La figura del Data Scientist negli ultimi anni, è arrivato anche da noi sia nell'ambito accademico, con percorsi universitari specifici o master di formazione, sia nell'ambito lavorativo. Il Data Scientist, ovvero colui che in azienda ha il compito di svolgere l'analisi dei dati: la sua attività è in particolar modo attinente a quei processi che richiedono l'utilizzo di tecniche avanzate, informatiche, statistiche i di machine learning. Egli porta con sé un insieme di competenze che consentono di utilizzare i dati per generare vantaggi competitivi. Il profilo ideale di un data scientist comprende numerose competenze, che è molto difficile trovare in un'unica persona[69]:

- Competenze informatiche riguardanti in particolare:
    - Database relazionali e linguaggio SQL
    - Hadoop e il suo "ecosistema" di software
    - Spark
    - Database NoSQL, in particolare quelli presenti nella piattaforma Hadoop
    - Linguaggi di programmazione fortemente orientati all'analisi dei dati, quali R o Python
    - Tool di analisi e visualizzazione avanzati.
- Competenze statistiche.
- Conoscenza delle tecniche di predictive analytics e machine learning.
- Conoscenza dei processi di gestione della qualità del dato.
- Conoscenza dei processi di business e dell'organizzazione aziendale.
- Conoscenza delle problematiche e delle sfide del dominio settoriale di interesse (industria, finanza, telecomunicazione…).
- Capacità di comunicare i risultati dell'analisi a tutti i destinatari.
- Capacità di colloquiare con i manager e di supportarli nelle decisioni critiche.

È ovvio, quindi, che il singolo data scientist avrà livelli elevati di competenze solo in alcune di questi campi e potrà, attraverso l'esperienza all'interno dell'azienda,

[69] A. Rezzani, *Big Data Analytics*, Maggioli Editore 2017, p.11

acquisire conoscenze ulteriori, in particolare legate al business e ai processi interni. Un approccio che funziona quasi sempre molto bene consiste nella creazione di un team di data scientist, ognuno dei quali apporta al gruppo competenze diverse e volte verso differenti aspetti informatici, statistici o di business.

In questo capitolo abbiamo compreso finalmente cosa si intende per Big Data, in che modo attraverso questi dati è possibile ottenere un vantaggio competitivo e quali sono le principali sfide che devono affrontare le aziende per ridurre il rischio di fallimento nell'adottare tale innovazione. Nel terzo capitolo andremo a delineare una metodologia su come implementare un sistema di BI che sfrutti tutte le potenzialità dei Big Data.

3 Progettare la Business Intelligence

3.1 L'importanza della metodologia

Come abbiamo avuto modo di scoprire nel primo capitolo, la BI ha un costo complessivo che può essere anche molto elevato. Se è oneroso implementare la BI, ancor più lo è intraprendere un progetto fallimentare, che genererebbe solo costi e nessun beneficio economico. Per tale ragione è importante prestare massima attenzione a ciascuna fase del ciclo di vita del progetto. Lo schema di lavoro proposto costituisce un approccio alla progettazione della BI che minimizza i rischi di esito negativo del progetto. Tale approccio si basa su alcuni punti fermi[70]:

1. Una fase preliminare di *assessment* (valutazione) della situazione corrente in azienda che permetta di definire quali sono i problemi dovuti alla mancanza del sistema di BI, quali, in linea generale, le esigenze degli utenti e quali benefici si avrebbero implementando la BI. L'assessment ha lo scopo di far conoscere l'azienda e la sua situazione a livello di patrimonio di dati, procedure e sistemi di analisi a tutti gli attori che prendono parte al progetto. Inoltre l'identificazione di costi e benefici rende possibile la stima del ROI, come visto nel primo capitolo.
2. La progettazione deve essere sempre guidata da un principio di massimizzazione del *valore del sistema di BI per il business*; per questo

[70] A. Rezzani, *Business Intelligence: Processi, Metodi utilizzo in Azienda*, Maggioli Editore 2012, p.63

non deve essere guidata dalla tecnologia, ma dai requisiti che provengono dagli attori aziendali.

3. La costruzione di *prototipi* finalizzati a validare la modellazione dei dati, confermarne la corrispondenza con le regole di business aziendali e mostrare agli utenti soluzioni a problemi di particolare importanza o complessità.
4. L'*approccio iterativo* alla raccolta dei requisiti, alla modellazione dei dati e all'implementazione di ETL, OLAP e reporting. Approccio iterativo significa che, pur avendo una chiara visione generale, si affrontano di volta in volta singoli processi di business evitando un approccio onnicomprensivo, che porta a risultati concreti solo dopo molto tempo e non permette una verifica periodica da parte degli utenti. Inoltre, ammettendo la possibilità di far validare dagli utenti quanto effettuato, si possono eseguire più cicli implementativi relativi allo stesso processo di business, in caso di errori o di discrepanze rispetto ai requisiti.
5. La *documentazione* delle varie fasi attraverso il modello concettuale, documenti tabellari in formato standard o documento in testo libero.
6. Il *tracciamento* della provenienza del singolo dato, in modo da poter sempre ripercorrere a ritroso la strada che ha portato un elemento di una dimensione, oppure una riga della tabella dei fatti dalla fonte alla destinazione.
7. La predisposizione di *più ambienti* di BI con finalità distinte:
    a. L'ambiente di prototipazione;
    b. L'ambiente di sviluppo;
    c. L'ambiente di test;
    d. L'ambiente di produzione.

    Ciascun ambiente dovrà essere completo e quindi includere il data warehouse, l'ETL, le strutture OLAP e il front-end. La suddivisione in ambienti con finalità diverse permette di tenere ben distinte le macro-fasi del progetto (prototipo, sviluppo, testing, produzione) consentendo a più figure di lavorare in parallelo, senza interferenza.

La metodologia di lavoro è quella schematizzata nella figura sottostante, si rileva che il flusso di lavoro è composto di quattro fasi[71]:

- L'analisi preliminare (assessment e giustificazione economica del progetto).
- L'attività di pianificazione, con cui si determinano tempi e scadenze per le fasi successive.
- La costruzione di prototipi.
- Il processo iterativo inizia con la raccolta dei requisiti e comprende, oltre a una fase analitica, la progettazione del data warehouse, la progettazione e creazione del front end e, da ultima, la verifica del lavoro svolto.

Inoltre, all'interno del processo iterativo, le fasi per l'implementazione dell'OLAP, reportining e data mining sono caratterizzate da un proprio ciclo implementativo.

3.2 Le figure professionali

La complessità del sistema di BI richiede un team di persone con competenze eterogenee, sia informatiche, sia relative a processi di business. Le figure professionali necessarie all'implementazione di un sistema di BI sono elencate nella seguente tabella, dove accanto alla figura professionale è presente anche una colonna che indica il ruolo svolto:

Tabella 2: Figure Professionali[72]

Figura/Ruolo	Attività
Project Manager	Pianifica, coordina e controlla le attività. Comunica con i responsabili dell'azienda, con i consulenti e con i fornitori di software.
Esperto di Architetture di BI	Progetta l'infrastruttura del sistema di BI.
Esperto del Business	Fornisce le regole del business e prende decisioni su problematiche del business.

[71] A. Rezzani, *Business Intelligence: Processi, Metodi utilizzo in Azienda,* Maggioli Editore 2012, p.64
[72] A. Rezzani, *Business Intelligence: Processi, Metodi utilizzo in Azienda,* Maggioli Editore 2012, p.66

Esperto di Data Quality	Svolge un'attività di profiling delle fonti dati e fornisce le regole di pulizia.
Data Modeler	Crea il modello concettuale e logico dei dati ed è coinvolto nella costruzione del modello fisico.
Database Administrator	Partecipa alla creazione del modello fisico dei dati e si occupa di tutti i problemi legati alle performance.
Sistemista/Esperto di rete	Predispone e amministra l'ambiente hardware.
Esperto Storage	Predispone e amministra l'ambiente di storage dei dati.
Esperto di ETL	Progetta e partecipa allo sviluppo di ETL.
Sviluppatore ETL	Sviluppa ETL.
Esperto di Data Mining	Predispone i modelli di data mining. Si occupa di definire i traning set e test set.
Esperto OLAP	Progetta e sviluppa il database multidimensionale.
Sviluppatore report/Front-end	Sviluppa report o altre componenti di front-end (es. applicazioni web, portali).
Addetti al Test	Verificano le procedure di ETL, il processo di pulizia dei dati e il funzionamento del front-end.
Manager/Utilizzatori	Contribuiscono alla definizione dei requisiti, verificano quanto implementato e forniscono un feedback.
Sponsor	Si tratta di una figura interna all'azienda in grado di sostenere e promuovere la BI. Il suo compito principale, è quello di rimuovere eventuali ostacoli posti da elementi interni all'azienda.

Va notato come ruoli diversi possono essere svolti dalla stessa persona, a patto che ne abbia le competenze. Si sottolinea inoltre l'importanza dell'esperto del business e il feedback da parte dei manager e degli utenti che usufruiscono dei servizi offerti dalla BI. Come avremo modo di evidenziare nel corso del capitolo, per creare un sistema di BI efficiente ed efficace è fondamentale che sia il business a orientare il

dipartimento IT (l'ufficio preposto alla gestione e allo sviluppo dei sistemi di BI) durante le fasi di sviluppo di un sistema di BI, e non viceversa.

3.3 Le fasi del progetto

Esaminiamo adesso le fasi che compongono la metodologia di progettazione discussa nel primo paragrafo del capitolo. Non entreremo troppo nel dettaglio tecnico, ma si cercherà di dare una visione d'insieme su come implementare un sistema di BI, approfondendo in particolare il rapporto tra business e IT.

3.3.1 Analisi preliminare e giustificazione del progetto

Nella fase preliminare si effettua un'attività di verifica della situazione aziendale sotto diversi punti di vista. Bisogna prima di tutto stabilire quali sono le necessità analitiche e quali problemi sono presenti in azienda a causa della loro mancata soddisfazione. Successivamente bisogna capire quali strumenti hardware e software siano già presenti in azienda, sia per acquisire la conoscenza dei sistemi operazionali, sia per comprendere l'attuale entità delle mansioni amministrative e manutentive. Tale valutazione può essere effettuata attraverso un'indagine o un questionario da somministrare al personale tecnico, agli analisti e al management. Un esempio di questionario può essere quello proposto qui di seguito[73]:

Tabella 3:Questionario per Assesment[74]

Area	**Domande**
Il Business	▪ Qual è il business principale? ▪ Vi sono altre attività oltre al business principale? ▪ Quali sono e che responsabilità hanno le business unit? ▪ Quante persone lavorano in azienda? ▪ Com'è composto il personale?
Problemi di analisi	▪ Qual è il tempo medio di realizzazione di un report? ▪ Come accedete ai dati?

[73] A. Rezzani, *Business Intelligence: Processi, Metodi utilizzo in Azienda*, Maggioli Editore 2012, p.68
[74] Ivi.

	▪ Quante persone sono dedicate al reporting o alla costruzione di analisi? ▪ Quanto tempo ciascuna di esse dedica, al mese, all'attività di reporting? ▪ Vi sono problemi di accesso alla fonte dati? ▪ Occorre unire dati provenienti da più fonti? ▪ Esistono più anagrafiche clienti/prodotti/fornitori? ▪ Come giudicare la qualità dei dati? ▪ Dovete fare attività di pulizia, uniformazione dei formati, verifica della correttezza dei dati? ▪ Quanto tempo richiede l'attività di pulizia dei dati? ▪ Come sono estratti i dati dai sistemi sorgente?
Problemi di Business	▪ I report e le analisi sono prodotti in modo tempestivo? ▪ I report e le analisi conducono a prendere decisioni corrette? ▪ Avete la certezza o perlomeno la sensazione di aver perso clienti a causa di decisioni errate? ▪ Avete la certezza o perlomeno la sensazione di non aver acquisito clienti a causa di decisioni errate? ▪ Con che criterio sono svolte le campagne di marketing? ▪ Come controllate i centri di costo? ▪ Ritenente probabile la mancata identificazione di inefficienze che portano a maggiori costi? ▪ Ritenete possibile migliorare il processo produttivo, tenendo sotto controllo costi, scorte e utilizzo del personale? ▪ Come potrebbe essere impiegate le risorse oggi dedicate allo sviluppo di report/analisi e pulizia dati?
Strumenti - Hardware	▪ Quali piattaforme hw utilizzate? ▪ In caso di hw aggiuntivo per la BI, occorreranno maggiori risorse umane per gestirlo? ▪ Com'è implementata la rete LAN?
Strumenti – Software	▪ Com'è strutturato il software operazionale? ▪ In caso di acquisto di un nuovo RDBMS, il personale IT ha le competenze per gestirlo? ▪ In caso di acquisto di un nuovo RDBMS, il personale IT ha sufficiente da dedicare alla sua gestione? ▪ Con quali strumenti sono attualmente implementati i report?

Requisiti Analitici	▪ Quali sono i report o le analisi che realizzate con maggior frequenza? ▪ Quali sono i report o le analisi che vorreste realizzare, ma che non siete in grado di produrre con il sistema attuale? ▪ Quale grado di dettaglio hanno le vostre analisi? ▪ Avete l'esigenza di effettuare analisi previsionali?
Standard Aziendali	▪ Avete degli standard di sviluppo? ▪ Avete degli standard nella nomenclatura di database, tabelle e OLAP.
Desiderate	▪ Quali sono le caratteristiche che il sistema di BI dovrebbe avere? ▪ Quanto è importante la facilità d'uso? ▪ Quanto è importante la pronta disponibilità dell'informazione? ▪ Quanto è importante la qualità dei dati? ▪ Che profondità storica desiderate mantenere? ▪ Siete interessati a una reportistica di tipo statico? ▪ Siete interessati a una reportistica di tipo dinamico?

L'intervista può essere condotta sia attraverso domande a risposta aperta, sia con domande a risposta chiusa. Si evidenzia che l'ausilio di domande aperte pur fornendo risposte dettagliate può presentare il rischio di divagazioni inutili che comportano una perdita di tempo. Le domande chiuse riducono i tempi di risposta e di controllo del questionario, inoltre permettono di porre questioni mirate sugli obbiettivi del progetto; tuttavia non consentono di entrare nel dettaglio come fanno le domande aperte[75].

L'attività di valutazione dovrebbe permettere di identificare quali sono i problemi correlati alla mancanza di un sistema di BI, consentendo di stimare la riduzione dei costi e i maggiori ricavi che si realizzerebbero con la sua implementazione. Il progetto di implementazione di un sistema di BI trova la sua giustificazione sia nella soluzione alle problematiche di analisi, che possono essere identificate

[75] A. Rezzani, *Business Intelligence: Processi, Metodi utilizzo in Azienda*, Maggioli Editore 2012, p.67

attraverso il questionario visto precedentemente, sia nei benefici economici, che abbiamo visto nel primo capitolo parlando del valore della BI. È chiaro che fornire una solida giustificazione economica aumenta le motivazioni del committente, degli utenti e del team che si occupa della realizzazione del sistema di BI[76].

Nella fase preliminare è anche importante tenere in considerazione i fattori di rischio che possono insorgere durante il progetto, e che conducono ad un esito insoddisfacente. Il primo "tipo" di rischio in cui certamente ci si può imbattere è il rischio tecnologico, che dipende dal grado di maturità della piattaforma prescelta: un conto è lavorare con una tecnologia matura e ben conosciuta, un altro è utilizzare tecnologie nuove, appena apparse sul mercato. Un altro tipo di rischio e collegato alla complessità del sistema da implementare. Per esempio, si potrebbe trovare difficoltà in un sistema che richiede l'estrazione dati da moltissime fonti eterogenee, la quali hanno forti problemi a essere integrate e pertanto necessitano di complesse regole di pulizia e trasformazione. A quanto detto finora, si aggiungono rischi legati alla componente umana del progetto. Da un lato abbiamo la parte manageriale dell'azienda che ha dei limiti di tolleranza circa la tempistica e i costi del progetto, limiti che devono essere oltrepassati. Dall'altro esiste un rischio legato alla capacità, alle competenze e all'impegno del team di progettazione e sviluppo: occorre valutare bene gli skill tecnologici, progettuali e di business dei componenti del gruppo e la loro esperienza nella realizzazione di sistemi di BI[77].

3.3.2 Pianificazione delle attività

Pianificare le attività significa individuare ciò che deve essere consegnato, la scadenza, l'impegno in termini di ore di lavoro e quindi i costi, e chi dovrà svolgere le attività. Senza addentrarci nella vasta materia del *project management*, possiamo dire che nel pianificare il progetto occorre svolgere le seguenti operazioni[78]:

- Dividere il lavoro complessivo in macro-attività, e a sua volta in compiti di ambito più ridotto.
- Effettuare una stima delle ore necessarie per svolgere ciascun compito.

[76] A. Rezzani, *Business Intelligence: Processi, Metodi utilizzo in Azienda,* Maggioli Editore 2012, p.68
[77] Ivi, p.69
[78] Ivi, p.69

- Allocare le persone.
- Dare un ordine di priorità alle attività e verificare eventuali dipendenze tra i vari compiti.
- Creare un piano di lavoro dettagliato e condividerlo.

3.3.2.1 Raccolta dei requisiti:

Con questa fase inizia il processo iterativo che terminerà con la creazione del front-end e con la misurazione del successo degli elementi di BI rilasciati. Come abbiamo visto precedentemente una prima raccolta dei requisiti generici avviene nella prima fase del progetto, durante l'analisi preliminare nello specifico nel momento della valutazione. La fase di valutazione ha lo scopo di portare alla luce i problemi generati dall'assenza del sistema di BI e di far emergere quali sono le caratteristiche generali che il sistema dovrà avere. Invece, la raccolta dei requisiti che trattiamo ora avviene a ciascuna iterazione e riguarda un singolo processo di business (o parte di esso), raggiungendo un elevato livello di dettaglio. Lo scopo di questa attività di dettaglio consiste nel comprendere a fondo il processo di business e interpretare le esigenze analitiche per arrivare a creare modelli dati e strumenti di presentazione efficaci e coerenti. Anche in questa fase l'individuazione dei requisiti avviene attraverso interviste che possono utilizzare domande a risposta aperta, grazie alle quali potranno emergere molti particolari relativi al processo di business. Spesso accade che i requisiti non siano espressi in maniera esplicita, ma possono manifestarsi come lamentele di manager e utenti nei confronti di una situazione problematica. In qualsiasi forma venga condotta l'intervista, essa deve portare all'identificazione degli eventi da misurare e delle entità di business rilevanti per le analisi. Dalla raccolta dei requisiti deve dunque scaturire una documentazione completa che comprenda[79]:

- La descrizione dettagliata del processo di business;
- I requisiti analitici e le loro proprietà;
- Le entità che entrano in gioco per ciascun processo (o attività del processo);
- I fatti da misurare.

[79] A. Rezzani, *Business Intelligence: Processi, Metodi utilizzo in Azienda,* Maggioli Editore 2012, p.69

L'esplicitazione delle entità (clienti, prodotto, tempo, venditore…) e la loro associazione ai processi a parte di essi è proposta da Kimball[80]nella cosiddetta *bus matrix*. Questo approccio è molto utile all'identificazione di tutte le dimensioni d'analisi e delle loro relazioni con le attività del processo. Un esempio di tale matrice è mostrato nella seguente figura:

A ogni iterazione, un nuovo processo di business è analizzato e la matrice è arricchita con le attività e le entità che lo riguardano. Alcune entità possono essere le stesse di processi già analizzati. La matrice mostra in modo immediato la condivisione delle entità tra i vari processi di business. Con la raccolta dei requisiti è possibili iniziare la creazione del modello concettuale dei dati e, di fatto, la matrice entità\processi ne rappresenta una parte[81].

3.3.2.2 Creazione del prototipo:

Come detta in precedenza, la costruzione di un prototipo è utile a far comprendere agli utenti come saranno implementati certi aspetti del sistema, oppure a dimostrare la fattibilità e l'esattezza di una soluzione sia dal punto di vista tecnologico sia da quello della modellazione dei dati. Il prototipo ha necessariamente un ambito limitato sia di funzionalità, sia di dati con cui è alimentato. Ciò consente la rapida realizzazione di un modello da sottoporre agli utenti. Nello schema visto ha inizio capitolo, la creazione del prototipo è inserita dopo la fase di raccolta dei requisiti e in parallelo all'analisi delle fonti dati. Infatti la creazione di un prototipo è possibile solo dopo aver compreso le richieste degli utenti e dopo aver esaminato le fonti dati, anche se in modo parziale. Il prototipo è molto utile se, una volta realizzato, coinvolge gli utenti di business, che potranno così validarlo, toccando con mano un applicativo già nelle prime fasi del progetto. Tuttavia esso non deve essere identificato con il prodotto finito: il rischio è che gli utenti comincino a utilizzare un applicativo che è stato pensato per fornire un esempio di funzionamento, ben

[80] R. Kimball and M. Ross, *The data warehouse toolkit: the complete guide to dimensional modelling*, Jhon Wiley and Sons, New York, 2002
[81] A. Rezzani, *Business Intelligence: Processi, Metodi utilizzo in Azienda*, Maggioli Editore 2012, p.69

lontano quindi da rappresentare un'applicazione completa, ma da esso si aspettino tutte le caratteristiche espresse nei requisiti. È possibile che l'azienda, solo per il fatto di tagliare i costi del progetto, decida consapevolmente di adottare il prototipo come soluzione definitiva, incurante del fatto che questo manchi di un testing accurato, non sia completo e non copra dunque tutti i requisiti. Pertanto è fondamentale esplicitare in maniera chiara l'ambito del prototipo, le funzionalità implementate e il suo scopo. Vi sono alcuni semplici principi da seguire per costruire un prototipo utile al successo del progetto[82]:

1. Sviluppare il prototipo velocemente (l'ambito del prototipo è limitato, pertanto non serve molto tempo per svilupparlo).
2. È importante riceve un feedback da parte degli utenti per migliorarlo iterativamente, pertanto mostrare spesso il prototipo agli utenti è molto utile.
3. Non cambiare lo scopo per cui era stato creato il prototipo in corsa.

Questa fase è fondamentale per evitare problemi che sarebbero noti solo alla fine degli sviluppi, o comunque in una fase avanzata degli stessi, causando allungamenti dei tempi di realizzazione e incremento dei costi.

3.3.2.3 Analisi delle fonti dati:

La fase più importante nello sviluppo di un progetto di BI riguarda l'analisi delle fonti dati relative al processo di business coinvolto. Essa consente di documentare tutti gli aspetti che serviranno come punto di partenza per la definizione dei processi di pulizia e trasformazione, che fanno parte dell'ETL. Prima di tutto bisogna documentare le caratteristiche principali delle sorgenti dei dati, che sono[83]:

- Denominazione del sistema sorgente;
- La tecnologia su cui è basato il sistema sorgente;
- Il nome delle tabelle/file/web service/ da cui recuperare i dati;
- Il numero di campi che costituiscono il record;

[82] A. Rezzani, *Business Intelligence: Processi, Metodi utilizzo in Azienda*, Maggioli Editore 2012, p.71
[83] A. Rezzani, *Business Intelligence: Processi, Metodi utilizzo in Azienda*, Maggioli Editore 2012, p.72

- Il numero di righe presenti nella fonte dati al momento dell'analisi;
- La dimensione in byte o kilobyte della singola riga;
- La frequenza di aggiornamento dei dati (giornaliera, settimanale, mensile);
- Il numero di righe che variano a ogni aggiornamento;
- Una o più note descrittive relative a problematiche particolari (commenti, osservazioni circa la qualità dei dati…).

La conoscenza approfondita della fonte dati si ottiene però attraverso il *profiling* degli attributi, che consiste nell'analizzare i valori contenuti in ciascun campo. Le informazioni principali che è possibile rilevare per ciascun campo sono[84]:

- Nome del campo;
- Tipo di dato (intero, data, stringa…);
- Valore minimo;
- Valore massimo;
- Numero di valori distinti;
- Numero di valori nulli (valore non presente);
- Note varie;
- Considerazioni relative alla data quality.

Nel censire i dati a disposizione è necessario identificare i dati più importati (master data), ossia le entità importanti per il business (cliente, prodotto…), caratterizzate da numerosi attributi e con cardinalità elevata. Di solito i master data si trovano in più sistemi (il cliente, per esempio, lo troviamo nel sistema ERP, e nel CRM e nei database dell'area marketing), ognuno dei quali può presentare differenze nelle anagrafiche: vi può essere un numero di attributi diverso, oppure valori diversi per lo stesso elemento. Per questo, durante l'analisi delle fonti dati, l'identificazione dei dati master e la verifica della loro qualità sono di fondamentale importanza, in quanto un contributo alla realizzazione del sistema di data quality[85].

[84] Ivi, p.72
[85] A. Rezzani, *Business Intelligence: Processi, Metodi utilizzo in Azienda*, Maggioli Editore 2012, p.73

3.3.2.4 Definizione delle regole di data quality:

Il livello di tolleranza dei dati "sporchi" è differente per ciascuna delle figure aziendali che li utilizzano: generalmente il livello manageriale non tollera in alcun modo dati imprecisi, mentre gli analisti o chi lavora a contatto con i dati hanno imparato con il tempo a interpretare dati poco accurati. Di conseguenza i requisiti di qualità che scaturiscono dalle interviste in azienda sono diversi per ogni figura professionale interpellata. Tuttavia, l'analisi e il profiling dei dati rendono evidenti, in maniera oggettiva, il livello di qualità dei dati. Inoltre, basandoci sulle richieste più stringenti che emergono dalle interviste, possiamo ottenere delle regole di pulizia che soddisferanno pienamente tutti i requisiti di qualità. Per raccogliere i requisiti di qualità dei dati è necessario prima di tutto effettuare interviste che propongano domande generiche sulla percezione della qualità e sulle esigenze da soddisfare. In seguito si scenderà nel dettaglio per determinare, assieme ai rappresentati dell'azienda, le regole di controllo e di pulizia dei singoli campi[86].

Nel secondo capitolo, parlando dell'importanza dei dati, abbiamo già visto quali sono gli elementi che determinano la qualità dei dati: completezza, consistenza, accuratezza, integrità e assenza di duplicazioni. Tali caratteristiche vengono definite fattori di qualità intrinseca del dato, in quanto riguardano il dato in sé e sono facilmente e oggettivamente misurabili. La qualità dei dati è analizzabile anche in base a fattori che non sono intrinseci del dato, ma riguardano il servizio dati, cioè le modalità con cui sono messi a disposizione degli utilizzatori[87]:

- *Accessibilità*: l'accessibilità fa sì che ciascun utente, nell'ambito delle proprie prerogative di sicurezza, possa facilmente fruire dei dati.
- *Disponibilità*: per disponibilità del sistema si intende la possibilità di accedere al sistema, e pertanto dovrebbe essere posto al riparo da guasti o da inconvenienti tecnici di ogni genere.
- *Copertura*: riguarda la capacità del sistema di descrivere i business nella sua interezza, è possibile che il servizio dati descriva solo una parte del business, lasciando così scoperte, oppure parzialmente interessate, altre aree.

[86] Ivi, p.74
[87] Ivi, p.186

- *Facilità d'uso*: dati disponibili e completi sono utili se il loro modo d'utilizzo è complesso e poco comprensibile agli utenti.

Tutti i fattori di qualità elencati sono misurabili in maniera oggettiva attraverso indicatori che l'attività di valutazione consente di valorizzare. Assicurare la qualità dei dati è un'attività complessa che coinvolge e comporta la definizione di regole e processi, la designazione di persone responsabili della qualità e l'utilizzo di tecnologie. Tali elementi costituiscono il sistema di data quality. Senza addentrarci troppo nel sistema di data quality, possiamo dire che implementando tale sistema, di certo l'azienda otterrà dei benefici, uno su tutti la disponibilità di dati puliti e consistenti per l'alimentazione dei sistemi di BI[88].

La raccolta dei requisiti, la creazione di prototipo, l'analisi delle fonti dati e la definizione delle regole di data quality pongono le basi per la modellazione del data warehouse, lo sviluppo delle procedure ETL, la progettazione dei cubi OLAP e la creazione di report e del data minig. Nel primo capitolo abbiamo già parlato ampiamente dell'architettura di un datawarehouse, delle procedure ETL e dello sviluppo di database multidimensionali OLAP, pertanto nei prossimi paragrafi ci soffermeremo sugli ultimi aspetti del progetto di BI.

3.3.2.5 Progettazione e sviluppo di report (e dashboard):

I report costituiscono il modo più comune di diffondere la Business Intelligence. Possiamo definire un report come un documento, elettronico o cartaceo, che contiene dati consultabili dagli utenti, presenti per mezzo di semplici tabelle oppure attraverso elementi grafici. Un report può contenere analisi, aggregazioni dei dati o calcoli che l'utente potrà valutare e interpretare. Con il progredire delle tecnologie il reporting ha subito un'evoluzione che ha portato uno strumento, dapprima esclusivamente cartaceo, a diventare consultabile attraverso i dispositivi mobili (tablet e smartphone). L'attività di reportistica nasce come reporting direzionale, volto a produrre l'insieme di rendiconti con i quali il management prende le proprie decisioni, rivedendo periodicamente le proprie azioni. Il reporting direzionale si

[88] A. Rezzani, *Business Intelligence: Processi, Metodi utilizzo in Azienda*, Maggioli Editore 2012, p.187

differenzia in base al livello di management a cui è rivolto: strategico, tattico o operativo. Ciascuna delle tre tipologie è caratterizzata da contenuti e da modalità di presentazione dei dati diversi[89].

- Il reporting strategico contiene dati molto aggregati e espone l'andamento di alcune grandezze chiave che riassumono la situazione aziendale. Questo tipo di reporting si presenta spesso in forma grafica. La frequenza con cui i report strategici sono prodotti non è quasi mai giornaliera, ma settimanale o addirittura mensile.
- Il reporting tattico ha un livello di aggregazione meno elevato e presenta i dati utilizzando in misura minore gli elementi grafici. Il reporting tattico ha spesso finalità di controllo e presenta confronti tra l'andamento delle grandezze analizzate e gli obbiettivi previsti. I report possono avere frequenza giornaliera, settimanale o mensile in base agli obiettivi da monitorare.
- Il reporting operativo presenta un livello di aggregazione minimo, esponendo anche dati di dettaglio, utilizzabili nelle attività operative.

Le diversità del business, le esigenze analitiche e la varietà di informazione utili al management di qualsiasi livello per prendere le decisioni, rendono praticamente impossibile definire a priori un insieme di report che possa essere valido per qualsiasi azienda. Anche all'interno di una singola realtà le esigenze di reporting possono essere molto variabili. Infatti, accanto a report che potrebbero essere utili per molto tempo, può nascere un insieme di report specifici volti a soddisfare esigenze analitiche di breve termine[90].

Oltre ai classici report vengono utilizzate anche le dashboard, che non sono altro che una rappresentazione semplice e sintetica degli indicatori di performance aziendale attraverso strumenti visivi quali grafici, diagrammi, etc. Sono destinate prevalentemente a utenti del livello strategico e spesso sono il punto di partenza di analisi più approfondite svolte con altri strumenti (i report per esempio). Le dashboard ricomprendono al loro interno diverse componenti tra le quali troviamo

[89] A. Rezzani, *Business Intelligence: Processi, Metodi utilizzo in Azienda*, Maggioli Editore 2012, p.240
[90] A. Rezzani, *Business Intelligence: Processi, Metodi utilizzo in Azienda*, Maggioli Editore 2012, p.244

anche KPI (*Key Performance Indicator*), che, attraverso un numero o un indicatore grafico, presentano la situazione di un indicatore di sintesi dell'andamento di una particolare grandezza[91].

La figura a inizio capitolo descrive questo punto come un vero e proprio ciclo di sviluppo, effettivamente, l'implementazione del front-end della BI prevede più fasi, che spesso costituiscono un processo attivo[92]:

- Raccolta dei requisiti specifici per la costruzione di report e dashboard.
- Recupero dei dati creando le query opportune sul data warehouse o sul cubo OLAP, costruendo così l'insieme dei dati da visualizzare.
- Disegno del layout grafico (tabelle, diagrammi, grafici, immagini) per la presentazione dei dati.
- Pubblicazione del report o della dashboard.
- Amministrazione, che consiste, per esempio, nella gestione dei permessi di visualizzazione oppure nella programmazione degli eventuali invii automatici di report agli utenti.
- Raccolta di feedback da parte degli utenti ed effettuazione di eventuali modifiche.

È chiaro che tali punti possono essere adattati in base alle tecnologie a disposizione e alle richieste degli utenti.

3.3.2.6 Progettazione e sviluppo del data minig:

La necessità di effettuare analisi previsionali e la grande quantità di dati a disposizione oggi, ha reso necessaria l'adozione di tecniche di analisi efficienti e in grado di lavorare su valori numerici, testuali o binari (per esempio le immagini). Le tecniche di analisi a cui ci riferiamo prendono il nome di data mining, poiché consentono di "scavare" nei dati ed estrarre informazioni, *pattern*[93] e relazioni non immediatamente identificabili e non note a priori. Il data mining può essere

[91] Ivi, p.258
[92] Ivi, p.84
[93] Per pattern intendiamo una struttura, un modello, o più in generale una rappresentazione sintetica dei dati. A. Rezzani, *Business Intelligence: Processi, Metodi utilizzo in Azienda,* Maggioli Editore 2012, p.267

utilizzato in qualsiasi settore economico per risolvere molteplici problemi di business[94].

- *Ricerche di anomalia*. Il data minig può essere impiegato per l'identificazione di comportamenti fraudolenti nell'utilizzo di carte di credito.
- *Churn Analysis*. La churn analysis consiste nell'analisi della clientela per determinare i clienti che presentano un'alta probabilità di passare alla concorrenza, al fine di intervenire in anticipo ed evitare migrazione.
- *Segmentazione della clientela*. Le tecniche di segmentazione possono essere, per esempio, a determinare il profilo comportamentale dei clienti. Una volta identificati i segmenti di clienti simili, è possibile studiare strategie di marketing differenziate per ciascun gruppo.
- *Previsioni*. Le analisi predittive dell'andamento delle vendite, o, genericamente dell'andamento di serie temporali, sono un altro degli ambiti di impiego del data minig.
- *Campagne pubblicitarie mirate*. L'utilizzo del data minig nell'ambito delle campagne di marketing, consente di stabilire a priori quali siano, tra i prospect, quelli con maggior probabilità di acquistare i prodotti dell'azienda, in modo da impiegare si di esse le risorse del marketing.
- *Market basket analysis*. Le tecniche di market basket analysis sono utili a suggerire, a un certo cliente, ulteriori prodotti da acquistare in base ai suoi comportamenti d'acquisto abituali, oppure a definire il layout dei prodotti sugli scaffali.
- *Sentiment analsysis*. Si tratta di un'applicazione di data mining ai social network. Per usare una definizione più semplice si potrebbe spiegare come un metodo di analisi che raccoglie in tempo reale le reazioni degli utenti o trend davanti a un qualsiasi evento, locale o globale che sia. Rappresenta uno strumento accurato per individuare ed ascoltare le conversazioni online fornendo alle aziende un'interpretazione del mercato molto realistica. Infatti, rispetto ai ben noti sondaggi telefonici o alle analisi di mercato che riguardano ristretti focus group dove, la possibilità di ottenere opinioni

[94] A. Rezzani, *Business Intelligence: Processi, Metodi utilizzo in Azienda*, Maggioli Editore 2012, p.267

"pilotate" è molto alto, l'utilizzo della sentiment analysis è in grado di fornire all'azienda una definizione veritiera sulla reputazione del brand e sul gradimento dei suoi prodotti[95]. Approfondiremo tale tipologia di analisi nel prossimo capitolo, effettuando anche una sentiment analysis su dati provenienti da Twitter.

Il data minig comporta l'utilizzo integrato di diverse discipline, come il data warehousing, la statistica, l'intelligenza artificiale, le tecniche di visualizzazione, l'analisi delle serie temporali e l'analisi di dati geo-spaziali. La componente tecnologica riveste una grande importanza, poiché gli algoritmi di data minig richiedono una certa potenza di calcolo e tecniche di ottimizzazione delle performance sono essenziali, soprattutto in presenza di una mole di dati elevata. Nel processo di data mining è però la figura dell'utente ad assumere un ruolo centrale; si tratta infatti, di un processo che richiede l'iterazione di un esperto del business, che deve sfruttare la propria conoscenza per la preparazione dei dati, per la costruzione dei modelli e per la valutazione dei risultati[96].

Anche il data minig è rappresentato attraverso un ciclo di sviluppo iterativo, le cui fase possono essere riassunte nei seguenti punti[97]:

- Analisi dei requisiti e piena consapevolezza del business e dei dati a disposizione.
- Campionamento e preparazione dei dati.
- Preparazione del modello.
- Test e valutazione del modello, da cui può risultare la necessità di effettuare dei cambiamenti.
- Consegna agli utenti.
- Feedback da parte degli utenti ed eventuali revisioni.

La fase di preparazione riveste particolare importanza, in quanto, anche se si lavora su dati puliti che soddisfano i requisiti di data quality aziendali, i modelli di data mining necessitano di dati opportunamente trasformati. Ciò comporta l'utilizzo di procedure ETL appositamente concepite per la preparazione dei dati di input dei

[95] Definizione dal sito: http://www.beantech.it
[96] A. Rezzani, *Business Intelligence: Processi, Metodi utilizzo in Azienda,* Maggioli Editore 2012, p.268
[97] Ivi,p.84

modelli. Altrettanto importante è la valutazione del modello, che avviene attraverso strumenti specifici. Se i risultati del test sul modello non sono soddisfacenti, si ritorna alla fase di preparazione del modello, variandone i parametri e ripetendo poi la valutazione, fino a raggiungere i risultati soddisfacenti.

3.3.2.7 Misurazione del successo:

L'ultima fase del ciclo di sviluppo consiste nell'ottenere un feedback da parte degli utenti che consenta di capire in che misura i requisiti siano stati soddisfatti. Il livello di soddisfazione e i commenti sono utilizzati per effettuare eventuali aggiustamenti, che saranno pochissimi o nulli se sono state svolte con attenzione e precisione le prime fasi del progetto. Vi sono varie modalità per misurare il successo di un progetto: raccogliendo feedback dagli utenti, attraverso interviste dirette o questionari, e anche tramite la misurazione di alcuni eventi. In quest'ultimo caso il successo del progetto può essere ricavato dalle seguenti grandezze[98]:

- Percentuale di utenti che effettivamente utilizzano il sistema. È possibile ottenere questa informazione misurando i log di accesso al sistema di reporting o all'applicazione OLAP.
- Il numero di applicazioni che sfruttano le informazioni provenienti dal sistema di BI. Infatti un modo per valorizzare i sistemi di BI è quello di creare applicativi che si basino sui dati contenti nel DW o che espongano semplicemente dei link a report.
- Numero di nuove richieste. Se gli utenti giudicano positivamente il lavoro svolto dal team BI, sicuramente porranno numerose nuove richieste.
- Numero di report *ad hoc* sviluppati dagli utenti. Gli strumenti di reportistica *ad hoc* consentono agli utenti avanzati di sviluppare report per sé o per altri utenti, in autonomia, senza passare per il team di sviluppo del sistema di BI.
- Dismissione dei vecchi sistemi di analisi. Se la percezione del sistema di BI è positiva, gli utenti abbandoneranno i vecchi sistemi di analisi.

[98] A. Rezzani, *Business Intelligence: Processi, Metodi utilizzo in Azienda*, Maggioli Editore 2012, p.85

Si comprende se il progetto di BI ha avuto successo o meno anche valutando se si sono effettivamente ottenuti dei benefici economici dall'implementazione del sistema. Di conseguenza è importante misurare gli effetti del sistema di BI in termini di maggiori ricavi e di minori costi[99].

3.4 Regole per il successo ed errori da evitare

Terminiamo il capitolo con l'analisi sia degli elementi che contribuiscono al successo del progetto di implementazione della BI, e di quelli che invece ne possono determinare il fallimento. Possiamo affermare come principio generale che, l'obiettivo da raggiungere sia la creazione di un sistema di Business Intelligence che abbia un'ottica di lungo periodo. L'implementazione del data warehouse, la creazione e il controllo delle regole di data quality e la realizzazione di un ETL efficiente sono evidentemente operazioni costose, ma pongono le basi per una BI solida e duratura. Al contrario, la tentazione di "scavalcare" il data warehouse e creare un sistema di reportistica basato direttamente sulle fonti dati operazionali può portare risparmi sull'investimento iniziale, ma non generare i benefici tipici di un sistema di BI[100].

Qui di seguito verranno elencati una serie di "regole" che possono aiutare ad avere successo e una serie di azioni da evitare per non rischiare di fallire. Per la definizione di tali elementi è stato preso spunto dal libro di Alessandro Rezzani "Business Intelligence: processi, metodi utilizzo in azienda", che è stato utilizzato anche in altre parti dell'elaborato, e dalla esperienza da me acquisita in questo ambito[101].

---

[99] A. Rezzani, *Business Intelligence: Processi, Metodi utilizzo in Azienda,* Maggioli Editore 2012, p.86

[100] Ivi,p.86

[101] Ho effettuato uno stage di 6 mesi presso l'azienda Bata nella sede di Limena (PD), che è la sede europea per quanto riguarda il brand Bata ed è la sede mondiale del brand AW. Ho lavorato nell'ufficio IT a stretto contatto con il project manager responsabile di tutti i sistemi di BI, avendo modo di apprendere tutte le dinamiche che vanno dall'acquisizione dei dati sino alla creazione di report. Successivamente sono stato assunto come consulente dalla societa MHT che fa parte del gruppo Engineering, e che propone consulenza in ambito IT (sistemi ERP, CRM e BI) di tutti i prodotti Microsoft. Lavorando nel team di BI all'interno di MHT ho avuto modo di approfondire tutte le tematiche inerenti la business intelligence (data waherhouse, ETL, data quality e reportisca) e ho iniziato a conoscere e studiare argomenti legati al Data Mining.

Regole per il successo[102]:

- *Ottenere una forte sponsorship in azienda.* Per progetti molto grandi, come l'implementazione di un sistema di BI, occorre un appoggio interno all'azienda in grado di sostenere il progetto e rimuovere gli ostacoli che le persone, a vari livelli, potranno porre.
- *Coinvolgere gli utenti.* A partire dalle fasi iniziali del progetto è assolutamente necessario avere un apporto di idee, commenti e considerazioni da parte degli utenti, al fine di raccogliere i requisiti e problematiche nel modo più completo possibile.
- *Utilizzare un metodo di sviluppo collaborativo.* Il coinvolgimento degli utenti deve continuare nelle fasi successive del progetto, anche attraverso la veloce realizzazione di prototipi.
- *Utilizzare un metodo di sviluppo iterativo.* La focalizzazione su un singolo processo di business permette di consegnare una soluzione in tempi abbastanza brevi. Particolare attenzione va posta sulle dimensioni che saranno utilizzate anche in altri processi di business.
- *Puntare alla qualità dei dati.* La qualità dei dati deve essere l'aspetto predominante che guida la costruzione del data warehouse. La qualità deve essere preferita alla quantità, quindi piuttosto che utilizzare dati non validi, imprecisi o incompleti, è meglio lasciarli fuori dal data warehouse.
- *Disciplina nella documentazione.* La documentazione di ciascuna fase permette di tenere una precisa traccia dei requisiti e della loro implementazione. Di fondamentale importanza è la documentazione relativa alle trasformazioni ETL e alla mappatura dei campi tra sorgente e destinazione, che fornisce una sorta di tracciabilità del dato.
- *Fare attenzione alle performance di caricamento e consultazione.* Le performance di caricamento dei dati rivestono una certa importanza, in quanto la finestra temporale per eseguire il caricamento del data warehouse potrebbe essere limitata. È bene quindi progettare in anticipo il sistema per favorire la velocità di inserimento dei dati. La performance di consultazione riveste un'importanza ancora maggiore, visto che essa è direttamente

[102] A. Rezzani, *Business Intelligence: Processi, Metodi utilizzo in Azienda,* Maggioli Editore 2012, p.87

percepita dagli utenti. Dovremo dunque utilizzare tutte le funzionalità che l'RDBMS e il motore OLAP mettono a disposizione per velocizzare il recupero dei dati, cercando di disegnare il data warehouse e la base dati OLAP in modo da privilegiare le soluzioni più performanti.

- *Promozione degli strumenti di BI implementati.* Una volta implementati gli strumenti di BI, è necessario favorirne la diffusione e l'utilizzo da parte di tutti gli utenti, inviando notifiche della disponibilità del nuovo sistema, o facendo dimostrazioni pratiche direttamente agli interessati.
- *Formazione regolare degli utenti.* Alcuni strumenti di BI richiedono un ciclo di formazione degli utenti. La formazione va programmata e portata a termine, pena un rischio subottimale degli strumenti.

Errori da evitare[103]:

- *Non relazionarsi direttamente con gli utenti.* Già nella fase di valutazione, non relazionarsi con gli utenti e fidarsi solo di consulenti o esperti esterni è molto rischioso e non consente di cogliere pienamente i requisiti e i problemi del sistema attuale.
- *Non coinvolgere il management.* Tra gli attori aziendali da coinvolgere, sicuramente troviamo il management aziendale che, pur non essendo direttamente utente del sistema di BI, deve essere messo a conoscenza dei benefici economici che esso porta all'azienda. Ciò è strettamente legato alla ricerca di uno sponsor adeguato. Saltare questo passaggio può comportare il rischio di non ottenere una sponsorship forte e di non riuscire a giustificare il progetto di realizzazione della BI.
- *Progettare tutto nei minimi dettagli, prima di iniziare gli sviluppi.* Raramente si sono visti finire con successo progetti di BI e data warehousing partiti con la progettazione totale e dettagliata di tutto il sistema. Vi sono due motivi principali per questo: innanzitutto è molto difficile prevedere ogni dettaglio e ogni problematica di tutti i processi aziendali al fine di produrre un documento di progetto che, dato in mano agli sviluppatori, porti alla realizzazione dell'intero sistema di BI. In

[103] A. Rezzani, *Business Intelligence: Processi, Metodi utilizzo in Azienda,* Maggioli Editore 2012, p.87

secondo luogo, è possibile che in mancanza di risultati, anche parziali, consegnati in tempi abbastanza brevi, il committente mostri insofferenze e decida di ridurre l'ambito del progetto o cessarlo definitivamente. È chiaro che una visione generale di quanto si dovrà implementare è necessaria, ma questo non significa che si debba per forza arrivare a dettagliare minuziosamente ogni aspetto del data warehouse.

- *Non implementare il DW.* Seguendo la volontà di contenere i costi o seguendo le affermazioni di alcuni venditori che promettono l'implementazione della BI direttamente sul sistema operazionale, si può decidere, erroneamente, di implementare un sistema di BI senza il disegno e l'implementazione del data warehouse. Bisogna però tenere in mente che, senza il dw molti dei vantaggi tipici del sistema di BI verrebbero meno, in quanto:
    - Non si implementa il sistema di data quality;
    - Non si ha una struttura denormalizzata che agevola l'interrogazione;
    - Data la mancanza del dw, i dati rimangono sparsi in più basi dati eterogenee;
    - Per ogni analisi che si realizza, occorre ripetere i procedimenti di uniformazione, pulizia e integrazione dei dati.
- *Disegnare il data mart per produrre un report specifico*. Spesso si incorre nell'errore di modellare un data mart sulla base di un report che gli utenti di business presentano come requisito. Questa situazione può portare a un'estrema rigidità del disegno, che sarà in grado di soddisfare quell'unico esigenza, ma non sarà in grado di adattarsi ad altre richiesta di analisi.
- *Considerare la Business Intelligence come un insieme di report*. Progettare il sistema di BI con l'unico scopo di produrre report è molto limitativo. È vero che il reporting costituisce una parte del front-end, tuttavia non bisogna dimenticarsi che esistono anche altri mezzi efficaci per comunicare fatti rilevanti per il business, quali i KPI, le scorecard e le dashboard. Inoltre il mancato utilizzo di strumenti OLAP fa perdere la caratteristica della flessibilità al front-end della BI.

È chiaro che questo elenco di fattori che possono aiutare a raggiungere il successo ed errori da evitare se non si vuole fallire nell'implementazione di un sistema di BI,

non è esaustivo, ma da un'idea dell'approccio da perseguire. Secondo la mia esperienza personale un elemento fondamentale nella creazione e nella gestione di un sistema di BI è sicuramente avere una profonda conoscenza della base dati da trattare e una visione di medio lungo termine, su quale compito deve espletare il sistema di BI.

Dopo aver delineato una metodologia nell'implementazione di un sistema di BI, nel prossimo capitolo vedremo un caso pratico di analisi dei Big Data, e nell'implementazione di un semplice sistema di BI.

4 Sentiment Analysis

4.1 Data Mining

In questo ultimo capitolo andremo a vedere un caso pratico di gestione e analisi dei Big Data, valutando il valore aggiunto da loro offerto ai vari processi di business e nella creazione di nuova conoscenza. Prima di addentrarci nel caso pratico, è opportuno chiarire alcuni concetti già visti nei capitoli precedenti e aggiungere nuove nozioni.
Partiamo col definire cosa si intende per "Data Minig". Come abbiamo già visto nel terzo capitolo, una delle attività da svolgere nel processo di BI è la progettazione e lo sviluppo del data minig. Per definire il Data Mining (DM), senza creare confusione, riportiamo alcune definizioni estrapolati dalla letteratura che insieme permettono di chiarire il concetto di DM:

- *Il data mining è uno step nel processo di Knowledge Discovery in Database (KDD), che consiste nell'applicare analisi di dati e algoritmi di individuazione che, sotto accettabili limiti di efficienza computazionale, producono una particolare enumerazione di patterns (modelli) sui dati*[104].

- *Il Data Mining è la ricerca di relazioni e modelli globali che sono presenti in grandi database, ma che sono nascosti nell'immenso ammontare di dati, come le relazioni tra i dati dei pazienti e le loro diagnosi mediche. Queste relazioni rappresentano una preziosa conoscenza del database e, se il database è uno specchio fedele del mondo, tali relazioni sono una valida rappresentazione della realtà*[105].

[104] Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "*From data mining to knowledge discovery in databases.*" AI magazine Volume 17 Number 3 (1996), p.41
[105] Holsheimer, Marcel, and Arno Siebes. *Data mining: The search for knowledge in databases*. Centrum voor Wiskunde en Informatica, 1994, p.1

- *Il Data Mining è un processo aziendale per esplorare grandi quantità di dati per scoprire patterns e regole significativi*[106].

- *Il termine data mining si riferisce quindi all'intero processo che si articola nella raccolta e nell'analisi dei dati, nello sviluppo di modelli matematici e nell'adozione di decisioni e azioni concrete basate sulle conoscenze acquisite.* [107].

Dalle definizioni sopra citate si possono discernere alcuni punti chiave che ci chiariscono il concetto di DM:

- Il DM è un processo aziendale;
- È uno step del processo di KDD;
- Analizza grosse mole di dati;
- Si appoggia a dei modelli matematici;
- Si basa su metodi di apprendimento induttivo;
- I risultati ottenuti devono essere patterns (modelli) e regole significative.

In questo paragrafo si cercherà di chiarire tutte queste caratteristiche che nell'insieme rappresentano il processo di DM. Alcuni dei punti sopra elencati possono essere risolti velocemente. Innanzitutto, come abbiamo visto nel secondo capitolo con l'avvento dei Big Data, grosse mole di dati eterogenei, le attività di Business Intelligence si sono dovute trasformare adeguandosi hai nuovi dati, una diretta conseguenza di questa evoluzione dei processi di BI è l'uso di algoritmi di DM per l'analisi dei dati. Agli albori del DM (anni '60 e '70), i dati erano scarsi. A quei tempi, un set di dati tipico poteva avere pochi attributi ed essere composto da poche centinaia di record. Ciò nonostante qualunque sia il set di dati, le tecniche di DM funzionano correttamente. Poiché non è difficile oggi disporre di macchine con

[106] Linoff, Gordon S., and Michael JA Berry. *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons, 2011, p.1
[107] Vercellis C., *Business Intelligence. Modelli matematici per le decisioni*, Mc Graw Hill, Milano, 2006, p.77

una buona potenza di calcolo, una grande quantità di dati non è un handicap anzi è un vantaggio. In altre parole, il DM consente ai computer di fare ciò che i computer sanno fare meglio: scavare attraverso un sacco di dati. Questo a sua volta, consente alle persone di fare ciò che sanno fare meglio, ovvero impostare il problema e comprendere i risultati[108].

Con il termine *modelli matematici di apprendimento* (matematical learning theory) si fa riferimento all'insieme di modelli e metodi matematici che sono il fulcro di ogni analisi di DM e che vengono utilizzati per generare nuove conoscenze[109]. Inoltre, il processo di DM si basa su metodi di apprendimento induttivo, in quanto cerca di ricavare regole generali a partire da una sequenza di esempi rappresentati dalle osservazioni riferite al passato e accumulate nei database. In altre parole, le analisi di DM si propongono di trarre alcune conclusioni a partire da un campione di osservazioni passate, che vengono generalizzate all'intera popolazione, in modo che possano essere il più accurate possibili.

4.1.1 Data Mining è un Business Process

Il data mining è un processo aziendale che interagisce con altri processi aziendali. Nello specifico, un processo non ha un inizio né una fine: è continuo. Il DM inizia con i dati, la cui analisi informa o ispira l'azione aziendale che crea ulteriori dati dando origine a più attività. La conseguenza pratica è che le organizzazioni, che desiderano eccellere nell'utilizzare i propri dati per migliorare le loro attività, non considerano il data mining come una attività secondaria. Infatti, la loro strategia aziendale include la raccolta e l'analisi dei dati per ottenere risultati e benefici a lungo termine. Inoltre, il data mining si adatta facilmente a molte strategie per la valutazione e segmentazione di mercati e clienti. Ricerche di mercato, panel di clienti e altre tecniche, sono compatibili con il data mining. La chiave è focalizzarsi sui clienti e condividere i dati e le informazioni da essi estrapolate all'interno dell'azienda[110].

---

[108] Linoff, Gordon S., and Michael JA Berry. *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons, 2011, p.2

[109] Vercellis C., *Business Intelligence. Modelli matematici per le decisioni*, Mc Graw Hill, Milano, 2006, p.77

[110] Linoff, Gordon S., and Michael JA Berry. *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons, 2011, p.1

4.1.2 Modelli e regole Significativi

La parte più importante della definizione di data mining è la parte sugli schemi significativi. Sebbene il data mining possa certamente essere un'attività divertente, aiutare l'azienda è più importante che divertire il data scientist. In molti casi la ricerca di schemi nei dati non è così difficile da attuare. Le attività operative dell'azienda generano dati e, di conseguenza, generano dei patterns. Tuttavia, l'obiettivo del data mining non è quello di trovare ogni schema presente nei dati, ma di estrarre quei patterns che si rivelano utili per il business. Questo significa trovare schemi per aiutare le operazioni aziendali di routine. O trovare patterns per ottimizzare l'acquisizione dei clienti, sia per i guadagni nel breve termine (aumento del numero di clienti) sia per benefici a medio e lungo termine (aumento del valore dei clienti). Alcune aziende aggregano dati da diverse fonti e li riuniscono per ottenere un'immagine più completa dei mercati e dei clienti. Altre aziende, come LinkedIn, utilizzano le informazioni fornite dai propri utenti per fornire servizi premium ad altri; i benefici sono di tutti, degli utenti in cerca di lavoro e dei reclutatori che possono così trovare i candidati giusti per le posizioni di lavoro aperte. In tutti questi casi, l'obiettivo è indirizzare prodotti e servizi alle persone che hanno maggiori probabilità di averne bisogno, rendendo il processo di acquisto e vendita più efficiente per tutti i soggetti coinvolti[111].

4.1.3 Knowledge Discovery in Database (KDD)

Senza addentrarci troppo nell'argomento discutiamo dell'ultima caratteristica del DM da esaminare il processo KKD. Come è stato detto il DM fa parte di un più ampio processo chiamato Knowledge Discovery in Database (KKD). Lo scopo di tale processo, come è facilmente intuibile, è di estrarre la conoscenza dai dati. Il dato è il risultato della misurazione di un certo evento e di per sé non ha grande utilità. Tuttavia, valutando i dati in un preciso contesto e applicandovi opportune elaborazioni, è possibile ricavare informazioni: in questo modo il dato diventa utile e in grado di rispondere ad una richiesta specifica. La conoscenza è quindi il

[111] Linoff, Gordon S., and Michael JA Berry. *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons, 2011, p.3

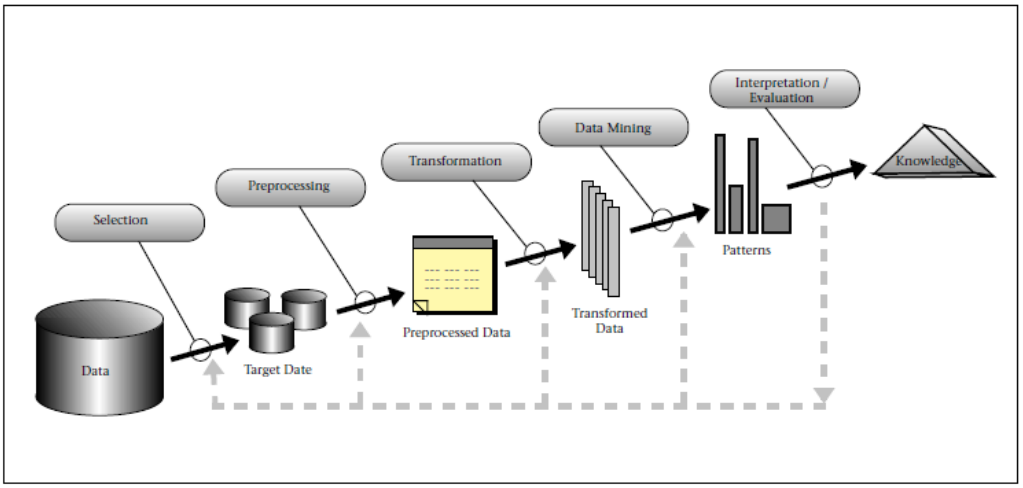
risultato di un percorso che parte dai dati grezzi e termina con l'interpretazione e lo sfruttamento produttivo dei risultati[112].

Una definizione di KKD ci viene fornita da Usama Fayyad, Gregory Piatetsky-Shapiro e Padhraic Smyth, nel loro saggio: "*From Data Mining to Knowledge Discovery in Databases*":

- *La KDD è il processo non banale di identificazione di patterns validi, nuovi, potenzialmente utili e comprensibili nei dati[113].*

Oltre a fornisci una definizione gli autori sopra citati hanno identificato cinque fasi che caratterizzano il processo di KDD:

Figura15:Una panoramica dei passaggi che compongono il processo KDD[114].



1- **Selezione dei dati**: è importante comprendere il dominio applicativo determinato dagli obiettivi dell'utente finale.
2- **Preelaborazione**: nella fase di preelaborazione rientrano le attività di pulizia dei dati e di definizione del trattamento dei dati mancanti. Qualora i dati provenissero da un data warehouse, molte delle operazioni di pulizia dovrebbero essere già state svolte nella fase di ETL.
3- **Trasformazione**: del tipo da numero a stringa, o per esempio la discretizzazione di valori continui o la normalizzazione dei valori.

[112] A. Rezzani, *Business Intelligence: Processi, Metodi utilizzo in Azienda*, Maggioli Editore 2012, p.269
[113] Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "*From data mining to knowledge discovery in databases.*" AI magazine Volume 17 Number 3 (1996), p.40
[114] Ivi, p.41

4- **Data Mining**: tale fase vede la determinazione dell'algoritmo da utilizzare, la costruzione e il testing di un modello.
5- **Interpretazione dei risultati**: i risultati del modello di DM esprimono schemi (patterns) nascosti nei dati, bisogna verificare se tali schemi sono utili e possono effettivamente portare un beneficio al business[115].

Da come si evince dalla descrizione delle varie fasi del processo di KKD, l'estrazione della conoscenza è un processo lungo e spesso articolato, in cui sono fondamentali le scelte per il trattamento delle anomalie o errori nei dati e l'identificazione chiara degli obiettivi che si vogliono perseguire. Non vi è una fase più importante o più significativa dell'altra, ma certamente la fase di DM è quella che ha ricevuto maggior attenzione in letteratura e che è oggetto di approfondimento in questo paragrafo.

4.1.4 Uno sguardo ravvicinato al data mining

Il DM implica l'uso integrato di diverse discipline (Statistica, Informatica, Economia, etc..), generalmente esso costituisce un'attività cooperativa tra esperti dell'ambito applicativo e analisti dei dati, che si avvalgono di modelli matematici di apprendimento.
È possibile suddividere le attività di DM in due gruppi distinti in base a due orientamenti principali di indagine: l'interpretazione e la predizione[116]:

- L'interpretazione è volta a identificare schemi di regolarità nei dati e a esprimerli attraverso regole e criteri che risultano chiari agli esperti del contesto applicativo. Le regole generate devono essere originali e quindi non scontate, in modo da incrementare il livello di conoscenza e comprensione nei riguardi del sistema analizzato.
- La predizione si propone di prevedere il valore che una variabile casuale assumerà in futuro, ossia cerca di stimare la probabilità dell'accadimento di eventi futuri. Le varie tecniche di DM ricavano la

[115] A. Rezzani, *Business Intelligence: Processi, Metodi utilizzo in Azienda,* Maggioli Editore 2012, p.269
[116] Vercellis C., *Business Intelligence. Modelli matematici per le decisioni*, Mc Graw Hill, Milano, 2006, p.79

predizione sulla base del valore di alcune variabili associate alle entità presenti in un database.

A prescindere dall'orientamento principale di indagine, è bene evidenziare le differenze che intercorrono tra analisi di DM e analisi di statistica classica o di tipo OLAP. Le indagini effettuate attraverso le tecniche di statistica classica e i cubi multidimensionali (OLAP), si limitano a fornire elementi a conferma o a smentita delle ipotesi formulate a priori dai decision maker, secondo un flusso di analisi di tipo *top-down*. Invece i modelli di DM, svolgono una funzione attiva, ossia sono in grado di generare predizioni e interpretazioni che costituiscono effettivamente nuova conoscenza a disposizione degli utilizzatori. Il flusso di indagine in questo caso è di tipo *bottom-up*, specie in presenza di grosse mole di dati, è indispensabile disporre di modelli capaci di svolgere un ruolo attivo, perché risulterebbe molto difficile per il knowledge worker formulare a priori ipotesi significative e attendibili[117].

Nel terzo capitolo abbiamo già evidenziato vari ambiti applicativi e metodologie di analisi del DM e, ciò che manca da descrivere in conclusione di questo paragrafo, è la metodologia richiesta per la costruzione di un modello di DM. La costruzione di un modello di DM si articola in più fasi[118]:

- *La scelta dell'algoritmo di calcolo*: tale step parte dall'analisi del problema di DM da risolvere e, in base alle caratteristiche del problema evidenziate, si sceglie l'algoritmo che meglio si adatta a tale situazione. I principali algoritmi utilizzati nelle analisi di DM sono: Naive Bayesian Classifier (classificazione bayesiana), Cluster Analysis, Decision Tree (alberi decisionali), Association Rule, Reti neurali e Analisi di serie temporali.
- *Il completamento della fase di preparazione dei dati*: spesso l'algoritmo richiede che i dati siano puliti o in formati particolari; a tal fine è necessario ultimare la fase di preparazione dei dati attraverso tecniche specifiche quali: la normalizzazione del dato, il trattamento dei dati mancanti e degli outlier (errori nei dati o dati anomali), la riduzione del dataset (riduzione della

[117] Vercellis C., *Business Intelligence. Modelli matematici per le decisioni*, Mc Graw Hill, Milano, 2006, p.80
[118] A. Rezzani, *Business Intelligence: Processi, Metodi utilizzo in Azienda*, Maggioli Editore 2012, p.280

dimensionalità del data set, del numero di valori nel dominio degli attributi, del numero di record per la costruzione del modello).

- *La scelta dei parametri base di configurazione dell'algoritmo*.
- *La suddivisione dei dati disponibili in traning set e test set*: Nel costruire un modello di DM è necessario operare una suddivisione dei dati disponibili in due insiemi: uno contenente buona parte dei dati costituisce il training set, cioè l'insieme dei dati su cui l'algoritmo scelto è calibrato. L'altro rappresenta il test set, ossia l'insieme di dati su cui si eseguirà il test del modello per verificarne la bontà. Ovviamente, il test set contiene anche l'attributo o gli attributi che sono il risultato dell'attività predittiva del modello. Perseguendo questa metodologia sarà possibile confrontare i dati reali con quelli previsti ed eseguire così una valutazione. È molto importante che la suddivisione in traning set e test set avvenga mantenendo la stessa distribuzione degli attributi in entrambi gli insiemi di elementi, in modo che essi siano ugualmente rappresentativi.
- *L'avvio della fase di training dell'algoritmo*: Nella fase di training l'algoritmo analizza le relazioni nascoste nei dati e imposta il modello di DM.

Come detto precedentemente il processo di DM è un processo iterativo, pertanto è spesso necessario ripetere più volte il processo di costruzione del modello, al fine di provare algoritmi alternativi o anche per poter operare degli aggiustamenti sui parametri o sulle trasformazioni del data set[119].

Dopo aver ben definito l'ambito di applicazione del DM e i vari step che compongo il suo processo di analisi, nel prossimo paragrafo ci concentreremo su uno specifico modello di DM, il Text Mining o meglio in un caso particolare di Text Mining la Sentiment Analysis.

[119] A. Rezzani, *Business Intelligence: Processi, Metodi utilizzo in Azienda,* Maggioli Editore 2012, p.280

4.2 Introduzione al Text Mining

Come abbiamo visto nel secondo capitolo, l'epoca che stiamo vivendo può essere definita l'era dei Big Data; mai come in questo periodo c'è stata una così alta disponibilità di dati tra loro eterogenei. Un contributo notevole alla creazione e diffusione dei dati è dovuto all'avvento di Internet e, in particolare, alla genesi delle varie forme di social network: dai piccoli blog ai grandi social network quali Facebook o Twitter. La costante crescita di volume e la varietà dei dati sono un fenomeno in perenne movimento. Questo processo ha reso disponibili non solo molti dati ma soprattutto dati eterogenei, ossia di formati diversi (numerici, testuali, immagini etc, ridondanti e non strutturati). Questa varietà nei dati ha spinto gli analisti a ricercare nuove forme di analisi o ad adattare le analisi convenzionali a queste nuove tipologie di dati. In questo paragrafo parleremo di una particolare analisi di Data Mining che agisce sui dati testuali, nello specifico Text Mining e Sentiment Analysis.

I social media sono nati non solo come mezzi di comunicazione personali, ma anche come mezzi per divulgare e condividere opinioni su prodotti e servizi, eventi politici e generali, tra i suoi utenti. La diffusione e la popolarità dei social media ha comportato un forte aumento del numero di recensioni e di opinioni prodotte e condivise quotidianamente dagli utenti[120]. È diventato pertanto indispensabile poter analizzare documenti scritti, recenti stime affermano che l'85% delle informazioni aziendali vive sotto forma di testi. Essere in grado di analizzare dati testuali è diventato oggi un vantaggio competitivo per le aziende. È possibile analizzare i dati testuali attraverso l'uso di algoritmi di Text Mining. Gli algoritmi di Text Mining mirano a svelare le informazioni nascoste nei testi attraverso metodi che da un lato sono in grado di far fronte al gran numero di parole e strutture in linguaggio naturale, dall'altro permettono di gestire le confuse e ambigue relazioni presenti nei testi[121].

Il Text Mining è un processo automatizzato che estrae nuove conoscenze, relazioni e schemi (pattern) da dati testuali non strutturati. Questo campo di ricerca utilizza algoritmi di data mining come classificazione, clustering, regole di associazione e

[120] YOUNIS, Eman MG. *Sentiment analysis and text mining for social media microblogs using open source tools: An empirical study*. International Journal of Computer Applications, 2015, p.45

[121] HOTHO, Andreas; NÜRNBERGER, Andreas; PAAß, Gerhard. *A brief survey of text mining*. In: Ldv Forum. 2005. p. 1

molti altri ancora nell'esplorazione e nella scoperta di nuove informazioni e relazioni in fonti testuali. Si tratta di un campo di ricerca interdisciplinare che combina l'information retrieval, il data mining, il machine learning, la statistica e la linguistica computazionale[122].

All'interno del Text Mining rientrano quindi problemi di classificazione dei testi, clustering, riconoscimento ed estrazione di informazioni ed entità, sentiment analysis e opinion mining, riconoscimento del topic e information retrieval engine. Per poter risolvere questi problemi legati all'analisi dei testi, si fa ricorso a vari campi e tecniche di ricerca come abbiamo visto, tra cui analisi di Natural Language Processing (NLP), o trattamento del linguaggio naturale[123]. Le tecniche di NLP possono essere utilizzate per analizzare i testi a livelli differenti, ad esempio:

- Analisi lessicale;
- Analisi sintattica;
- Analisi semantica, che analizza i possibili significati del discorso;
- Analisi del discorso.

Generalmente le analisi sui testi si basano sulla semantica, e pertanto sull'estrazione o parsing di un testo per poter analizzarne gli elementi costitutivi: soggetto, verbo, complemento. Tuttavia, è prassi in ambito di Text Mining focalizzarsi sul concetto di "**Bag of Words**". Con tale concetto si fa riferimento al fatto che un testo viene analizzato in funzione delle parole che lo compongono, facendo particolare attenzione alle parole chiave, alle entità, cioè all'estrazione di nomi propri, luoghi, eventi, relazioni e date, oppure sull'analisi del sentiment di un testo[124].

Affinché una collezione di testi possa essere oggetto di analisi di Text Mining, i testi devono possedere alcune caratteristiche in comune, per esempio: devono trattarsi di opere dello stesso autore, o raccolta di documenti su un determinato tema, messaggi di una determinata comunity di internet o riguardanti un brand o un prodotto, oppure ancora un campione di interviste. È necessario inoltre, che i testi oggetto di analisi abbiano un livello minimo di pulizia, di normalizzazione del testo e alcuni standard per la preparazione e lo scambio di testi in formato elettronico[125].

[122] YOUNIS, Eman MG. *Sentiment analysis and text mining for social media microblogs using open source tools: An empirical study*. International Journal of Computer Applications, 2015, p.45
[123] PORCU, Valentina. Text mining e sentiment analysis con R. Valentina Porcu, 2016, p.3
[124] Ivi
[125] PORCU, Valentina. Text mining e sentiment analysis con R. Valentina Porcu, 2016, p.4

Le applicazioni classiche di Text Mining comprendono, anche, la classificazione dei documenti. Un esempio di tale applicazione può essere l'uso di dataset di e-mail etichettati come spam o non spam per classificare ulteriori e-mail non etichettate, attraverso la creazione di modelli predittivi. Anche le tecniche di Clustering fanno parte del Text Mining, ad esempio per raggruppare testi cha hanno lo stesso argomento. Il processo di clustering permette di raggruppare in maniera automatica degli elementi (in questo caso dei testi) basandosi sulle loro similarità. I documenti pertanto non sono etichettati all'inizio come appartenenti ad un gruppo, ma in base ai termini che contengono è possibile classificarli[126].

L'ambito in cui vengono utilizzate maggiormente tali applicazioni di Text Mining sono le ricerche di marketing, è possibile infatti utilizzare le tecniche di clustering a dati testuali per la segmentazione della clientela. L'applicazione delle tecniche di clustering permette di individuare gruppi omogenei di clienti, in termini di comportamenti d'acquisto e di caratteristiche sociodemografiche. La possibilità di segmentare la clientela in gruppi omogenei è molto importante perché permette di effettuare campagne marketing mirate, ottenere indicazioni su come modificare la propria offerta, e rende possibile monitorare nel tempo l'evoluzione della propria clientela e l'emergere di nuove tipologie. Un'altra delle applicazioni più comuni è l'analisi delle associazioni, la quale permette di trovare associazioni di dati di vendita per scoprire quali prodotti sono acquistati tipicamente in maniera congiunta o alternativa. Ottenere questo tipo di informazioni permette di migliorare l'offerta dei prodotti e di incrementare le vendite attraverso le vendita dei prodotti complementari. Oltre alle applicazioni citate, il Text Mining ha altre svariate applicazioni e ambiti applicativi; è evidente che il Text Mining è un campo vastissimo e non è possibile affrontarne tutte le tematiche nel corso di questa sola tesi.[127]

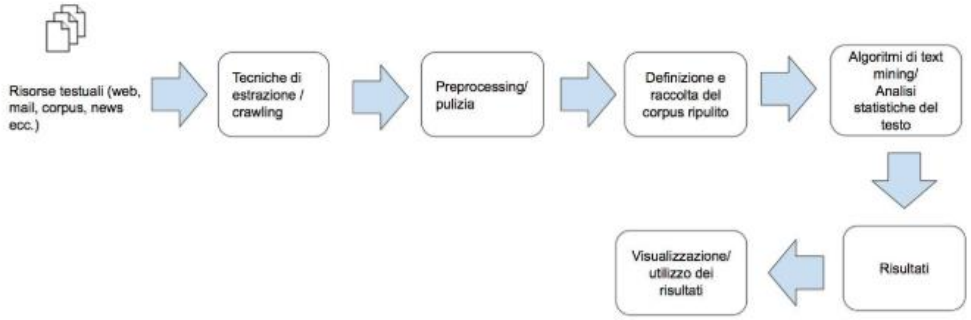
L'obiettivo di questo paragrafo è quello di introdurre il concetto di Text Mining con i suoi ambiti applicativi e dare un rapido sguardo al suo processo. Ci concentreremo in maniera più approfondita nel prossimo paragrafo su una delle applicazioni di Text Mining: la Sentiment Analysis.

[126] Ivi, p.5
[127] Ivi

4.2.1 Il Processo di Text Mining

L'oggetto delle analisi di Text Mining, come abbiamo visto, sono dati testuali i quali rientrano nella categoria di dati non strutturati. Al fine di applicare le procedure di Text Mining, è necessario trasformare i dati testuali in dati strutturati. Spesso l'estrazione dei dati e la loro preparazione per l'analisi occupano la maggior parte del tempo rispetto alle fasi di analisi vere e proprie. Il primo passo è quelli di raccogliere i dati e quindi i documenti di analisi; la fonte di dati per eccellenza al giorno d'oggi è il web, da cui derivano la maggior parte dei documenti di analisi. Questa è anche una delle ragioni per cui il Text Mining è diventato così importante per le aziende[128].

Figura 16: Step del processo di Text Mining[129]



La figura qui in alto mostra i vari step del processo di Text Mining. Una volta che i dati testuali sono stati raccolti, la prima procedura da effettuare comprende il preprocessing del documento, o del corpus di documenti. La fase di preprocessing comprende diverse attività di pulizia e trasformazione dei dati testuali. I principali step che vengono sempre effettuati sono[130]:

1. la tokenizzazione del testo;
2. la rimozione delle stopwords;
3. lo stemming e il rooting;
4. la normalizzazione;
5. l'identificazione dei limiti di frasi i preposizione all'interno di periodi complessi, e l'identificazione dei limiti delle parole;

[128] PORCU, Valentina. Text mining e sentiment analysis con R. Valentina Porcu, 2016, p.24
[129] Ivi.
[130] Ivi, p.25

6. la rimozione dei numeri e degli spazi di troppo;
7. la rimozione della punteggiatura;

Brevemente descriveremo ogni step, va detto però che il preprocessing può essere differente a seconda degli obiettivi della ricerca e quindi necessita di una certa flessibilità che va decisa in base agli scopi dell'analisi.

Uno dei primi step del fase di preprocesing è la tokenizzazione di un testo, o corpus di documenti. **Tokenizzare** un testo significa dividere il documento in parole, dove ogni parola rappresenta un token[131]. Se considero la frase:

"Francesco va in vacanza mentre Roberto va a lavorare"

avremo 8 token, tuttavia all'interno della frase ci sono anche delle parole ripetute, in questo caso "va". Pertanto, abbiamo sì nove token, ma otto type. Il type, a differenza del token, è la singola parola, che può essere ripetuta più volte, anche con varie declinazioni. Per esempio, se consideriamo la frase:

"Francesco va in vacanza mentre Roberto andrà a lavorare"

il numero di token rimarrebbe nove in ogni caso, ma anche il numero di type rimarrebbe sempre otto, perché abbiamo la stessa parola, anche se in una diversa forma verbale, ossia "va" e "andrà". Per riassumere il type è il singolo elemento del vocabolario, pertanto se all'interno di una frase una parola è ripetuta più volte, sarà comunque contata una volta sola. Il token invece, sono tutte le parole, quindi nel conteggio vengono comprese anche le parole ripetute. L'obiettivo della tokenizzazione è quello di vettorializzare il testo, ossia di dividerlo in singoli termini. Una volta vettorializzato il testo, sarà possibile analizzarlo tramite il conteggio delle parole, o eliminare i termini più comuni e così via[132].

Il secondo step, nel processo di preprocessing è la rimozione delle **stopwords**, ossia l'eliminazione delle cosiddette "parole vuote". Per parole vuote si intende tutte le preposizioni, avverbi e congiunzioni che sono comuni a tutti i testi. I vari software preposti al Text Mining, o più in generale al Data Minig, come per esempio R, hanno a disposizione vari dizionari, specie in inglese, che consentono di eliminare

[131] PORCU, Valentina. Text mining e sentiment analysis con R. Valentina Porcu, 2016, p.26
[132] Ivi, p.27

da un corpus le parole vuote più diffuse. È chiaro che la rimozione delle stopwods sarà differente in relazione agli scopi dell'analisi[133].

Il terzo step, nel processo di preprocessing è lo stemming. Per **stemming** si intende la riduzione della forma flessa di una parola alla sua radice, o tema, che non sempre corrisponde alla radice morfologica della parola. Gli algoritmi utilizzati per effettuare lo stemming saranno diversi a seconda della lingua del testo che si desidera analizzare. Gli stemmer consistono in una serie di regole di sostituzione per ridurre le varie forme di un termine ad un'unica radice, per esempio, "avrò", "avremo", "ebbe", verranno tutte sostituite con la radice del verbo origine, "avere"[134].

Così come per la rimozione delle stopwords anche la **normalizzazione** dipende a seconda degli scopi dell'analisi da effettuare. In generale la normalizzazione include la rimozione di errori di battitura e la conversione delle parole in minuscolo[135].

Nella fase di preprocessing di un testo è necessario anche determinare i limiti delle parole. Tale step si effettua tramite alcuni caratteri di punteggiatura (.,;:!?), ma soprattutto tramite gli spazi vuoti. Tuttavia, questo non è valido per tutte le lingue, come ad esempio cinese e giapponese.

Come abbiamo visto nei vari step, la fase di preprocessing dipende fortemente dalla lingua del testo da analizzare, di conseguenza si avranno liste di stopword a seconda della lingua del nostro testo, così come si avranno differenti modi di indentificare i limiti di parole e frasi.

A conclusione della fase di preprocessing vi è la necessità di fornire una rappresentazione al documento da analizzare. Esistono più modi per rappresentare un testo, tutto dipende dall'obiettivo che si cerca di perseguire. In ogni caso per rappresentare un documento o un corpus di documenti si utilizzano fondamentalmente delle matrici. Le due matrici più utilizzate sono[136]:

- la **Document Term Matrix** (o DTM), in tale matrice si hanno per riga i vari documenti del corpus, per colonne le parole, i valori della matrice non sono

[133] PORCU, Valentina. Text mining e sentiment analysis con R. Valentina Porcu, 2016, p.27
[134] Ivi, p.28
[135] Ivi, p.29
[136] PORCU, Valentina. Text mining e sentiment analysis con R. Valentina Porcu, 2016, p.18

altro che le frequenze delle parole nei vari documenti del corpus. La figura qui sotto ne da una rappresentazione:

Figura 17: Docuement Term Matrix[137]

	Parola 1	Parola 2	Parola 3
Documento 1	0	2	1
Documento 2	1	0	0
Documento 3	2	0	0
Documento 4	1	3	0

- la seconda matrice utilizzata per rappresentare un documento è la **Term Document Matrix** (o TDM). Tale matrice non è altro che l'inversa della DTM, nel senso che per riga abbiamo le parole e per colonna i documenti, i valori della matrice sono sempre le frequenze delle parole. Eccone un esempio qui sotto:

Figura 18: Term Document Matrix[138]

	Documento 1	Documento 2	Documento 3
Parola 1	1	0	1
Parola 2	2	0	2
Parola 3	0	1	0
Parola 4	1	2	1

La rappresentazione dei documenti attraverso le matrici DTM e TDM permette, per esempio di clusterizzare i documenti, ossia di vedere qual è la distanza tra i vari documenti in base alla differenza delle parole in esso contenute e di creare dei raggruppamenti di termini in base a quanto sono omogenei tra loro.

Un'altra modalità attraverso cui è possibile rappresentare un documento è la vettorializzazione del testo; tale procedura permette di rappresentare un documento come un vettore, in cui gli elementi del vettore indicano il numero di occorrenze delle parole nel testo. Come si vede il **Vector-Space model**, metodo così chiamato, ha degli elementi di similitudine con la rappresentazione in matrice, tuttavia a differenze delle matrici TDM e DTM, tale metodo permette di determinare quanto un documento è relativo a una determinata query di ricerca. La ricerca sui testi, fondata sull'unità base delle parole, permette alcune delle più comune analisi di Text Mining, come Topic Analysis, Sentiment Analysis, e Word Relation Analysis.

[137] PORCU, Valentina. Text mining e sentiment analysis con R. Valentina Porcu, 2016, p.18
[138] Ivi.

Il vector-space model effettua un'assunzione implicita che risiede nel concetto di bag of words. Il modello della borsa di parole, o bag of words, come già visto è un metodo che permette di rappresentare e analizzare un testo in base alle parole che lo compongono, senza preoccuparsi della parte sintattica. Quindi ogni documento è la somma dei termini che lo compongono, non importa in quale posizione si trovino, l'importante è solo la frequenza. Le frequenze vengono generalmente e convenientemente espresse tramite la funzione di peso TF-IDF, **Term Frequency - Inverse Document Frequency**, che misura l'importanza di un termine rispetto ad un documento o ad un collezione di documenti. La funzione è composta da due fattori:

- il primo termine della funzione è il numero dei termini presenti nel documento (TF). Generalmente tale numero viene suddiviso per la lunghezza del documento stesso, questo per evitare di privilegiare i documenti più lunghi.

$$TF_{i,j} = \frac{n_{i,j}}{|d_j|}$$

  dove $n_{i,j}$ è il numero di occorrenze del termine t_i nel documento d_j , mentre il denominatore è semplicemente la dimensione, espressa in numero di termini, del documento d_j.

- L'altro fattore della funzione indica l'importanza generale del termine nella collezione:

$$IDF_{I,D} = \log \frac{|D|}{|\{d \in D : t_i \in d\}|}$$

  Dove $|D|$, cardinalità di D, è il numero totale di documenti nella collezione, mentre il denominatore $|\{d \in D : t_i \in d\}|$ rappresenta il numero dei documenti dove il termine t_i compare. Infine, il valore finale del termine composto viene calcolato come:

$$TFIDF_{i,j,D} = TF_{i,j} \times IDF_{i,D}$$

In sintesi, TF misura la densità delle parole in un documento, mentre l'IDF misura l'importanza di una parola rispetto al corpus. Le frequenze così calcolate vengono utilizzate nel Vector Space Model (VSM), nello specifico nella misurazione tramite

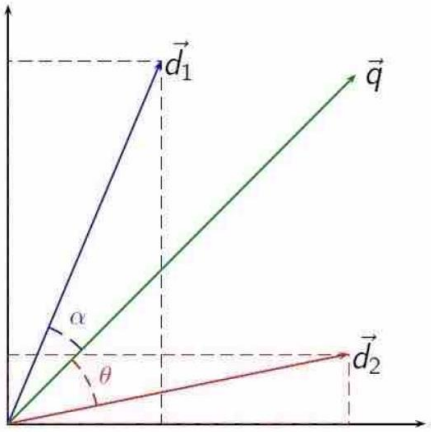
coseno, dove i punteggi TF-IDF dei documenti vengono comparati. Come abbiamo visto, l'idea di base del VSM è che il significato del documento derivi dai termini che lo costituiscono. Inoltre, sappiamo che i documenti e le query vengono rappresentati come vettori di termini:

$$d = (t_1, t_2, t_3, t_4, \dots, t_n)$$

$$q = (t_{1q}, t_{2q}, t_{3q}, t_{4q}, \dots, t_{nq})$$

Dove ogni t_i è un valore che rappresenta l'occorrenza di un termine nel documento d.

Figura 19: Vector Space Model[139]



Il Vector Space Model si basa su tre step:

1. L'indicizzazione di un documento, questo sta a significare che i termini di un documento vengono listati in base alla frequenza. Precedentemente sono state eliminate le parole vuote e le stopwords.
2. Una volta indicizzato il documento, i termini vengono pesati in base alla frequenza.
3. Infine, si vanno a calcolare i coefficienti di similarità.

Vi sono varie misure di similarità da applicare per confrontare un documento e una query:

- il prodotto scalare, si basa sulla distanza Euclidea;

[139] PORCU, Valentina. Text mining e sentiment analysis con R. Valentina Porcu, 2016, p.22

- il coseno, che calcola l'angolo tra i documenti, questa è la misura che generalmente viene utilizzata. L'angolo tra i due vettori rappresenta la similarità tra documenti, e non la distanza;
- le direzioni dei vettori.

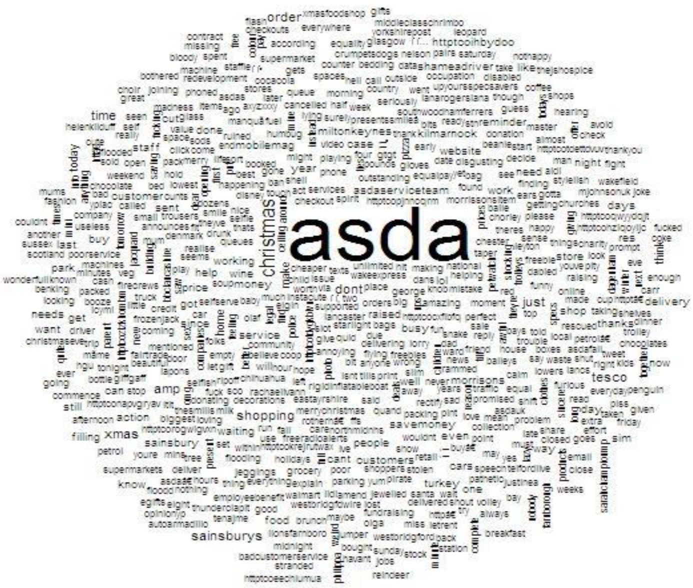
La formula del coseno è così espressa:

$$\cos \alpha = \frac{d_1 q}{||d_1|| \, ||q||}$$

Se viene utilizzato il coseno come metodo per calcolare la similarità tra un documento e una query, i punteggi tf-idf dei documenti devono essere comparati. I limiti principali del VSM, e di conseguenza della rappresentazione a bag of words, hanno a che vedere con la carenza di informazione sintattiche, ad esempio sulla struttura delle frasi, sull'ordine delle parole e sulla loro prossimità[140]. Per avere dei risultati di Text Mining più esaustivi, è necessario rappresentare i documenti attraverso modelli che prendono in considerazione anche gli aspetti semantici del testo. Per esempio, essere in grado di rappresentare un documento in modo da distinguere entità come persone, organizzazioni, località e le relative relazioni intercorrenti è di grande utilità al fine di scoprire patterns più interessanti. Tuttavia, ad oggi, le tecniche di natura language processing non sono abbastanza evolute da permettere rappresentazioni semantiche accurate di testi e o documenti. Per questa ragione molti approcci al Text Mining, si basano su rappresentazioni dei dati basati sul modello a bag of words che, pur comportando una perdita di informazioni circa la posizione dei termini all'interno del documento originario, generalmente sono più flessibili e ben si adattano a varie tipologie di algoritmi. Una volta effettuato il preprocessing e rappresentato i documenti attraverso l'approccio bag of words o la matrice document terms, il passaggio finale consiste nell'applicare le tecniche di data mining come clustering, classificazione, regole di associazione o altre applicazione di Text Minig per scoprire associazioni di termini e pattern nel testo. Infine, è importante poter rendere i risultati dell'analisi di Text Mining in una forma chiara ed efficace, e vi sono strumenti per la visualizzazione dei risultati di Text Mining specifici come word-cloud o tag-cloud.

[140] PORCU, Valentina. Text mining e sentiment analysis con R. Valentina Porcu, 2016, p.18

Fig 20: Word-Cloud di Tweet[141]



In questo paragrafo abbiamo visto le varie fasi del processo di Text Mining, e si può intuire che la parte centrale di tale processo, ossia l'applicazione degli algoritmi, dipenda molto dagli obbiettivi dell'analisi. Al fine di dar risalto ad una di queste applicazioni, nel prossimo paragrafo verrà descritta nello specifico la tecnica di Text Mining nota come Sentiment Analysis, e a conclusione di questa tesi verrà trattato un caso pratico di Sentiment Analysis su dati testuali provenienti da Twitter.

4.3 La Sentiment Analysis: Concetti e Metodologie

La Sentiment Analysis (SA), anche chiamata opinion mining, è il campo di studio che analizza le opinioni, i sentimenti, le valutazioni, le attitudini e le emozioni delle persone verso entità e i loro attributi espressi in un documento scritto. In letteratura si possono trovare molte applicazioni appartenenti al campo della Sentiment Analysisi, quali ad esempio opinion mining, opinion analysis, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, e review mining[142].

---

[141] YOUNIS, Eman MG. *Sentiment analysis and text mining for social media microblogs using open source tools: An empirical study*. International Journal of Computer Applications, 2015, p.46
[142] LIU, Bing. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press, 2015, p.1

Ma cosa si intende realmente per Sentiment Analysis?

All'inizio del paragrafo abbiamo dato una prima definizione, LIU (2015). In letteratura si possono trovare numerose definizioni di Sentiment Analysis, e la maggior parte di queste convergono nel ritenere la Sentiment Analysis il processo per determinare l'atteggiamento o la polarità delle opinioni delle persone in documenti testuali. L'elemento base della SA sono i dati testuali, pertanto gli algoritmi di SA rientrano nella famiglia del Text Mining. Inoltre, la SA è strettamente legata al concetto di Opinion Mining, tecnica che permette di elaborare una ricerca su parole chiave. Inoltre, permette di indentificare per ciascun termine degli attributi (positivo, negativo, neutro) tali per cui, una volta aggregate le distribuzioni di questi termini, diventa possibile estrarre l'opinione associata a ciascun termine chiave[143].

Il fondamento della SA è l'analisi testuale che, a partire da un semplice testo applicando specifici algoritmi, ci permette di estrarre il tipo di sentimento e l'opinione dietro di esso. Ogni tecnica o algoritmo utilizzato per effettuare una Sentiment Analysis, ha come riferimento i quattro principi dell'analisi testuale. Per spiegare tali principi ci affidiamo alla ricerca di Ceron et al. (2016):

1. Ogni modello linguistico quantitativo è sbagliato, ma qualcuno può essere utile. Con questa affermazione si vuol evidenziare che il solo approccio quantitativo mal si adatta ai processi mentali che si utilizzano per la trasposizione delle emozioni in un testo. Questi processi mentali sono ancora poco chiari e tutt'ora vengono studiati, in particolare per quanto riguarda il modo di esprimersi e l'estrema variabilità che c'è nel significato che può assumere un testo. A volte lo spostamento di una parola, l'assenza della punteggiatura, l'uso di espressioni ironiche, o metafore stravolgono il significato di un testo. In sostanza, la complessità del linguaggio è così vasta e in continuo mutamento, con nuove espressioni e neologismi, che difficilmente un testo può essere analizzato solamente con approcci automatici e quantitativi[144].

[143] CERON, Andrea; CURINI, Luigi; IACUS, Stefano Maria. *Social Media e Sentiment Analysis: L'evoluzione dei fenomeni sociali attraverso la Rete*. Springer Science & Business Media, 2014, p.28

[144] CERON, Andrea; CURINI, Luigi; IACUS, Stefano Maria. *Social Media e Sentiment Analysis: L'evoluzione dei fenomeni sociali attraverso la Rete*. Springer Science & Business Media, 2014, p.29

2. I metodi quantitativi aiutano l'uomo, non lo sostituiscono. Alla luce dei limiti dei metodi quantitativi, bisogna però considerare che tali metodi servono moltissimo per velocizzare le operazioni in cui sono presenti grandi quantità di dati, in questo caso testi, che non riuscirebbero ad essere gestite soltanto con le abilità dell'uomo. Possono essere considerati uno strumento che aumenta le capacità umane, ma non certo lo strumento che sostituisce l'uomo[145].
3. Non esiste la tecnica ideale di analisi testuale. È chiaro che ogni algoritmo è concepito sulla base di assunti definiti a priori. Di conseguenza vi sono molti fattori e variabili che portano a preferire una tecnica piuttosto che un'altra. Secondo Ceron i principali fattori che influenzano la scelta dell'algoritmo di SA sono:
    a. la lingua del testo, che sia in italiano o in inglese o in altre lingue straniere;
    b. l'argomento di discussione, il tema di riferimento: politica, economia, sport, etc;
    c. il periodo storico del testo, molte parole hanno un significato diverso alla luce del periodo temporale in cui è stato scritto il testo;
    d. l'età e il genere dell'autore del testo.

    Inoltre, anche il tipo di analisi influenza la scelta dell'algoritmo, vi sono modalità di analisi individuale e aggregata. L'obiettivo dell'analisi individuale è quello di attribuire un testo non ancora letto ad una categoria semantica o ad un autore. L'analisi aggregata invece può essere considerata come l'unione di più analisi individuali al fine di studiare in che modo si distribuiscono le varie categorie semantiche.
4. Validazione dell'analisi. Come abbiamo già avuto modo di vedere, ogni modello deve essere validato dai dati stessi. La validità dell'analisi dipende dalla natura del modello utilizzato. Vi sono due tipologie di metodi che si possono utilizzare[146]:
    a. I metodi supervisionati ossia quei algoritmi in cui le categorie semantiche (o nel nostro caso il sentiment del testo-autore) sono note a priori o vengono identificate tramite codifica manuale su un

[145] Ivi
[146] Ivi

sottoinsieme di testi, detto traning set, sono facilmente validati. Per validare tali metodi basta controllare l'attribuzione semantica generata dal metodo la oggettiva appartenenza semantica del testo leggendo il testo dopo aver effettuato la classificazione.

b. Nei metodi non supervisionati, invece, non si conoscono le categorie semantiche a priori, pertanto è necessario classificare i testi andando a cercare delle ricorrenze all'interno di gruppi di testi classificati come omogenei, oppure l'assegnazione delle categorie avviene attraverso l'uso di dizionari di termini o cataloghi. Per i metodi non supervisionati la validazione può essere un'attività alquanto gravosa. Un sistema che si adotta in questi casi è quello di creare degli esperimenti controllati, esclusivamente per validare la bontà dell'algoritmo. Ad esempio, si vanno a classificare testi di cui in realtà si conosce il contenuto semantico ma dei quali l'algoritmo ignora la classificazione, in questo modo è possibile verificare se l'algoritmo assegna correttamente i documenti al gruppo di effettiva appartenenza.

Dalla discussione sui principi di analisi testuale, emergono due concetti fondamentali per effettuare una qualunque catalogazione delle varie tecniche di analisi testuale: ossia gli algoritmi per estrarre il *sentiment* da testi o documenti (stima del sentiment individuale o aggregato), e la tipologia di algoritmo che si vuole utilizzare, supervisionato o non supervisionato. Qui sono si propone una tabella che sintetizza questa classificazione delle tecniche di analisi testuale:

Fig 22: Word-Cloud di Tweet[147]

		Tecnica di classificazione dei testi	
		unsupervised	supervised
Tecnica di stima della distribuzione delle opinioni	**individuale**	Analisi individuale unita a dizionari ontologici	Analisi individuale unita a codifica manuale senza a priori
	aggregata	Analisi aggregrata unita a dizionari ontologici	Analisi aggregata unita a codifica manuale senza a priori

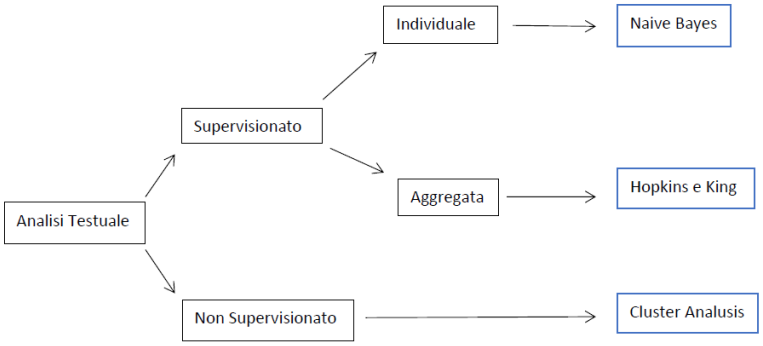
[147] CERON, Andrea; CURINI, Luigi; IACUS, Stefano Maria. *Social Media e Sentiment Analysis: L'evoluzione dei fenomeni sociali attraverso la Rete*. Springer Science & Business Media, 2014, p.31

Come abbiamo già detto per il Test Mining, prima di andare ad utilizzare le varie tecniche di analisi testuale e i vari algoritmi di SA, è necessario preparare i dati testuali attraverso il processo di preprocessing e fornire una struttura tale che possano poi essere tratti dagli algoritmi. I vari step di preprocessing, in sintesi, hanno come obiettivo finale quello di ridurre un testo in un dato quantitativo tale da poter essere trattato da un modello statistico[148].

4.3.1 Metodi di Classificazione dei testi

La Fig. 23 ci mostra come le tecniche di analisi testuale si suddividono a seconda della modalità di estrazione del sentiment (Individuale o Aggregata) e in base al tipo di algoritmo che si intende utilizzare (Supervisionato o Non Supervisionato). Nel nostro caso andremo a vedere tre tecniche di analisi del Sentiment, così catalogate:

Figura 23 : Tecniche di Analisi Testuale



Metodo Naive Bayes

I metodi bayesiani sono metodi supervisionati che si fondano sulla teoria della probabilità e, come si evince dal nome, sul Teorema di Bayes. L'algoritmo di Naive Bayes si basa sulla determinazione della probabilità di un elemento di appartenere a una certa classe[149]. Come detto, la tecnica si basa sul famoso Teorema di Bayes, il quale deriva da tre leggi fondamentali della probabilità:

[148] Ivi.
[149] A. Rezzani, *Big Data Analytics*, Maggioli Editore 2017, p.322

- **Probabilità Condizionata:** La probabilità di A condizionata da B è definita come la probabilità che si verifichi l'evento A a condizione che si verifichi anche l'evento B (naturalmente con P(B) ≠ 0). Enunciamo il teorema in formula.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- **Probabilità Composta:** La probabilità che due eventi A e B si verifichino contemporaneamente è uguale alla probabilità di uno dei due eventi (P(A) o P(B)) moltiplicato con la probabilità dell'altro evento condizionato dal verificarsi del primo (P(A|B) o P(B|A)). Enunciamo il teorema in formula.

$$P(A \cap B) = P(B)P(A|B) = P(A)P(B|A)$$

- **Probabilità Assoluta:** Se A_1,…..,A_n formano una partizione di tutti gli eventi possibili Ω, ossia $A_i \cap A_j = \emptyset$ con $1 \leq i, j \leq n$ e $\bigcup_{i=1}^{N} A_i = \Omega$ e B un qualsiasi evento dipendente dagli eventi A, allora:

$$P(B) = \sum_{i=1}^{N} (A_i)P(B|A_i)$$

Il teorema di Bayes nella sua forma più semplice dice che:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Dati due eventi legati da una relazione di tipo causa-effetto il teorema cerca di rispondere a questa domanda: dato un certo effetto, qual è la probabilità che sia stato prodotto da una certa causa? Il Teorema di Bayes in sostanza ci permette di apprendere dall'esperienza, mettendo insieme più fonti di informazioni allo scopo di effettuare previsioni sui dati. La probabilità che associamo a un evento dipende quindi dalle informazioni a nostra disposizione e quindi dal campione che stiamo analizzando[150].

Essendo un approccio supervisionato si parte da un traning set manuale, in cui a ogni oggetto viene assegnata una categoria (o etichetta) che servirà come apprendimento per l'algoritmo e che in seguito verrà applicato ai dati del test set. I

[150] PORCU, Valentina. Text mining e sentiment analysis con R. Valentina Porcu, 2016, p.22

metodi bayesiani sono inoltre metodi incrementali, pertanto ogni istanza nell'insieme di addestramento (traning set) modifica la probabilità che il classificatore funzioni correttamente[151].

Per ottenere un classificatore Bayesano Naive, bisogna innanzitutto partire da una classe di interesse C e un insieme di caratteristiche (o features) $F_1, \dots\dots, F_n$ di un individuo (o testo nel nostro caso), l'obiettivo è quello di conoscere la probabilità che questo appartiene a C, ossia si vuole conoscere $P(C|F_1, \dots\dots, F_n)$. Se si applica la regola della probabilità composta è possibile ottenere come risultato la forma base del teorema di Bayes:

$$P(C|F_1, \dots\dots, F_n) = \frac{P(C)P(F_1, \dots\dots, F_n|C)}{P(F_1, \dots\dots, F_n)}$$

A questo punto se applichiamo n volte la definizione di probabilità condizionata e se trascuriamo il denominatore, che viene definito constante di normalizzazione, abbiamo che:

$$P(C|F_1, \dots\dots, F_n) = P(C)P(F_1|C)P(F_2|C, F_1) \dots . P(F_n|C, F_1, \dots, F_{n-1})$$

Il generico termine $P(C|F_1, \dots\dots, F_n)$ diventa $P(F_i|C)$ per assunzione *Naive* di indipendenza tra i termini. In conclusione, applicando il teorema della probabilità composto abbiamo:

$$P(C|F_1, \dots\dots, F_n) = P(C)P(F_1|C)P \dots (F_n|C, F_1) = \prod_{i=1}^{n} P(F_i|C)$$

L'indicatore viene indicato come *Naive* perché si basa sull'assunzione secondo cui l'effetto di un attributo (features) su una data classe è indipendente dai valori degli attributi. Tale assunzione, chiamata indipendenza condizionale delle classi ha l'obiettivo di semplificare i calcoli proprio per questo l'algoritmo prende il nome di "Naive"[152]. Si noti che questa assunzione è violata nella maggior parte dei problemi reali, in ogni caso il Naive Bayes si comporta molto bene e risulta essere molto

[151] Ivi

[152] A. Rezzani, *Business Intelligence: Processi, Metodi utilizzo in Azienda,* Maggioli Editore 2012, p.282

efficace. In sintesi, il classificatore, dirà con quale probabilità un testo è associato ad una classe, la classe scelta per il testo in esame, sarà quella che ha avuto la probabilità maggiore.

Metodo Hopkins e King

Il metodo di Hopkins e King è sempre un metodo supervisionato come il classificatore Bayesano Naive, tuttavia si differenzia da quest'ultimo in quanto appartiene alla famiglia di analisi aggregate. L'obiettivo finale del metodo di Hopkins e King è quello di stimare la probabilità aggregata della classificazione di tutti i testi di un corpus, calcolandola direttamente e non utilizzando le classificazioni dei singoli testi per ricavarla. Infatti, uno dei limiti maggiori delle analisi individuali è quello di sommare i singoli errori compiuti nel test set, generando una stima non in linea con i dati reali. Per comprendere il metodo di Hopkis e King utilizzeremo l'approccio di Ceron et al. (2016) che spiega in quale modo avviene la classificazione aggregata delle opinioni, prima utilizzando l'analisi individuale e successivamente con attraverso il metodo aggregato di Hopkis e King.

Si consideri l'insieme $D_j = (D_1, D_2, \dots, D_K)$, con $j = 1, \dots, K$, delle diverse opinioni espresse o delle categorie semantiche già note. Si consideri anche l'insieme delle parole o stem utilizzate in un testo $S_i = (s_1, s_2, \dots, s_M)$ per esprimere un concetto semantico o un'opinione tra quelle del vettore D_j. Il testo ha indice i, con $i = 1 \dots, N$ dove N rappresenta il numero totale di testi nel corpus. Con M invece si indica il numero complessivo di stem distinti nell'intero corpus[153].

Dividiamo a questo punto l'insieme dei testi in due gruppi: il traning set, ossia il sottoinsieme dei testi del corpus che verrà codificato manualmente, e il test set cioè il sottoinsieme dei rimanenti testi nel corpus di cui si vuole predire il contenuto semantico attraverso un algoritmo. Una volta diviso l'insieme dei testi si effettua l'associazione di ogni testo S nel traning set alla corrispondente categoria D_j. Partendo dalla matrice completa degli stem si calcola la frequenza delle volte in cui una certa sequenza di stem o parole $(s_1, s_2, \dots)$ è associata ad una particolare

[153] CERON, Andrea; CURINI, Luigi; IACUS, Stefano Maria. *Social Media e Sentiment Analysis: L'evoluzione dei fenomeni sociali attraverso la Rete*. Springer Science & Business Media, 2014, p.38

categoria D_j. Se tale operazione viene ripetuta per tutte le sequenze **s** associate alla categoria D_j, possiamo stimare la probabilità complessiva $P(D_j|S_i)$[154]:

$$P(D_j|S_i) = P(D_1|s), P(D_2|s), \dots, P(D_k|s)$$

Ad esempio, se prendiamo un testo a caso dal test set con vettore di stem $s = S_i$, il modello così progettato ci dirà che quel testo è associato con probabilità 0,95 alla categoria D_1, con probabilità 0,03 alla categoria D_2, e con probabilità 0,02 alla categoria D_3, in questo caso si ha che K = 3. In questo caso il testo verrà associato alla categoria D_1 che ha ottenuto una percentuale di probabilità più alto, non bisogna dimenticarsi che stiamo parlando di probabilità, pertanto l'assegnazione effettuata non è certa, rimane sempre un margine di incertezza, nel nostro caso il 5%, che non è altro che il potenziale errore di classificazione. In questo esempio abbiamo classificato solo un testo, se immaginiamo di classificare tutti gli N testi del corpus in questo modo, cioè scegliendo di classificare un testo sulla base della categoria che riceve la probabilità più alta, passiamo dalla classificazione individuale a quella aggregata, semplicemente contando quante volte i testi sono classificati in tal modo. Formalmente avremo la seguente matrice[155]:

$$P(D) = P(D|S) * P(S)$$

dove $P(D) = (P(D_1), P(D_2), \dots, P(D_K)$, è un vettore di probabilità di dimensione $K\ x\ 1$ che indica la distribuzione aggregata delle opinione. $P(S)$ è un vettore di probabilità di dimensione $2^M x\ 1$, in quanto si ha che ogni vettore stem $s = (s_1, s_2, \dots, s_M)$ è una combinazione di M variabili binarie (0,1). Infine, $P(D|S)$ è una matrice di dimensioni $K\ x\ 2^M$. Il metodo qui presentato, tuttavia, ha un grosso problema presenta il maggiore errore. Fintantoché si considera la previsione individuale della categoria, con un margine di errore intorno al 3-5%, non vi è alcun problema, ma dal momento in cui le stime effettuate vengono aggregate per ottenere $P(D)$, l'errore di ogni singola previsione viene amplificato arrivando anche attorno al 20%. Per ovviare questo problema Hopkins e King decisero di invertire il

[154] Ivi, p.39

[155] CERON, Andrea; CURINI, Luigi; IACUS, Stefano Maria. *Social Media e Sentiment Analysis: L'evoluzione dei fenomeni sociali attraverso la Rete*. Springer Science & Business Media, 2014, p.39

problema e quindi di concentrarsi sulla distribuzione degli stem $P(S)$, al fine di ottenere direttamente la distribuzione aggregata $P(D)$[156].

$$P(S) = P(S|D) * P(D) \qquad (3.5)$$

si noti che la matrice $P(S|D)$, di dimensione $2^M \; x \; K$, rappresenta la probabilità che una particolare sequenza di stem s di S compaia all'interno dei testi che sono classificati secondo una particolare categoria D_j. Come dicevamo si inverte la visione del problema, l'obiettivo non è più quello di capire il sentiment dalle parole usate, ma identificare le parole che sono usate per esprimere una data opinione o emozione. Il cambiamento di visione ha un importante senso logico: normalmente una persona scrive su un qualche argomento avendo già chiaro in mente quale sia la sua opinione (positiva, negativa o neutra), il suo compito è quello di trovare le parole adatte per esprimere tale opinione[157]. Si osservi però che la distribuzione $P(S|D)$ si conosce solo per i testi del traning set, ossia per quei testi che sono già effettivamente codificati, pertanto è necessario assumere che il linguaggio utilizzato nel traning set (T) per discutere l'opinione D_j sia lo stesso linguaggio utilizzato da tutti i testi del corpus, cioè $P_T(S|D) = P(S|D)$. Alla luce di questa assunzione la formula (3.5) diventa:

$$P(S) = P_T(S|D) * P(D) \qquad (3.6)$$

Tuttavia, lo scopo ultimo è quello di desumere la distribuzione aggregata, pertanto invertendo la formula si ha:

$$P(D) = P_T(S|D)^{-1} * P(S) \qquad (3.7)$$

dove $P_T(S|D)^{-1}$ è una matrice inversa. In questo modo tutti i termini di destra della formula (3.7) possono essere calcolati attraverso il corpus di testi e il traning set. Il vantaggio nel metodo di Hopkis e King è che la stima della distribuzione aggregata $P(D)$ viene stimata direttamente abbattendo il margine di errore, ossia evitando

[156] Ivi

[157] CERON, Andrea; CURINI, Luigi; IACUS, Stefano Maria. *Social Media e Sentiment Analysis: L'evoluzione dei fenomeni sociali attraverso la Rete*. Springer Science & Business Media, 2014, p.39

quell'effetto di propagazione dell'errore tipico delle aggregazioni di tipo individuale[158].

Cluster Analysis

Vediamo infine la Cluster Analysis, che a differenza dei metodi visti precedentemente, è una tecnica di classificazione non supervisionata. In questo metodo pertanto le classi in cui raggruppare i vari testi non si conoscono a priori, ma devono essere dedotti tramite l'individuazione di similarità e dissimilarità tra i testi del corpus. I gruppi o classi vengono quindi a formarsi in base alla possibilità di definire una distanza, o meglio, una misura di dissimilarità tra oggetti che si vogliono classificare. In sostanza i testi del corpus vengono suddivisi in gruppi il più possibile omogenei tra loro secondo questa misura di dissimilarità[159].
La Cluster Analysis ha molti ambiti di applicazione, tuttavia nel nostro caso ci soffermeremo su quanto concerne la classificazione di testi, o documenti testuali. Gli algoritmi di clustering possono essere suddivisi in due gruppi principali:

- Algoritmi di clustering gerarchico;
- Algoritmi di clustering partizionale.

Il metodo gerarchico va ad organizzare i dati in sequenze nidificate di gruppi, che possono essere rappresentati tramite una struttura ad albero. Invece, gli algoritmi di clustering partizionato, determinano la suddivisione dei dati in cluster in modo da ridurre il più possibile la dispersione all'interno del singolo cluster, e viceversa, cercando di aumentare il più possibile la dispersione tra cluster. Non si può parlare in questo caso di metodo gerarchico perché i vari cluster ottenuti sono tutti allo stesso livello di partizionamento. I metodi di clustering partizionali ben si adattano a gestire grosse mole di dati, per le quali invece la costruzione di una struttura gerarchica dei cluster comporterebbe uno sforzo computazionale molto elevato[160].
Vi sono diversi metodi per calcolare la misura di similarità. I più utilizzati sono la distanza Euclidea o la distanza di Manhattan. A prescindere dal tipo distanza adottato, l'obiettivo finale della Cluster Analysis è quello di creare dei gruppi o

[158] Ivi
[159] CERON, Andrea; CURINI, Luigi; IACUS, Stefano Maria. *Social Media e Sentiment Analysis: L'evoluzione dei fenomeni sociali attraverso la Rete*. Springer Science & Business Media, 2014, p.36
[160] A. Rezzani, *Business Intelligence: Processi, Metodi utilizzo in Azienda*, Maggioli Editore 2012, p.285

appunto cluster, omogenei in modo che al loro interno gli oggetti abbiano massima omogeneità, ossia distanza minima, e che tra di loro ci sia massima eterogeneità, ossia distanza massima. L'analisi dei cluster può avvenire in due modalità: o in maniera dissociativa, e quindi la strategia è quella di separare gli elementi che hanno più distanza tra di loro; oppure in maniera agglomerativa, raggruppando gli elementi con distanza maggiore. In questo paragrafo abbiamo descritto alcune delle possibili tecniche di analisi testuale, tuttavia ve ne sono molte altre, come per esempio il *Natural Language Processing* (NPL), l'*Information retrieval* (IR), l'*Information extraction* (IE), il *Topic detection* e la *Text summarization* per ulteriori dettagli si rimanda al lavoro di Ceron et al. (2016).

4.4 Twitter Sentiment Analysis

Twitter è una piattaforma gratuita di microblogging e social network. Nel suo sito si definisce cosi:

"*Twitter is what's happening in the world and what people are talking about right now*[161]."

Twitter è oggi una delle reti sociali più utilizzate nel mondo, dove milioni di utenti ogni giorno condividono e visualizzano le più disparate informazioni, opinioni, news e molto altro. Secondo un'indagine condotta da Web Digitalic[162] di Aprile 2018 Twitter aveva:

- 330 milioni di utenti attivi;
- oltre 500 milioni i tweets che vengono postati ogni giorno;
- 9 i milioni di utenti attivi in Italia;
- 91,1 milione di dollari è l'utile registrato nel primo trimestre del 2018

Questi numeri danno la dimensione del traffico che viene generato da questo Social Network che, pur essendo indietro rispetto ai suoi diretti competitor, Facebook e Istagram, mantiene comunque un'ottima posizione nel mercato.

[161] https://about.twitter.com
[162] https://www.digitalic.it

La Struttura di Twitter è molto semplice. Ogni utente ha la possibilità di "seguire" un gruppo di utenti e di conseguenza essere anche "seguito" (seguire un utente significa ricevere e visualizzare tutti i messaggi che esso pubblica). Pertanto, su Twitter si distinguono due gruppi sociali:

- *Followers*: insieme degli utenti da cui si è seguiti;
- *Following*: insieme degli utenti che si seguono.

Ogni pagina profilo ha una propria *timeline* dove compaiono tutti i tweets pubblicati dai propri following. È chiaro che il servizio principale di Twitter è quello di pubblicare brevi messaggi, nell'ordine di 280 caratteri (fino al 2017 il numero di caratteri consentito per tweet era di 140, in seguito è stato raddoppiato). I tweet possono contenere opinioni, pensieri, immagini, video e link a contenuti multimediali sparsi nella rete. A differenza di altri Social Network, Twitter permette la relazione tra due utenti anche in un solo verso, cioè un utente può seguire un altro utente anche se quest'ultimo non lo segue a sua volta. Twitter è uno strumento di comunicazione unilaterale, in cui vi è qualcuno che scrive e un altro che legge, anche se nel tempo il concetto di conversazione si è andato ad allargare. Infatti, ad oggi è possibile creare un dibattito o una conversazione su un particolare tema aggiungendo il simbolo # anteposto al *topic* del dibattito. È possibile rispondere o nominare un utente attraverso il simbolo @ anteposto al nickname dell'utente. Per questo e altre ragioni Twitter è oggi utilizzato sempre più da aziende, personaggi famosi e politici come canale di comunicazione e interazione con i propri clienti, fans o elettori. Questa sua eterogeneità di contenuti, temi e profili, è uno dei motivi per cui i dati di Twitter ben si adattano ad essere oggetti di Sentiment Analysis. Un altro motivo è il fatto di essere un ambiente "pubblico" senza particolari limitazioni. Infatti, su Twitter è possibile leggere e condividere messaggi anche di utenti che non fanno parte della propria rete di amici. Questo è chiaramente dichiarato in una delle prime righe dell'Informativa sulla privacy di Twitter:

"Twitter è pubblica e i Tweet sono immediatamente visibili e consultabili da parte di chiunque nel mondo.[163]*"*

Per poter recuperare i dati, e quindi i tweets, si deve accedere ai server di Twitter tramite una serie di API (Application Programming Interfaces). Possiamo dire in

[163] https://twitter.com/it/privacy

generale che le API sono il modo in cui i programmi informatici "parlano" tra di loro per richiedere e fornire informazioni. Nascono per la realizzazione di applicazioni in grado di replicare le funzionalità presenti sulla propria pagina web di Twitter, per esempio visualizzare i *tweets* degli utenti che si seguono, visualizzare l'insieme di Followers o Following, ricercare tweet per contenuti, ecc. Pur essendo un ambiente "libero", vi sono comunque delle limitazioni nell'accesso ai dati tramite API, nello specifico vi sono delle limitazioni per scaricare grandi quantità di dati. Le API di Twitter sono integrate in molti linguaggi di programmazione quali per esempio Python o R (che è il software utilizzato per la nostra analisi)[164].

4.4.1 R: strumento di Analisi

Per effettuare la Sentiment Analysis sui dai di Twitter è stato utilizzato il software di analisi dei dati R. R è un software di elaborazione dati che, tramite uno specifico linguaggio di programmazione, consente di eseguire analisi statistiche base ed avanzate. R è un software open source, è disponibile per tutti i principali sistemi operativi e conta su una grande comunità di sviluppatori che, nel corso degli anni, hanno creato numerosi packages che ne estendono le funzionalità, rendendolo uno strumento completo e versatile[165].
In questo paragrafo elencheremo solo i packages utilizzati nell'analisi che verranno descritti nella presentazione dell'analisi. I packages utilizzati sono:

- tm;
- twitterR;
- Openxlsx;
- dplyr;

4.4.2 Sentiment Analysis in Twitter: Implementazione

In questo capitolo vengono descritte le fasi implementazione e messa in funzione del progetto di tesi. Per aiutare a comprendere al meglio i vari step dell'analisi

[164] https://help.twitter.com
[165] A. Rezzani, *Big Data Analytics*, Maggioli Editore 2017, p.243

effettuata, vengono mostrati e descritti i vari *script*, in linguaggio R, utilizzati per lo sviluppo delle varie funzionalità. Il lavoro di progettazione e sviluppo del sistema di Sentiment Analysis si divide in quattro fasi:

1. Filtraggio testuale;
2. Addestramento del sistema (Traning);
3. Creazione di un Classifcatore Bayesiano;
4. Validazione del Classificatore;

Alla luce del compito che il classificatore dovrà svolgere (determinare il sentiment di una notevole mole di tweet) si è utilizzato il metodo di classificazione bayesiano, metodo supervisionato che prevede la creazione di un classificatore a partire da elementi già etichettati. In questo modo l'algoritmo apprende come deve classificare i dati reali. Per la fase di addestramento e di test del classificatore utilizziamo un dataset[166] di circa 600000 tweets già etichettati; di questi 2/3 (400000) sono per la fase di training e 1/3 (200000) per la fase di testing. Ogni riga del dataset contiene il testo del tweet ed il relativo sentimento espresso (0 per sentimento negativo, 1 per sentimento positivo). Nello script qui sotto viene mostrato la fase di caricamento del dataset di traning nell'ambiente R, la separazione del dataset in quattro dataframe così formati:

1. **PositiveTrain**: formato da 200000 tweets positivi (fase di *training*);
2. **NegativeTrain**: formato da 200000 tweets negativi (fase di *training*);
3. **PositiveTest**: formato da 100000 tweets positivi (fase di *test*);
4. **NegativeTest**: formato da 100000 tweets negativi (fase di *test*).

Listato 1.1: Script Caricamento Dataset Traning/Test

```
#Carichiamo il file del Dataset per la fase di Traning e Test

Sentiment.Dataset <-read.csv ("C:/Users/glattanzi/Desktop/Dataset.csv")

#Dividiamo il dataset in due dataframe contenenti i tweet pos e neg
```

[166] http://thinknook.com/twitter-sentiment-analysis-training-corpus-dataset-2012-09-22/

```
data.positive <-subset(Sentiment.Dataset$Text,Sentiment.Dataset$Sentiment == 1)

data.negative <-subset(Sentiment.Dataset$Text,Sentiment.Dataset$Sentiment == 0)

#Creaiamo i quattro dataset per le fasi di traning e test

Positive.Train <-data.frame(data.positive[1:200000])

Negative.Train <-data.frame(data.negative[1:200000])

Positive.Test <-data.frame(data.positive[200001:300000])

Negative.Test <-data.frame(data.negative[200001:300000])
```

Filtraggio:

Una volta caricato e suddiviso il dataset, si passa alla fase di preprocessing per filtrare e pulire il dataset, eliminando tutti quei elementi non utili alla classificazione. Per fare questo si utilizzano tecniche di *detection* tramite espressioni regolari[167] per:

- Ricercare ed eliminare: *Hashstag* (#), *@user* (menzioni di uno specifico utente in un tweet), URL (link a pagine,"http");
- Ricercare e sostituire le *emoticon*. Che possono essere sfruttate, nel nostro caso, come indicatore aggiuntivo del sentiment. Per questo le emoticon che indicano un sentimento positivo vengono sostituite dalla stringa *"Positivo"*, mentre quelle che indicano un sentimento negativo vengono sostituite dalla stringa *"Negativo"*;

Tabelle 4: Alcune delle Emoticons rilevate

Chiave	**Emoticon**
Positive	(: :) :] [: :-) (-: [-: :-] (; ;) ;] [; ;-) (-; [-; ;-] :-D :D :-p :p (=: ;=D :=) :S @-) XD

[167] *"Le espressioni regolari sono dei pattern che permettono di descrivere delle stringhe generiche. Sono molto utili per cercare, sostituire, o estrarre delle stringhe o delle sottostringhe."* PORCU, Valentina. Text mining e sentiment analysis con R. Valentina Porcu, 2016, p.22

Negative	:( ): :-( )-: ;( ); :-[ ]-: ;-( )-; :'[:'( )':]: :[:| :/ |: /: :=(:=| :=[xo :| D: O:

- Ricercare ed eliminare serie di lettere ripetute più volte, come per esempio le lettere ripetute della frase *"I looooove youuuuu"*, che diventa *"I love you"*;
- Ricercare ed eliminare la punteggiatura, i numeri e le parole formate da meno di tre caratteri, che risultano poco significative ai fini dell'analisi;
- Eliminare gli spazi bianchi ripetuti.

Nello script *Listato 1.2* viene mostrato il codice per la fase di preprocessing, che riceve in input il dataset e restituisce come output un nuovo dataset filtrato e ripulito.

Listato 1.2: Script Fase di Preprocessing

```
#Preprocessing

#Funzione FilterDF

FilterDF <- function (arg) {
arg <-apply (arg,2,CleanFunction )
#NOTA : dopo il filtraggio eliminiamo tutti i tweets
#che hanno meno di 3 caratteri ( poco significativi )
arg <-data.frame(text=arg[nchar(as.character(arg))>2])
 return(arg)
}
# La funzione applyFun esegue il filtraggio su ogni singolo tweet

# Funzione CleanFunction per pulire i dataset

CleanFunction <- function(x) {
# Eliminazione degli @user
tmp<-(gsub("@.*?(\\s|$)"," ",x))
# Eliminazione Url ( http ...)
tmp<-gsub ("http[^[:blank:]]+","",tmp)
```

```
# Eliminazione #Tag
tmp<-gsub ("#[^[:blank:]]+"," ",tmp)
# detection emoticon positive
tmp<-gsub(":-)|:)|;)|\\(:|:]|\\[:|\\(-:|\\[-:|:-]|\\(;|;]|\\[;|;-)|\\(-;|\\[-;|;-]|:-D|:D|:
p|:p|\\(=:|;=D|:=)|:S|\\@-)|XD","Positive",tmp)
# detection emoticon negative
tmp<-gsub(":\\(|):|:-\\(|)-:|;\\(|);|:-\\[|]-:|;-\\(|)
;|:'\\[|:'\\(|)':|]:|:\\[|:\\||:/|\\|:|\\/:|:=\\(|:=\\||:=\\[|xo|D:|O:","Negative",tmp)
# Eliminazione lettere ripetute nelle singole stringhe
tmp<gsub("(a|b|c|d|e|f|g|h|i|l|m|n|o|p|q|r|s|t|u|v|z|j|k|w|x|y)\\1{3,}","\\1",t
mp)
# Eliminazione di numeri e metacaratteri
tmp<-removeNumbers(tmp)
tmp<-removePunctuation(tmp)
# Eliminazione spazzi eccedenti
tmp<-gsub ("\\s{2,30}"," ",tmp)
# Trasformo le lettere maiuscole in minuscole
tmp<- tolower (tmp)
return(tmp)
}
```

Un esempio del risultato che si ottiene a seguito della fase di filtraggio:

Tabelle 5: Esempio di filtraggio testuale

Tweet1 Originale	- @codinghorror See, three external monitors on one laptop, http://www.twitpic.com/6e3zp Win7RC 1 x 1920x1080, 2 x 1280x1024
Tweet1 Filtrato	see three external monitors on one laptop
Tweet2 Originale	ARRRRRRGGGH! I cant wait till next weeeeeeeeek! :-)
Tweet2 Filtrato	argh i cant wait till next week positive

Traning:

Come abbiamo visto nel sezione Data Mining, tutti gli algoritmi machine learning necessitano di una fase di addestramento, dove il sistema apprende informazioni e nozioni utili per future classificazioni, in modo da creare una base di conoscenza. Il nostro classificatore bayesiano non è da meno, pertanto una volta che i tweets sono stati filtrati (*Fase di Filtraggio*) vengono create due matrici dette *Term Document Matrix* (*tdm*), una per i tweets positivi l'altra per quelli negativi. Una tdm non è altro che una matrice in cui si hanno per colonna i vari documenti testuali, nel nostro caso i tweet, per riga invece si hanno i singoli termini che compongono il documento. L'elemento E_{ij} della matrice è settato a 1 se il termine ti compare nel documento d_j , 0 in caso contrario (*vedi Listato 1.3*). La tdm ci permette di ottenere la frequenza dei termini all'interno di un set di documenti.

Listato 1.3: Script Esempio Term Document Matrix

```
#Esempio Term Document Matrix

#Prendiamo come esempio i seguenti tweets:

doc1<-"Elon Musk posta su Twitter il modello di SpaceX astronave per viaggi intorno alla Luna, per trasportare le persone da una parte all'altra della Terra e un giorno per raggiungere Marte curiosa la somiglianza con i cartoon del secolo"
doc2<-"Secondo il Wall Street Journal l'iPhone diventerà presto un prodotto obsoleto come la Polaroid e il Walkman."
doc3<-"Io sono per le pari opportunità quindi volevo far sapere ad Amazon che le vostre scatole sono apprezzate non solo da gatti e cani ma anche dalle piccole tartarughe"

#Creazione di un VCorpus e vettorializzazione

Corpus<-c(doc1,doc2,doc3)

VDataFrame<-VCorpus(VectorSource(Corpus))

#Creazione di una TDM

TDM<-TermDocumentMatrix(VDataFrame)

#La term document matrix corrispondente è la seguente:

             Docs
Terms         1 2 3
  all'altra   1 0 0
  alla        1 0 0
  amazon      0 0 1
  anche       0 0 1
  apprezzate  0 0 1
```

```
astronave   1 0 0
cani        0 0 1
cartoon     1 0 0
che         0 0 1
come        0 1 0
con         1 0 0
curiosa     1 0 0
dalle       0 0 1
del         1 0 0
della       1 0 0
diventerà   0 1 0
elon        1 0 0
far         0 0 1
gatti       0 0 1
giorno      1 0 0
intorno     1 0 0
journal     0 1 0
l'iphone    0 1 0
luna,       1 0 0
marte       1 0 0
modello     1 0 0
musk        1 0 0
non         0 0 1
obsoleto    0 1 0
opportunità 0 0 1
pari        0 0 1
parte       1 0 0
per         3 0 1
persone     1 0 0
piccole     0 0 1
polaroid    0 1 0
posta       1 0 0
presto      0 1 0
prodotto    0 1 0
quindi      0 0 1
raggiungere 1 0 0
sapere      0 0 1
scatole     0 0 1
secolo      1 0 0
secondo     0 1 0
solo        0 0 1
somiglianza 1 0 0
sono        0 0 2
spacex      1 0 0
street      0 1 0
tartarughe  0 0 1
terra       1 0 0
trasportare 1 0 0
twitter     1 0 0
una         1 0 0
viaggi      1 0 0
volevo      0 0 1
vostre      0 0 1
walkman.    0 1 0
wall        0 1 0
```

#La matrice ha 3 colonne (una per ogni documento)
 # e 60 righe (una per ogni termine che compare almeno in un documento)

Abbiamo usato la funzione TermDocumentMatrix del pacchetto tm per la creazione della tdm, che prende come input una serie di dati testuali, nel nostro caso tweets, e da come output la matrice tdm. Creando le tdm per i nostri dataset di traning con i tweet positivi e negativi, possiamo ricavare i pesi (frequenza del termine all'interno dei documenti) che ci indicano quanto un termine influenza il sentimento, positivo o negativo, espresso da un tweet, o, in altre parole, quanto il sentimento è condizionato dalla presenza del termine. Il listato 1.4 ci mostra come ottenere una matrice Termini/Frequenza a partire da una tdm.

Listato 1.4: Script Creazione matrice Termini/Frequenza

```
#Creazione Matrice Termini Frequenza
# La funzione TermFreq prende in input la TDM

TermFreq <- function(TermMatrix){
 Matrix<-TermMatrix
 Termini <-data.frame(Termini= Terms(Matrix))
 Termini [ ,2] <-apply (Termini ,1, applyFreq )
 return (Termini)
}

#La funzione applyFreq esegue la sommatoria di ogni
# riga della tdm , in modo da ottenere la frequenza
# assoluta di apparizione per ogni termine

applyFreq <- function ( termine ){
 val <-as.numeric(tm_term_score(matr,termine,FUN = slam::row_sums))
 return (val)
}
```

Si noti che nella funzione applyFreq viene utilizzata la funzione tm_term_score del pacchetto tm che, permette di calcolare la frequenza di un termine nei documenti all'interno di una tdm. A questo punto, pertanto abbiamo tutti gli strumenti necessari per avviare la fase di Traning. Nel listato 1.5 mostreremo lo script che riassume tutte le operazioni necessarie per l'addestramento.

Listato 1.5: Script Traning

```
#Caricamento Dataset
#Carichiamo il file del Dataset per la fase di Traning e Test

Sentiment.Analysis.Dataset<-
read.csv("C:/Users/glattanzi/Desktop/MHT/Altro/Gio_BATA/Tesi/BigDataHadoop/Se
ntiment Analysis Dataset.csv")

#Dividiamo il dataset in due dataframe contenenti i tweet pos e neg

data.positive<-
data.frame(Text=subset(Sentiment.Analysis.Dataset$SentimentText,Sentiment.Analys
is.Dataset$Sentiment == 1))
data.negative<-
data.frame(Text=subset(Sentiment.Analysis.Dataset$SentimentText,Sentiment.Analys
is.Dataset$Sentiment == 0))

#Creiamo i quattro dataset per le fasi di traning e test

Positive.Train <-data.frame(Text=data.positive[1:200000,1])
Negative.Train <-data.frame(Text=data.negative[1:200000,1])
Positive.Test <-data.frame(Text=data.positive[200001:300000,1])
Negative.Test <-data.frame(Text=data.negative[200001:300000,1])

# Applichiamo la funzione di Filtraggio

Positive.Train.Clean<-FilterDF(Positive.Train)
Negative.Train.Clean<-FilterDF(Negative.Train)

#Creazione di un VCorpus e vettorializzazione

VCorpus.Positive.Train<-VCorpus(VectorSource(Positive.Train.Clean))
VCorpus.Negative.Train<-VCorpus(VectorSource(Negative.Train.Clean))
```

#Creazione delle TDM

```
Positive.Train.TDM<-
TermDocumentMatrix(VCorpus.Positive.Train,control=list(wordLengths=c(3,9),remove
Numbers=TRUE,stopwords=stopwords(kind="en"),removePunctuation=list(preserve_i
ntra_word_dashes=TRUE)))

Negative.Train.TDM<-
TermDocumentMatrix(VCorpus.Negative.Train,control=list(wordLengths=c(3,9),remov
eNumbers=TRUE,stopwords=stopwords(kind="en"),removePunctuation=list(preserve_
intra_word_dashes=TRUE)))
```

#Applichiamo la funzione per la creazione delle Matrici Termini/Frequenza

```
Positivi.TermFreq.Train<-TermFreq(Positive.Train.TDM)
Negative.TermFreq.Train<-TermFreq(Negative.Train.TDM)
```

I data frame Positive.TermFreq.Train e Negative.TermFreq.Train rappresentano la base di conoscenza del classificatore bayesiano che stiamo andando a definire. Tramite questi dataset siamo in grado di determinare quanto un termine influenza il sentimento generale espresso da un tweet. Nel listato qui di seguito (Listato 1.6), mostriamo alcune informazioni significative relative alla fase di traning.

Listato 1.6: Script Info Traning

#Tdm tweets positivi

```
> Positive.Train.TDM
<<TermDocumentMatrix (terms: 75080, documents: 1)>>
Non-/sparse entries: 75080/0
Sparsity: 0%
Maximal term length: 9
Weighting: term frequency (tf)
```

#Tdm tweets negativi

```
> Negative.Train.TDM
<<TermDocumentMatrix (terms: 72998, documents: 1)>>
Non-/sparse entries: 72998/0
Sparsity : 0%
```

Maximal term length: 9
Weighting : term frequency (tf)

#Come si nota la base di conoscenza è composta da 75080

 # termini positivi e da 72998 termini negativi

Classificazione

Con classificazione intendiamo l'effettiva implementazione di un classificatore Bayesiano. Nel dettaglio ogni tweet verrà scomposto nei termini (token) di cui è costituito e per ognuno di essi si valuterà la probabilità che il tweet a cui il termine appartiene, esprima un sentimento positivo o negativo. Riepilogando, ogni termine ci fornisce due informazioni:

1. La probabilità che il tweet di appartenenza abbia un sentimento positivo;
2. La probabilità che il tweet di appartenenza abbia un sentimento negativo;

Tabella 6: Frequenza di apparizione di alcuni termini

Termine	**Frequenza Termini di Apparizioni**	
	Tw Positivi	**Tw Negativi**
"like"	10728	10624
" worst"	101	522
" positive"	3017	1836
"negative"	749	1216
" world"	1188	749
" market"	105	92
"iphone"	466	1203

Per sempio, se nella fase di training la parola "iphone" compare in 466 tweet positivi e in 1203 tweet negativi, un generico tweet che contiene la parola "iphone" avrà maggior probabilità di risultare negativo durante la classificazione; la tabella

6 mostra alcuni esempi di termini con relative frequenze di apparizione nei 200000 tweets positivi e nei 200000 tweets negativi, utilizzati nella fase di training. Nel listato 1.7[168] vediamo lo script di classificazione, mediante l'uso delle matrici di frequenza.

Listato 1.7: Script Classificazione

```
#Classificazione

#Calcolo Probabilità di avere un elemento positivo o negativo nel tot del traning set

PrPositive<-percent(DimPositive[1,]/(DimPositive[1,]+DimNegative[1,]))
PrNegative<-percent(DimNegative[1,]/(DimPositive[1,]+DimNegative[1,]))

#La funzione Classification prende in imput i tweets

Classification<-function(tweets){
 # Calcolo Tempo di Esecuzione
 ptm<-proc.time()
 print("...Aspetta classificazione in corso…")
 matr<-apply (tweets,1,ApplyClass)

 Results<-(data.frame(tweets,matr,TempoTrascorso=signif(((proc.time()-ptm)/60)[3],7
),row.names = NULL))
 colnames(Results)[1]<-"Tweets"
 colnames(Results)[2]<-"Sentiment"
 return(Results)
}

#La funzione ApplyClass prende in input un singolo tweet e lo va a tokenizzare

ApplyClass<-function(tweet){
#Tokenizzazione dei tweet
tok<-MC_tokenizer(tweet)
Token<-data.frame(tok)
numberToken<-length(tok)

 if(numberToken>1){
 matri<-apply(Token,1,applyToken)

 #Freq tweet positivo
 ProbP <-prod(matri[1,])

 #Freq tweet negativo
 ProbN <-prod(matri[2,])

 if(ProbP>ProbN)
```

[168] Lo script per script di classificazione si basa alcune metodologie apprese nel corso online di Udemy "Text Mining e Sentiment Analysis con R", di Valentina Porcu. Lo script tuttavia non è la copia del codice visto durante il corso, ma è stato da me più volte modificato al fine di meglio adattarlo alla natura dei dati e del processo.

```
  Sent<-1

 if(ProbN>ProbP)
  Sent<-(-1)

 if(ProbP==ProbN)
  Sent<-0

 return(c(Sent))
 }
ProbP=0
ProbN=0
Sent=0
return(c(Sent))
}

#La funzione applyToken prende in imput il singolo token
#e determina la prob che il tok appartenga ad un tweet positivo
#o negativo nella base di conoscenza (Tranin Set)
applyToken <- function (tok){

retN=1;
retP=1;

# Prob che il termine sia in un tweet positivo
posP<-which(PrPosTerm[,1]==tok)
if(length(posP)!=0)
retP<-PrPosTerm[posP,2]

# Prob che il termine sia in un tweet negativo
posN<-which(PrNegTerm[,1]==tok)
if(length(posN)!=0)
retN<-PrNegTerm[posN,2]

return(c(retP,retN))
}

#Il risultato finale della Classificazione sarà un data
#Frame con i singoli tweet e il valore 1 se il tweet è positivo
#-1 se negativo e 0 se è neutro
```

Validazione

Le fasi di validazione e valutazione servono a determinare la bontà e precisione del nostro modello di analisi, andando a calcolare l'accuratezza e l'errore del classificatore. Il test di validazione avviene tramite l'incremento graduale della dimensione del traning set e il calcolo delle relative prestazioni ad ogni passo. Le misure di accuratezza ed errore sono date dalle seguenti formule:

$$\hat{A} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\hat{E} = \frac{FP + FN}{TP + TN + FP + FN} = 1 - \hat{A}$$

Dove si ha che:

- FP (Falsi Positivi): numero di tweet classificati non correttamente nella classe Positivi;
- TP (Veri Positivi): numero di tweet classificati correttamente nella classe Positivi;
- FN (Falsi Negativi): numero di tweet classificati erroneamente nella classe Negativi;
- TN (Veri Negativi): numero di tweet classificati correttamente nella classe Negativi.

Abbiamo quindi calcolato tutti i valori sopra esposti e misurato i valori di Accuratezza e Errore, che sono riassunti in tabella 7. Nel dettaglio per ogni traning set vengono mostrati i valori di accurency e error delle classi Positive e Negative, e il totale del sistema.

Tabella 7:Validation Test

Tweet	**Ac. Pos**	**Ac. Neg**	**Er. Pos**	**Er. Neg**	**Accurrency**	**Error**
2000	0.633	0.838	0.367	0.162	0.736	0.264
5000	0.615	0.818	0.385	0.182	0.717	0.283
10000	0.626	0.812	0.374	0.188	0.719	0.281
20000	0.645	0.792	0.355	0.208	0.718	0.282
40000	0.653	0.782	0.347	0.218	0.718	0.282
80000	0.660	0.778	0.340	0.222	0.719	0.281
120000	0.659	0.801	0.341	0.199	0.730	0.270
160000	0.643	0.817	0.357	0.183	0.730	0.270

180000	0.638	0.820	0.362	0.180	0.729	0.271
200000	0.632	0.825	0.368	0.175	0.729	0.271

Se inseriamo i dati del Validation Test all'interno di un grafico possiamo notare che all'aumentare del volume del Traning set di riferimento i valori di accurency e error si stabilizzano verso un valore soglia.

Figura 24:Validation Andamento Accurency

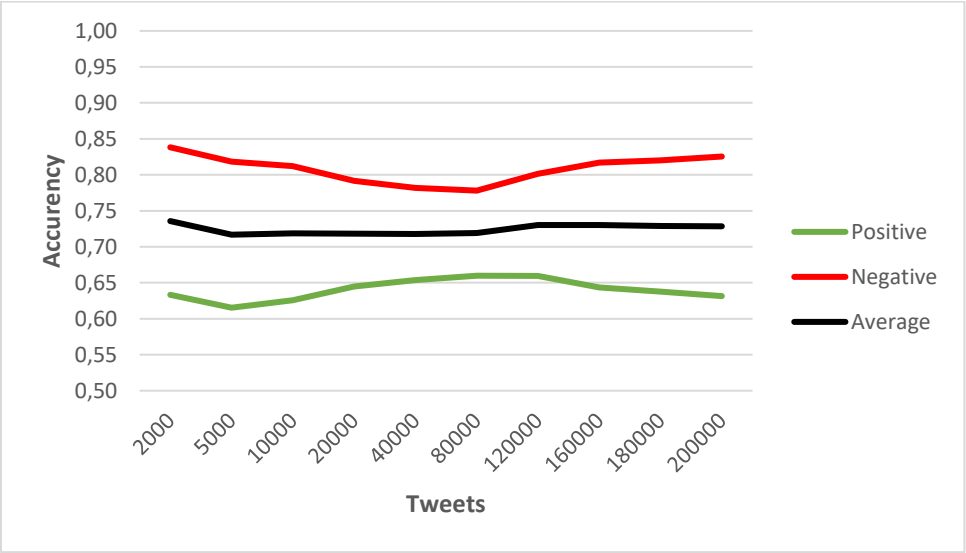
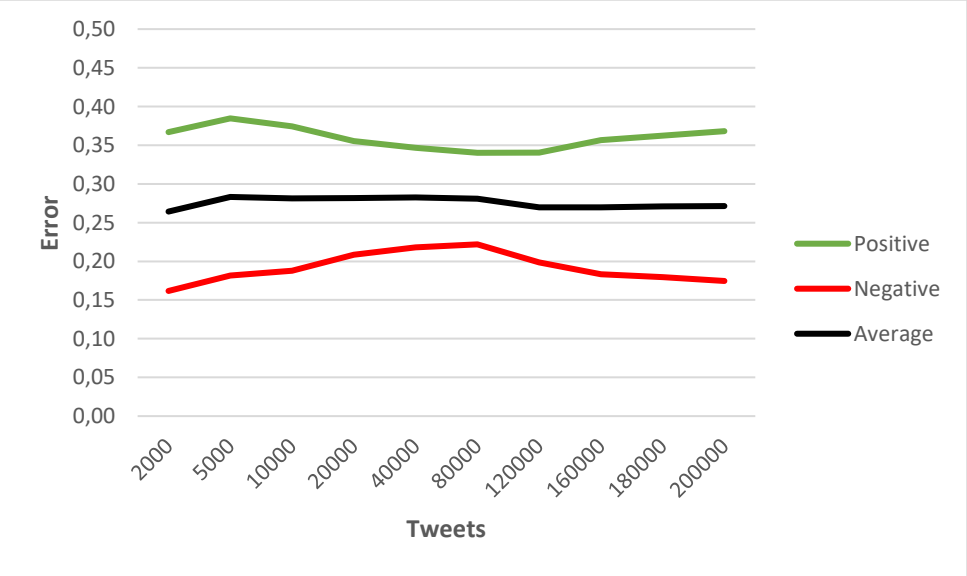


Figura 25:Validation Andamento Error



Dai grafici notiamo anche che il nostro modello ha un riscontro migliore verso gli elementi Negativi, infatti abbiamo un livello di accurency pari al 82% e un errore

che si attesta al 18%. Invece per quanto concerne gli elementi Positivi abbiamo un livello di accurency al 63% e un errore al 37%. Questo vuol dire che il nostro classificatore è "leggermente pessimista", ossia che difronte ad un risultato negativo abbiamo maggiore certezza che esso corrisponda al sentimento realmente espresso.

Finita la fase di validazione, non ci resta che effettuare un test di Valutazione, per verificare ancora una volta la bontà del nostro classificatore. Il test di valutazione viene fatto sul Test set che è composto da 100000 tweets positivi e altrettanti tweets negativi. È importante che il test set sia composto da elementi differenti rispetto al traning set da cui si è creato il modello.

In tabella 8 si possono vedere i risultati del test di classificazione per i tweet positivi, nella tabella 9 invece si hanno i risultati dei tweet negativi. Osservando i risultati in tabella, si nota che il processo di classificazione è molto oneroso per classificare 100000 tweets servono circa 6 ore. Infine, nella tabella 10 si hanno i dati di accurency e error del test di classificazione.

Tabella 8: Test Classificazione tweets Positivi

	Positivi	**Negativi**	**Neutrali**	**Tempo di Classificazione (minuti)**
abs.	59353	38095	2404	390
%	0.59	0.38	0.02	

Tabella 9: Test Classificazione tweets Negativi

	Positivi	**Negativi**	**Neutrali**	**Tempo di Classificazione (minuti)**
abs.	14075	84473	1324	360
%	0.14	0.85	0.01	

Tabella10: Prestazioni classificatore

	Accuratezza	**Errore**
%	0.72	0.28

4.4.3 Il sistema ed il Mondo Reale: Sentiment in Azienda

Abbandoniamo il mondo ovattato e irreale della fase dei test dove le informazioni (*tweets*) che ci servivano erano contenute in un *dataset* precostruito e pre-classificato, e proviamo a proiettare il nostro sistema nel mondo reale in cui le informazioni che vogliamo provare a classificare, ci vengono date direttamente dai server di Twitter. Il lavoro di progettazione e sviluppo del sistema di Sentiment Analysis su dati Twitter si articola in tre fasi:

1- Estrazione dei dati;
2- Architettura del Sistema di Sentiment Analysis;
3- Conclusione e Valutazione dei Risultati.

Estrazione dei Dati:

L'estrazione dei dati da Twitter è avvenuta utilizzando il packages twitterR e nello specifico la funzione *searchTwitter*:

Listato 1.8: Script Estrazione dati Twitter

```
#Programma Tesi Estrazione dati Twitter
#Autore: Giovanni Lattanzi

#Begin Script

#Connessione a Twitter
api_key <-"S0SThSjrzLazZOcY1udk2boVX"
api_secret <-"gA937I8ALbMLrAR92v5mU3aEjHzeDBOt0jlHvZRB6ZX3gCtIpJ"
access_token <-"3053583430-6Qrodk9oe5WlGUBZkISxCWiSQrGn0s5KWFcMwHf"
access_secret <-"BZHz6gmameuRQmI2mfNrT1hnOlr0460KR5idPrPhMG0e0"

#richiamiamo i pacchetti


library(twitteR)
library(openssl)
library(httpuv)
library(base64enc)
```

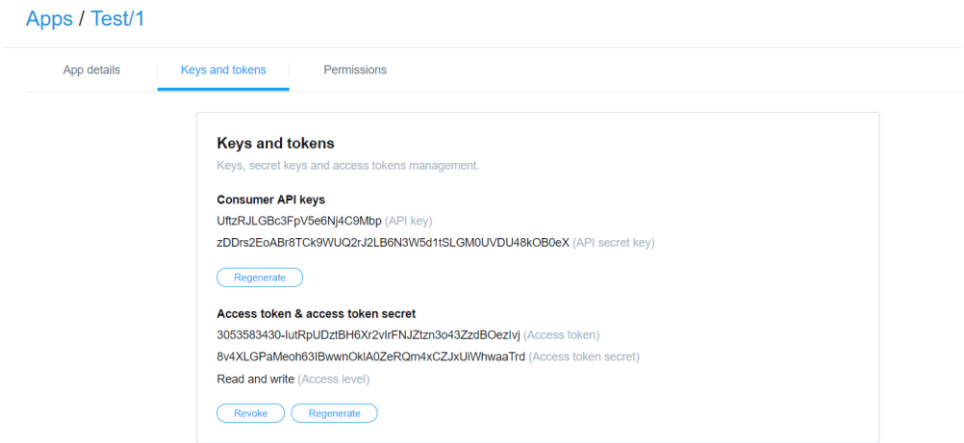
```
#Impostiamo l'accesso ai Server di Twitter (funzione setup_twitter_oauth)


origop <- options("httr_oauth_cache")
options(httr_oauth_cache = TRUE)
setup_twitter_oauth(api_key,api_secret,access_token,access_secret)


#Parole di ricerca

Words<-
c('#Apple','#iPhone','#iPad','#TimCook','#ElonMusk','#Tesla','#Model3','#ModelX','#A
mazon','#AmazonPrime','#AmazonPrimeVideo','#JeffBezos')


#Ricerca

#Funzione ResearchTweet(Con ciclo per applicare ad ogni wordkey la funzione Search)


a=0

ResearchTweet<-function(word){
for(word in Words){
print("...Aspetta ricerca in corso…")
Res<-SearchTweet(word)
Results<-data.frame(Res,word)

a<-a+1

 if(a==1)
 df1<-Results
 if(a==2)
 df2<-rbind.data.frame(df1,Results)
 if(a==3)
 df3<-rbind.data.frame(df1,df2,Results)
 if(a==4)
 df4<-rbind.data.frame(df1,df2,df3,Results)
 if(a==5)
 df5<-rbind.data.frame(df1,df2,df3,df4,Results)
 if(a==6)
 df6<-rbind.data.frame(df1,df2,df3,df4,df5,Results)

 if(a==7)
 df7<-rbind.data.frame(df1,df2,df3,df4,df5,df6,Results)
 if(a==8)
 df8<-rbind.data.frame(df1,df2,df3,df4,df5,df6,df7,Results)
 if(a==9)
 df9<-rbind.data.frame(df1,df2,df3,df4,df5,df6,df7,df8,Results)
 if(a==10)
 df10<-rbind.data.frame(df1,df2,df3,df4,df5,df6,df7,df8,df9,Results)
 if(a==11)
 df11<-rbind.data.frame(df1,df2,df3,df4,df5,df6,df7,df8,df9,df10,Results)
 if(a==12)
```

```
 df12<-rbind.data.frame(df1,df2,df3,df4,df5,df6,df7,df8,df9,df10,df11,Results)
}
DfFinale<-df12[!duplicated(df12$id),]
return(DfFinale)
}
```

```
#Ricerca Tweet per specifica wordKey

SearchTweet<-function(x){
Search<- twListToDF(searchTwitter(x,n=10000,since='2019-02-05',until ='2019-02-06'
,lang="en"))
df<-subset(Search,isRetweet=='FALSE')
Results<-df
return(Results)
}

#Applichiamo la funzione alle parole chiave

TweetScaricati<-ResearchTweet(Words)
```

L'estrazione dei dati tramite API richiede autenticazione tramite il protocollo OAuth[169]. Una volta registrata la propria App sul sito https://developer.twitter.com, vengono fornite dei token per l'accesso:

Figura 26: Keys e Tokens dell'Apps di Twitter[170]



I token di accesso vengono caricati in R e poi passati alla funzione setup_twitter_oauth per accedere a Twitter.

Una volta effettuato l'accesso, per scaricare i tweets basta utilizzare la funzione searchTwitter. Per facilitare la ricerca la funzione searchTwitter dispone di una serie

[169]https://dev.twitter.com/docs/auth/obtaining-access-tokens
[170] https://dev.twitter.com/docs/auth/obtaining-access-tokens

di parametri attraverso cui è possibile settare la ricerca in maniera precisa, i parametri disponibili sono:

- *searchString*: la striga contente l'argomento di ricerca, che può essere una singola parola, una frase o un hashtag;
- *n*: il numero massimo di tweet da estrarre;
- *lang*: limita l'estrazione dei tweet nella lingua specificata;
- *since*: limita i tweet a quelli dalla data specificata. La data deve essere formattata come YYYY-MM-DD;
- *until*: limita i tweet a quelli fino alla data specificata. La data deve essere formattata come YYYY-MM-DD.

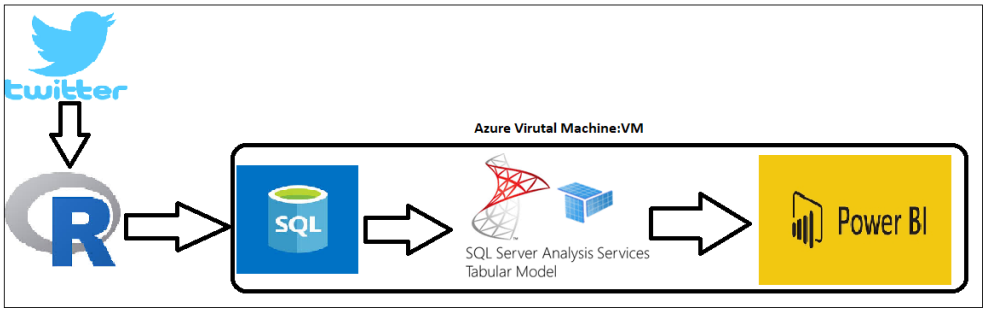
Questi sono i parametri utilizzati nel progetto di Tesi, tuttavia la funzione serchTwitter dispone anche di altri parametri, quali per esempio *geocode* che permette di estrarre i tweet di utenti ubicati in una specifica città (in questo parametro è necessario inserire le coordinate di latitudine e longitudine). Le funzione *ResearchTweet* e *SearchTweets* (che al suo interno sfrutta la funzione serchTwitter del pacchetto di TwitteR) da noi create, ci permettono di eseguire la ricerca dinamicamente nel tempo. Inoltre, la ricerca viene eseguita a partire da una lista di parole chiave, hashtag nel nostro caso, questo ci permette di effettuare una ricerca multipla ogni volta che eseguiamo il programma. L'estrazione dei dati avviene ogni giorno e si riferisce ai dati, tweet, relativi al giorno precedente.
È importante sottolineare, che sono stati estratti solo tweet, non sono stati presi in considerazione eventuali retweet estratti. Anzi proprio per eliminare eventuali retweet dal dataset è stato eseguito un subsetting in modo escludere tutti i retweet. Si vanno a escludere i retweet per una ragione molto semplice: scopo dell'analisi è quello di determinare il Sentiment espresso nei tweet, di conseguenza l'oggetto di analisi è il testo, avere anche i retweet all'interno del dataset vorrebbe dire calcolare due volte il sentiment dello stesso tweet. In fase di aggregazione dei dati, per eventuali valutazioni, la presenza del sentiment dei retweet falserebbero i risultati.

Architettura del Sistema di Sentiment Analysis:

Il progetto di Tesi non si limita a costruire un modello di analisi e ad applicarlo su dati reali, quello che si prefigge è la creazione di un Sistema di BI automatico che

estrae, analizza e costruisce report di sintesi giornalmente. Per fare questo ci siamo affidati a strumenti di Business Intelligence di Microsoft[171]: Integration Service, come strumento ETL, Analysis Service, come strumento per la costruzione di database multidimensionali e tabulari, e Power BI, come strumento di reportistica finale. Inoltre, sia per ragioni di spazio fisico, archiviazione dati, sia per ragioni di automazione, tutto il sistema è stato definito all'interno di una macchina virtuale AZURE. In fig. 27 è rappresentato uno schema dell'architettura del sistema di Sentiment Analysis:

Figura 27: Architettura Sistema



Come si vede dallo schema in figura 4: i dati vengono scaricati, tramite lo script R di estrazione, da Twitter e vengono storicizzati all'interno di un database (Microsoft Sql Server). La storicizzazione dei dati viene eseguita per due motivi principali. Innanzitutto, bisogna sottolineare che l'estrazione dei tweet tramite le API di twitter, ci permette di estrarre solo una parte dei tweet realmente presenti, inoltre possiamo risalire ai dati di 10 giorni precedenti dalla data corrente. Quest'ultima limitazione comporta la necessita di mantenere comunque uno storico dei dati estratti, così in caso di errori durante il filtraggio o la classificazione è sempre possibile rielaborare il tutto. Il secondo motivo per tenere uno storico dei dati è dato dalla necessità di analisi (che comunque richiede di avere il sentiment delle aziende giorno per giorno), in particolare nella necessità di rielaborare la base di conoscenza del classificatore[172].

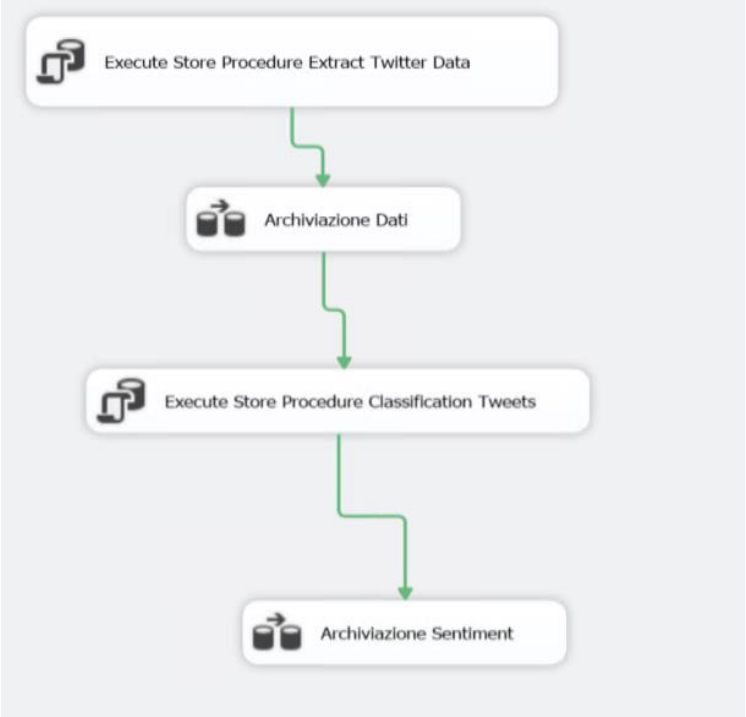
Il flusso di dati da R al database Sql Microsoft viene eseguito attraverso lo strumento ETL di Microsoft Integration Service (SSIS). La creazione di un flusso

[171] Si è prediletto l'uso di strumenti Microsoft per una ragione di convenienza: utilizzandoli a lavoro era per me la scelta più ragionevole per ottimizzare il tempo.

[172] Il nostro classificatore è stato costruito sulla base di un dataset precostituito e classificato, in questo modo abbiamo evitato di effettuare un tagging manuale, che costa molto tempo. Tuttavia, qualora il classificatore non dà risultati soddisfacenti o poco coerenti, può essere necessario effettuare un tagging manuale, di conseguenza tenere uno storico dei dati è fondamentale.

di caricamento oltre a trasferire dati da R a Sql Server ci permette di effettuare delle trasformazioni nei dati e di automatizzare l'estrazione e l'archiviazione dei dati. Questo avviene grazie alla possibilità di eseguire degli script di R all'interno delle store procedure di Sql[173]. Inoltre, la creazione di un pacchetto SSIS ci dà la possibilità di schedulare un Job giornaliero, che lancia l'esecuzione del flusso e quindi l'estrazione e classificazione dei tweet. In questo modo il sistema di estrazione ed analisi dei dati avviene dinamicamente giorno per giorno, è da sottolineare che i dati estratti e analizzati in un dato giorno si riferiscono al giorno precedente. La figura qua sotto mostra il flusso del pacchetto SSIS:

Figura 28: Flusso Dati



Come si vede in figura 5 il flusso dati è composto di quattro step, in cui vengono utilizzati strumenti di SSIS per la creazione del flusso:

1- Il primo step del flusso dati utilizza l'oggetto Execute SQL Task per lanciare la store procedure per l'estrazione dei dati;
2- Il secondo step invece usa Data Flow Task, che fornitogli una base dati di input e una base dati di output, esegue il trasferimento dei dati da input a output. In questo modo archiviamo lo storico dei dati estratti da Twitter;

[173] Nello specifico,

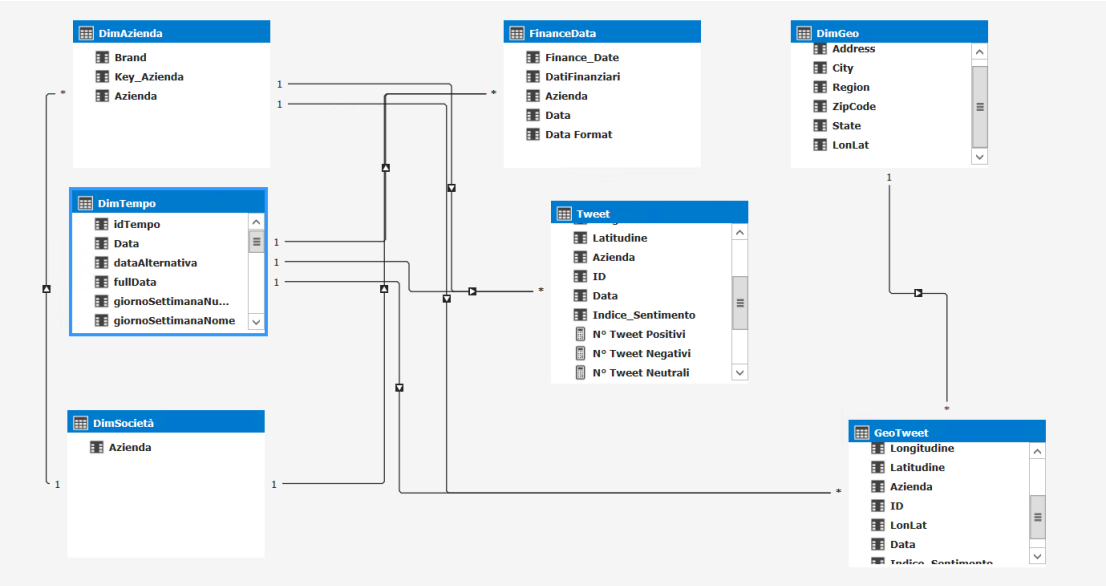
3- Il terzo step lancia la store procedure per la classificazione dei dati. Ovviamente la classificazione non viene ripetuta giornalmente su tutto lo storico dati, ma solo sugli ultimi dati estratti. Questo è molto importante, data la mole di tempo necessaria a classificare i dati.
4- L'ultimo step non fa altro che archiviare i dati di Sentiment ottenuti tramite la classificazione.

Quindi in conclusione il pacchetto SSIS è un ottima soluzione per automatizzare il caricamento e analisi dei dati. Il risultato finale è la creazione di un database, denominato TwitterDB, in cui sono archiviati sia tutti i tweet estratti giornalmente sia i dati di sentiment calcolati.

Una volta costituito il database TwitterDB con i nostri dati di analisi, si è costruito un database Tabulare attraverso il software Analysis Service. I motivi per cui aggiungere questo step all'interno dell'architettura sono primariamente due:

1. Prima di tutto, un ambiente tabulare ci permette di creare il modello dati su cui basare la reportistica finale. Inoltre, possiamo creare misure di sintesi (KPI) ed effettuare trasformazioni nelle singole tabelle. In figura si vede il modello dati con i legami tra le tabelle:

Figura 29: Modello Dati



come si evince dalla figura il modello dati si caratterizza da 3 tabelle dei fatti e 3 tabelle di dimensioni. Le tabelle dei fatti sono:

- TweetData: in cui sono presenti tutti i tweet con il sentiment corrispondente;
- GeoTweet: in cui sono presenti solo i tweet che si possono geolocalizzare, ossia che hanno valorizzato i campi latitudine e longitudine;
- DatiFinanziari: in questa tabella invece ci sono i dati dei titoli azionari delle azien
- de per cui si è ricercato il sentiment (questi valori ci serviranno da Benchmark, come metro di confronto con i valori del sentiment).

Le Dimensioni invece sono semplicemente:

- DimAzienda: in questa tabella sono presenti i nomi delle aziende per cui si è ricercato il sentiment. In particolare, vengono associati i vari keywords usati per ricercare i tweet alle aziende di riferimento (per esempio: l'azienda Apple avrà associato le keywords: #Apple,#TimCok,#Iphone,#Ipad);
- DimTempo: questa è la dimensione temporale, sono presenti informazioni temporali con diverse granularità: anno, mese, data. Questo ci permette di effettuare analisi a differenti livelli temporali;
- DimGeo: in questa dimensione invece sono presente informazioni geografiche più dettagliate, quindi i valori di latitudine e longitudine vengono tradotti in città, regione, stato. Questo ci permette di ottenere una distribuzione geografica a livello geografico.

2. Il secondo motivo per cui viene sfruttato il modello tabulare, è semplice e intuitivo. Il sistema creato va a caricare nuovi dati ogni giorno: questo comporta un aumento di dimensione del database. Il modello tabulare ci permette di sfruttare un particolare tipo di collegamento con Power BI per la creazione dei report. Il collegamento è chiamato "Direct Query" e, sostanzialmente, va in lettura sui dati del modello tabulare e questo non comporta il caricamento di nessun dato. In questo modo i report sono file molto 'leggeri' con performance di caricamento ottime che non cambiano all'aumentare del numero dei dati.

Conclusione e Valutazione dei Risultati:

Come abbiamo accennato nel precedente paragrafo, i dati di Sentiment hanno poco significato se valutati singolarmente, è necessario avere dei dati con cui confrontarli. Nel nostro caso dovendo confrontare i dati di sentiment di aziende, abbiamo preso come benchmark l'andamento dei titoli azionari. L'idea è quella di capire se vi è una qualche relazione tra il sentimento della gente comune verso le aziende (espresso nei Social Network), e l'andamento aziendale. In sostanza cerchiamo di rispondere a questa domanda: se il sentimento comune verso l'azienda è positivo, l'azienda avrà una buona quotazione in borsa? Cerchiamo di capire se c'è una correlazione tra il pensiero comune della gente e l'andamento in borsa.
Per ottenere i dati finanziari delle aziende ci siamo affidati al sito Yahoo Finanza, nello specifico abbiamo utilizzato il package di R **pdfetch**, che ci permette di connetterci al sito ed estrarre lo storico dei dati. Lo script di estrazione dei dati finanziari è il seguente:

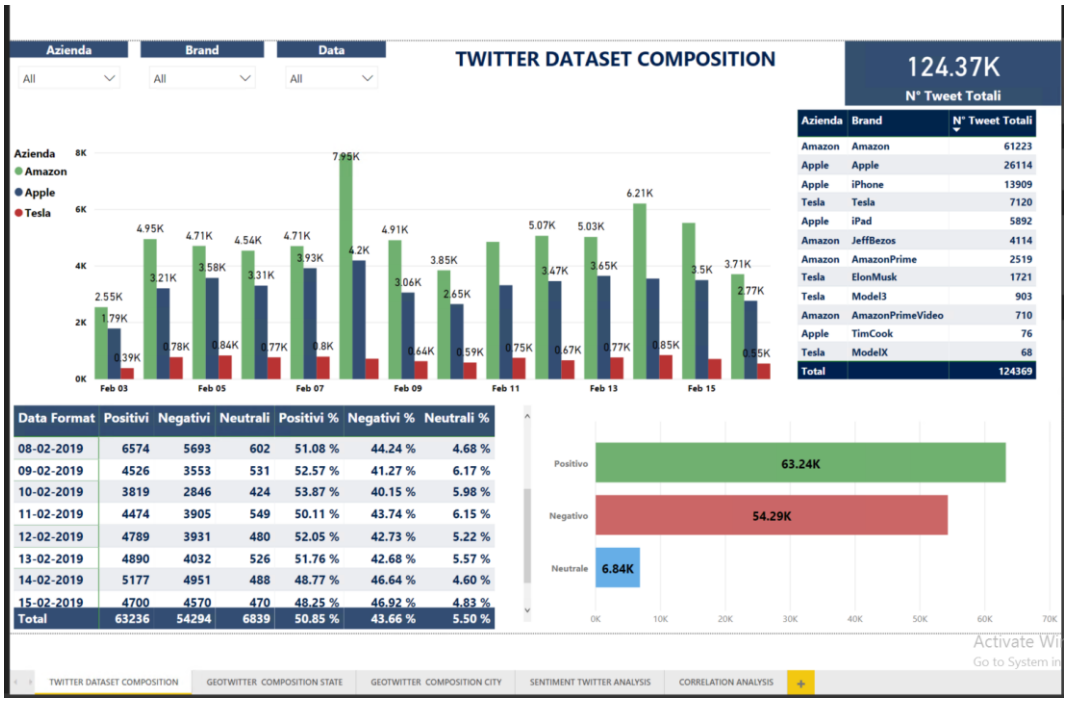
Listato 1.9: Script Estrazione dati Finanziari

```
#Estrazione dati Yahoo Finance
#carico pacchetto per scaricare dati finanziari da Yahoo Fianance

library(pdfetch)

Shares<-c("AAPL","AMZN","TSLA")

#Estrazione dati Finanziari per i vari titolo azionari
Finance_AAPL<-
data.frame(pdfetch_YAHOO("AAPL",fields="close",from=as.Date("2019-02-01")))
Finance_AMZN<-
data.frame(pdfetch_YAHOO("AMZN",fields="close",from=as.Date("2019-02-01")))
Finance_TSLA<-
data.frame(pdfetch_YAHOO("TSLA",fields="close",from=as.Date("2019-02-01")))
```

La funzione pdfetch_YAHOO ci permette di scaricare dati da Yahoo Finanza attraverso una serie di parametri: il nome del titolo azionario ricercato, quale tipologia di dati estrarre (se i dati di apertura, chiusura, volume dell'indice azionario), la data a partire dalla quale scaricare i dati. Nel nostro caso scarichiamo i dati a partire dal 3° Febbraio 2019, e prendiamo come riferimento solo i dati di chiusura del mercato azionario. Inserendo lo script R per l'estrazione dei dati finanziari all'interno di una store procedure e nel flusso dati SSIS possiamo automatizzare l'estrazione.

A questo punto abbiamo tutte le informazioni per la creazione e valutazione dei risultati finali. Per far questo ci affidiamo ai report preparati in Power BI, partendo dal report che descrive il nostro dataset di riferimento.

Figura 30: Report Twitter Dataset Composition

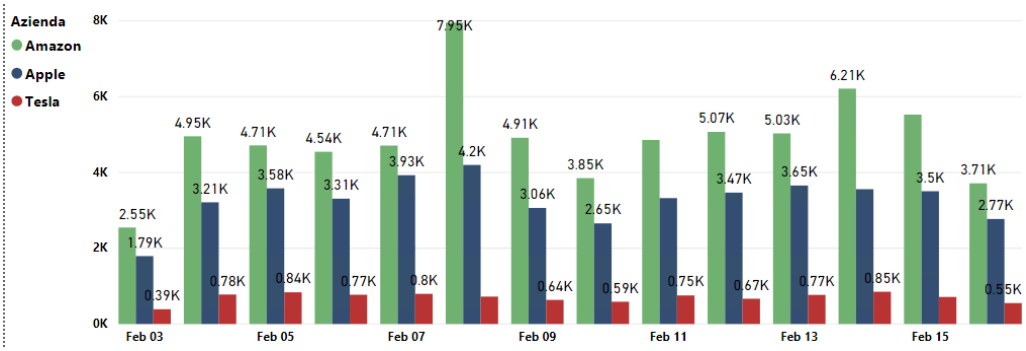


Azienda	Brand	N° Tweet Totali
Amazon	Amazon	61223
Apple	Apple	26114
Apple	iPhone	13909
Tesla	Tesla	7120
Apple	iPad	5892
Amazon	JeffBezos	4114
Amazon	AmazonPrime	2519
Tesla	ElonMusk	1721
Tesla	Model3	903
Amazon	AmazonPrimeVideo	710
Apple	TimCook	76
Tesla	ModelX	68
Total		124369

Data Format	Positivi	Negativi	Neutrali	Positivi %	Negativi %	Neutrali %
08-02-2019	6574	5693	602	51.08 %	44.24 %	4.68 %
09-02-2019	4526	3553	531	52.57 %	41.27 %	6.17 %
10-02-2019	3819	2846	424	53.87 %	40.15 %	5.98 %
11-02-2019	4474	3905	549	50.11 %	43.74 %	6.15 %
12-02-2019	4789	3931	480	52.05 %	42.73 %	5.22 %
13-02-2019	4890	4032	526	51.76 %	42.68 %	5.57 %
14-02-2019	5177	4951	488	48.77 %	46.64 %	4.60 %
15-02-2019	4700	4570	470	48.25 %	46.92 %	4.83 %
Total	63236	54294	6839	50.85 %	43.66 %	5.50 %

Il report Twitter Dataset Composition, ci fornisce una serie di informazioni circa i dati di Twitter estratti e analizzati. I dati visibili in fig. 30 si riferiscono al periodo che va dal 3 Febbraio al 16 Febbraio 2019. Tuttavia, come ribadito precedentemente, il sistema creato immette ogni giorno nuovi dati riferiti al giorno precedente[174]. Nel periodo di riferimento sono stati estratti 124.370 tweet. In fig. 31 è rappresentato il grafico che mostra il numero di tweet per giorno di estrazione, divisi per Azienda.

Figura 31: Grafico tweet giornalieri per azienda



[174] Di conseguenza il giorno dell'esposizione della tesi, verranno mostrati dati che si riferiscono ad un periodo più grande: che parte sempre dal 3 Febbraio e arriva sino al giorno prima della discussione.

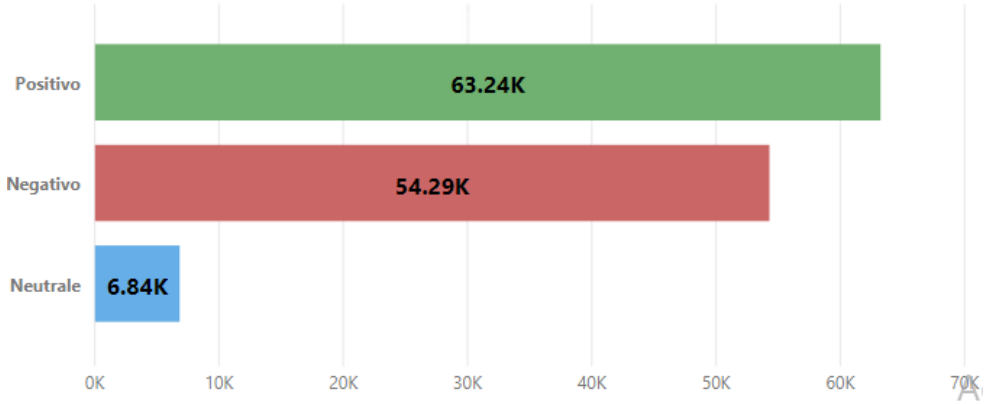
Come si vede nel grafico la maggior parte dei tweet sono riferiti ad Amazon, poi ad Apple infine a Tesla, inoltre osserviamo che il giorno in cui sono stati estratti la maggior parte dei tweet è l'8 Febbraio. In termini generali abbiamo la seguente composizione del dataset per Azienda:

Tabella 11: Composizione Dataset

Azienda	N° Tweet	% N° Tweet
Amazon	68,566	55%
Apple	45,991	37%
Tesla	9,812	8%

I due oggetti in basso nel report, invece, mostrano a livello generale la composizione del Sentiment calcolato tramite il classificatore Bayesiano. A livello generale abbiamo la seguente situazione:

Figura 32: Sentiment del Dataset Generale



Nella tabella sottostante, invece, possiamo vedere il sentiment dei tweet nel tempo, sia in valore assoluto che in valore percentuale.

Tabella 12: Valori Sentiment Assoluti e Percentuali

Data	Positivi	Negativi	Neutrali	Positivi %	Negativi %	Neutrali %
03/02/2019	2394	2056	275	50.67%	43.51%	5.82%
04/02/2019	4483	3880	576	50.15%	43.41%	6.44%
05/02/2019	4624	3977	526	50.66%	43.57%	5.76%
06/02/2019	4430	3727	463	51.39%	43.24%	5.37%
07/02/2019	4779	4124	523	50.70%	43.75%	5.55%
08/02/2019	6574	5693	602	51.08%	44.24%	4.68%
09/02/2019	4526	3553	531	52.57%	41.27%	6.17%

10/02/2019	3819	2846	424	53.87%	40.15%	5.98%
11/02/2019	4474	3905	549	50.11%	43.74%	6.15%
12/02/2019	4789	3931	480	52.05%	42.73%	5.22%
13/02/2019	4890	4032	526	51.76%	42.68%	5.57%
14/02/2019	5177	4951	488	48.77%	46.64%	4.60%
15/02/2019	4700	4570	470	48.25%	46.92%	4.83%
16/02/2019	3577	3049	406	50.87%	43.36%	5.77%
			Media	**50.92%**	**43.52%**	**5.57%**

Mediamente abbiamo i seguenti risultati:

- Il 50,92% dei tweet analizzati esprimono un sentiment generalmente positivo;
- Il 43,52% dei tweet analizzati esprimono un sentiment generalmente negativo;
- 5,57% dei tweet analizzati non sono classificabili, ossia non esprimono né un sentiment positivo né negativo.

I risultati neutrali sono quei tweet a cui il nostro classificatore non è stato in grado di attribuire un "*sentiment*", questo è dovuto per uno di questi due motivi:

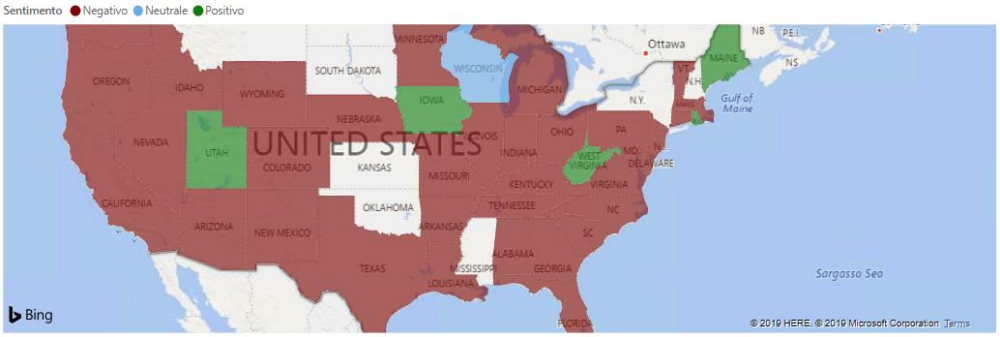
1. Il testo del tweet non contiene termini presenti nella nostra base di conoscenza (cioè l'insieme dei termini incontrati durante la fase di traning per cui esistono valori di probabilità in relazione al sentimento).
2. La seconda ipotesi, è che analizzando i termini presenti nel testo del tweet, il classificatore giunge alla conclusione che la probabilità di appartenere alla classe positivi è uguale alla probabilità di appartenere alla classe negativi (è quindi essere neutrali)

Un modo per ridurre il numero di dati neutrali è quello di ampliare la base di conoscenza su cui è costruito il classificatore.

Come visto precedentemente è stato possibile effettuare anche un'indagine geolocalizzata del Sentiment, anche se su una minima parte del dataset generale. Questo è dovuto al fatto che solo una minima parte dei tweet estratti, avevano informazioni relative alla posizione dell'utente (latitudine e longitudine). Il dataset

dei tweet geolocalizzati si compone di 1275 elementi[175]. La fig. 33 mostra il Sentiment generale del dataset per gli ogni Stato degli USA.

Figura 33: Sentiment USA



Come si vede dalla figura, la maggior parte degli Stati esprime un sentiment negativo. Lo stato del Wisconsin ha invece un sentiment neutrale, questo è dovuto ad uno delle ragioni esposte precedentemente. Si può notare come alcuni Stati non hanno un sentimento associato, ciò dipende dal fatto che non sono stati estratti tweet con una posizione a loro correlata. Si ricorda che sono stati estratti solo tweet di lingua inglese, di conseguenza la provenienza dei tweet è associata per gran parte a paesi di lingua inglese. La fig. 34 conferma quanto detto:

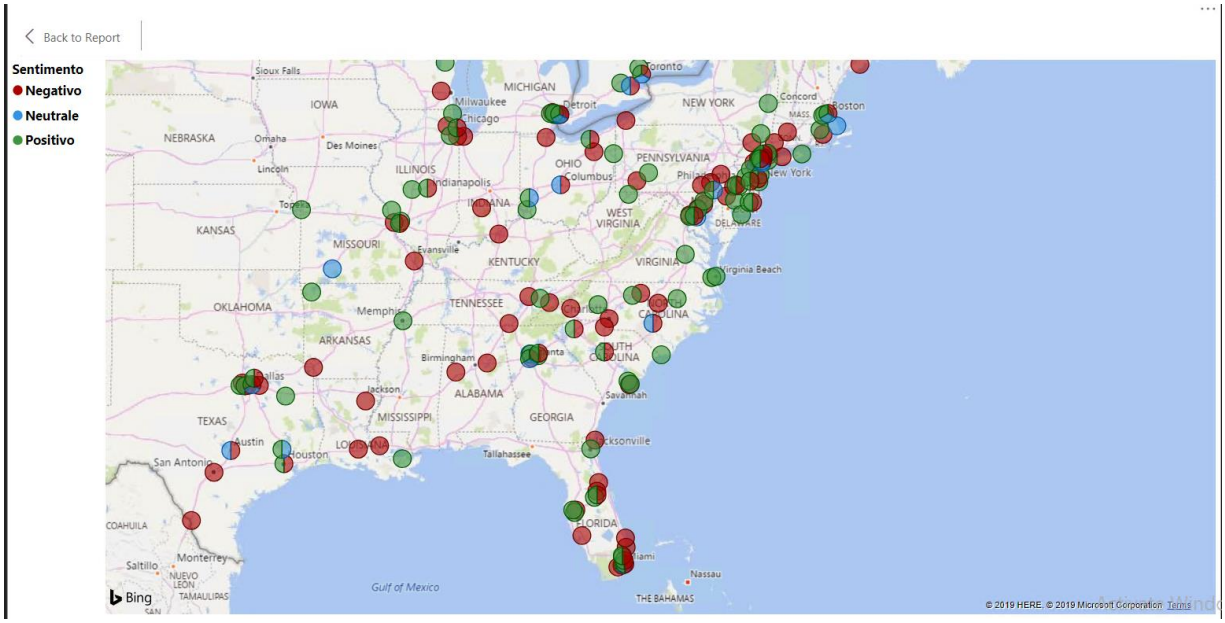
Figura 34: Dataset GeoTweet per Città

State	Region	City	N° GeoTweet
United States of America	New York	New York City	65
United States of America	California	Los Angeles	59
United Kingdom	England	London	33
Australia	New South Wales	Sydney	19
Nigeria	Lagos	Ikeja	19
United States of America	Michigan	Detroit	17
United States of America	Georgia	Atlanta	16
United States of America	Georgia	Lawrenceville	14
Indonesia	State Not Found	Jakarta Special Capital Region	13
France	Auvergne-Rhône-Alpes	Lyon	12
United States of America	Florida	Miami	12
India	Maharashtra	Navi Mumbai	11
United Kingdom	England	Derby	10
United States of America	Washington	Seattle	10

[175] Il periodo di estrazione è esattamente lo stesso del dataset generale, quindi dal 3 al 16 Febbraio. Come detto per il dataset generale, in fase di discussione del lavoro di tesi, si mostreranno dati relativi ad un periodo più ampio, per i nuovi dati che verranno estratti ogni giorno fino alla data della discussione.

La geolocalizzazione ci permette di arrivare ad un livello di granularità pari all'indirizzo, in fig. 35 è possibile vedere la distribuzione del sentiment delle città della Costa Est degli Stati Uniti:

Figura 35: Sentiment Costa Est USA



Come abbiamo detto a inizio paragrafo per dare maggior valore all'analisi abbiamo effettuato delle analisi in relazione a dei valori di benchmark. I valori di benchmark sono gli indici azionari delle varie aziende, lo scopo di questa analisi è quello di valutare se il Sentiment espresso dalle persone ha una qualche forma di relazione con l'andamento della borsa.

A livello generale il Sentiment valutato per azienda è cosi distribuito:

Figura 36: Distribuzione Sentiment Amazon

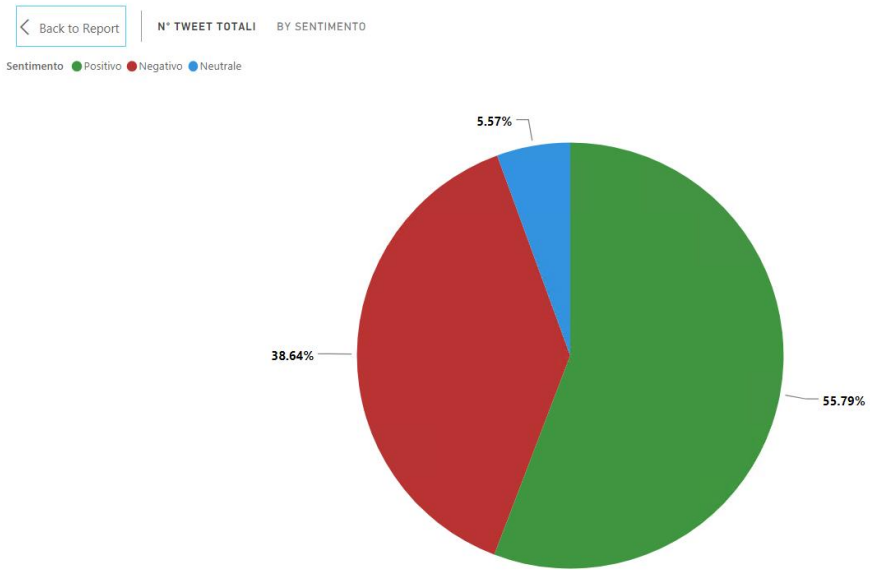


Figura 37: Distribuzione Sentiment Apple

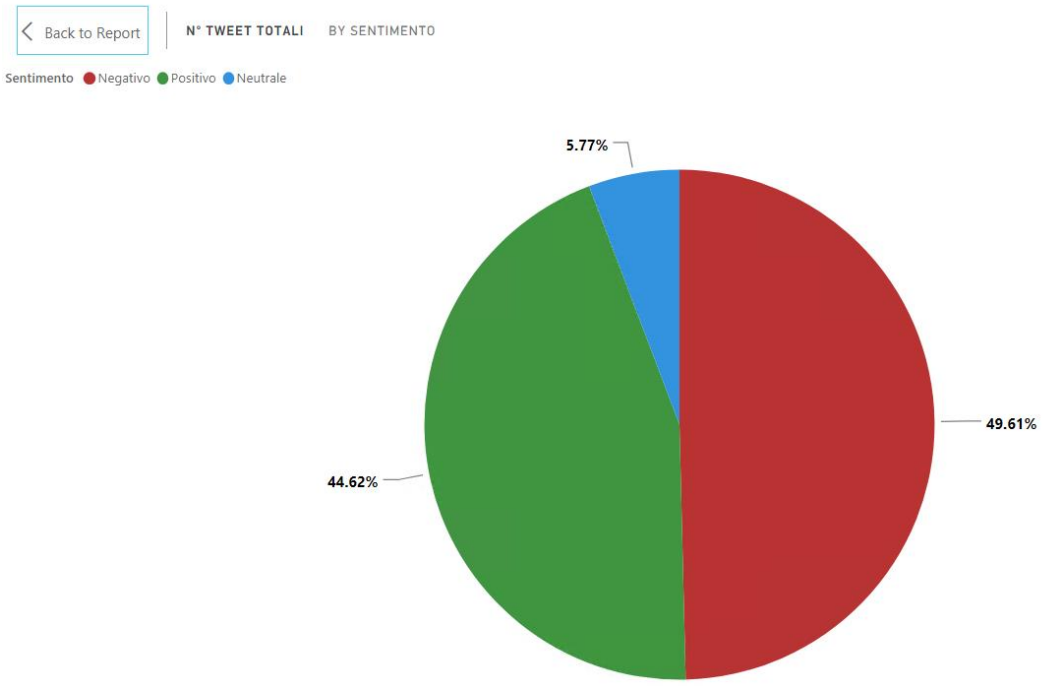
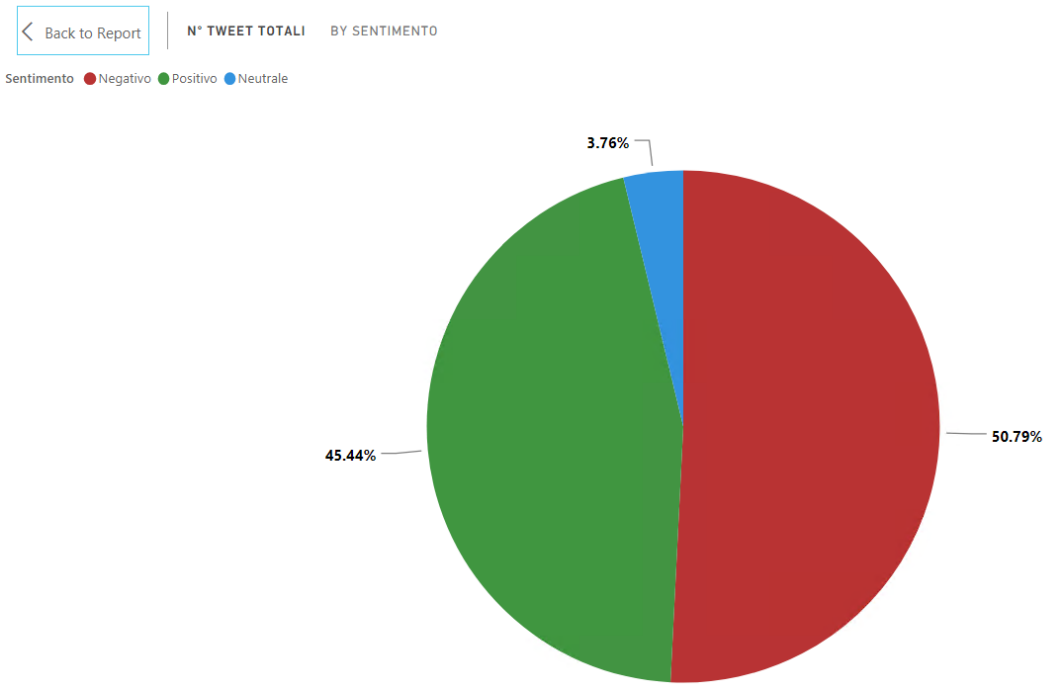


Figura 38: Distribuzione Sentiment Apple

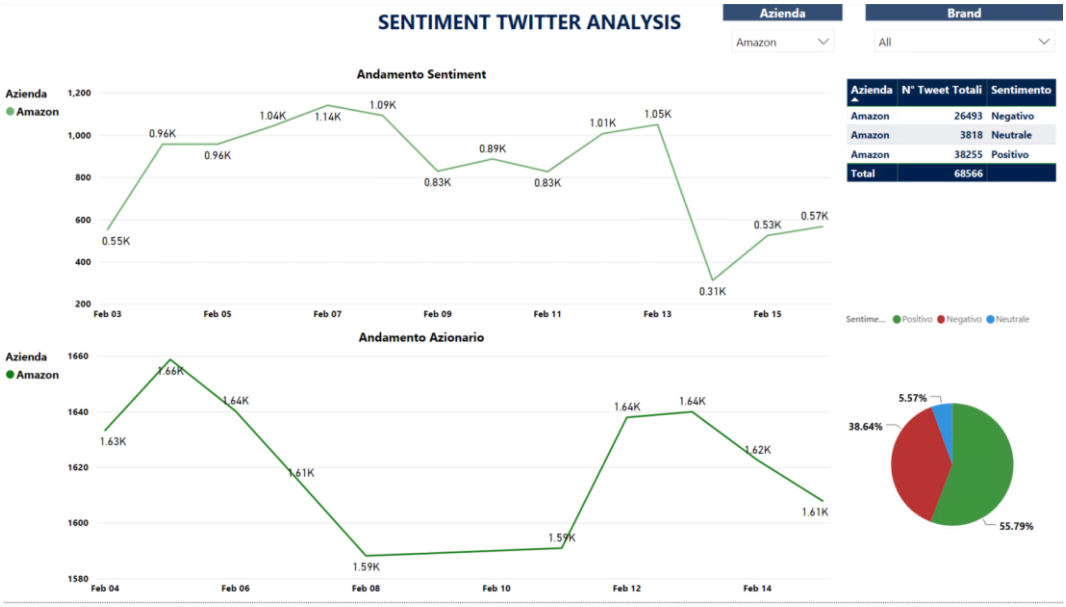


I grafici a torta qui sopra mostrano la distribuzione del sentiment per le varie aziende (calcolato su dati fino al 16 Febbraio). A livello generale la situazione è questa:

- Più della metà dei tweet associati a Amazon hanno un sentiment positivo (55,79%), i tweet non classificabili sono pari circa al 5,5%, che sono in linea con i valori neutrali del dataset generale.
- I dati di Apple e Tesla hanno una distribuzione del sentiment simile, in cui la distribuzione è abbastanza equa fra valori positivi e negativi, con una leggera tendenza verso i valori negativi. È da notare che i valori neutrali di Apple sono inferiori rispetto alla media generale, questo sta a significare che la base di conoscenza copre la maggior parte dei termini presenti nei tweet estratti per la keyword "Apple".

Per fare un confronto tra il Sentiment e gli indici di borsa, è stato creato un report specifico, in cui attraverso dei filtri di Azienda e Brand (che non sono altro che gli hashtag usati per l'estrazione), è possibile vedere l'andamento temporale del Sentiment e degli indici azionari delle singole aziende.

Figura 39: Sentiment Twitter Analysis Amazon



Come si vede nel report, è possibile valutare se l'andamento temporale del Sentiment è in tendenza con l'andamento dell'indice azionario. L'andamento del Sentiment è stato creato a partire dai valori di sentimento associato ad ogni tweet, 1 se positivo, -1 se negativo e 0 se neutrale. I valori così espressi sono stati poi sommati per ogni giorno di riferimento, così da avere un singolo dato aggregato per giorno.

Per capire meglio se vi è una qualche forma di relazione lineare tra i valori di Sentiment e i dati azionari si è andati a calcolare i valori di correlazione e si è costruito un grafico di dispersione che rappresentate la situazione. Gli indice di correlazione calcolati sono:

Tabella 13: Correlazione

Azienda	**Correlazione**
Amazon	0.22
Apple	-0.28
Tesla	-0.46

Ricordiamo che il coefficiente di correlazione va a determinare la presenza o meno di una relazione lineare tra due variabili. Quindi qualora i risultati mostrassero che non intercorre una relazione fra le due variabili, questo non significa che non vi è nessun tipo di relazione, ma che semplicemente non vi è una relazione lineare. Il coefficiente di correlazione è determinato dalla formula:

$$\rho_{\mathrm{AB}} = \frac{\operatorname{cov}(A, B)}{\sigma_A \cdot \sigma_B}$$

Dove **cov(A,B)** indica la covarianza di A e B e $\boldsymbol{\sigma_A}$ e $\boldsymbol{\sigma_B}$ indicano, rispettivamente la deviazione standard di A e di B. Il coefficiente di correlazione è un indice sempre compreso tra -1 e 1 ed in particolare:

- Se ρ_{AB}>0, le variabili A e B si dicono direttamente correlate, oppure correlate positivamente, e quanto più il valore si avvicina ad 1 tanto più la relazione è forte.
- Se ρ_{AB}=0, le variabili A e B si dicono non correlate.
- Se ρ_{AB}<0, le variabili A e B si dicono inversamente correlate, oppure correlate negativamente, e quanto più il valore si avvicina ad -1 tanto più la relazione è forte.

Detto questo, guardando i risultati ottenuti dalle nostra variabili di Sentiment e Indici Azionari, si ha che:

- Amazon e Apple hanno rispettivamente un indice di correlazione pari a 0.22 e -0.28, sono valori molto bassi come indice di correlazione. Nello specifico possiamo dire che Amazon ha una legame positivo molto debole, Apple

invece ha un legame negativo molto debole. Dai grafici di dispersione, si nota la scarsa correlazione che intercorre tra le variabili analizzate:

Figura 40: Grafico di Dispersione Amazon

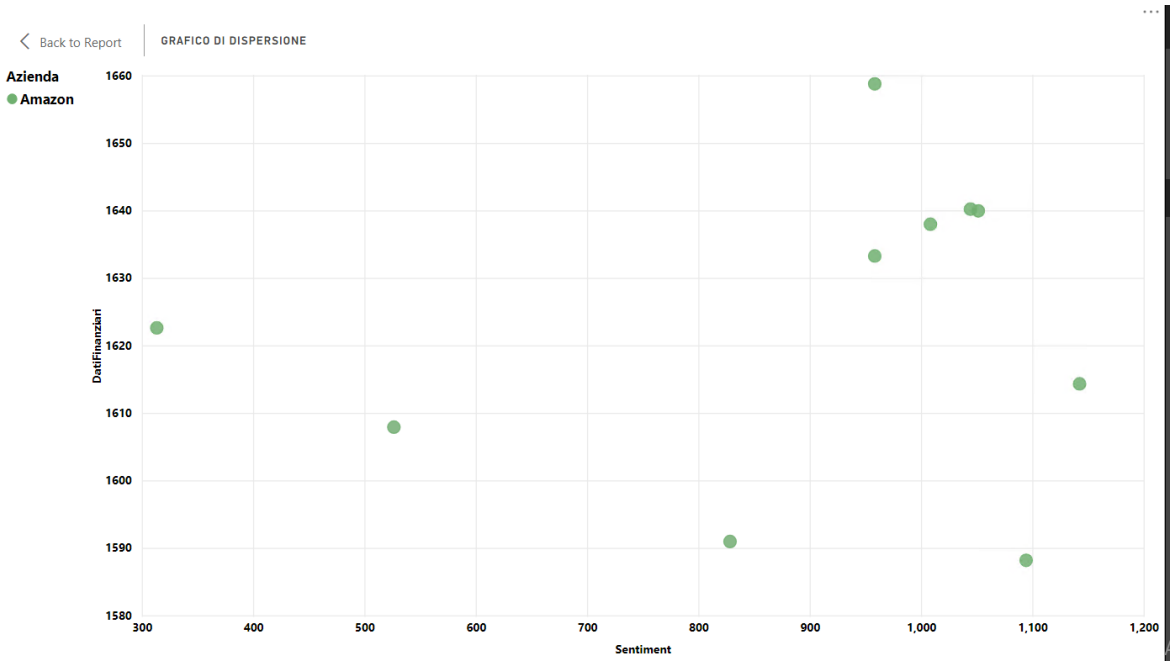
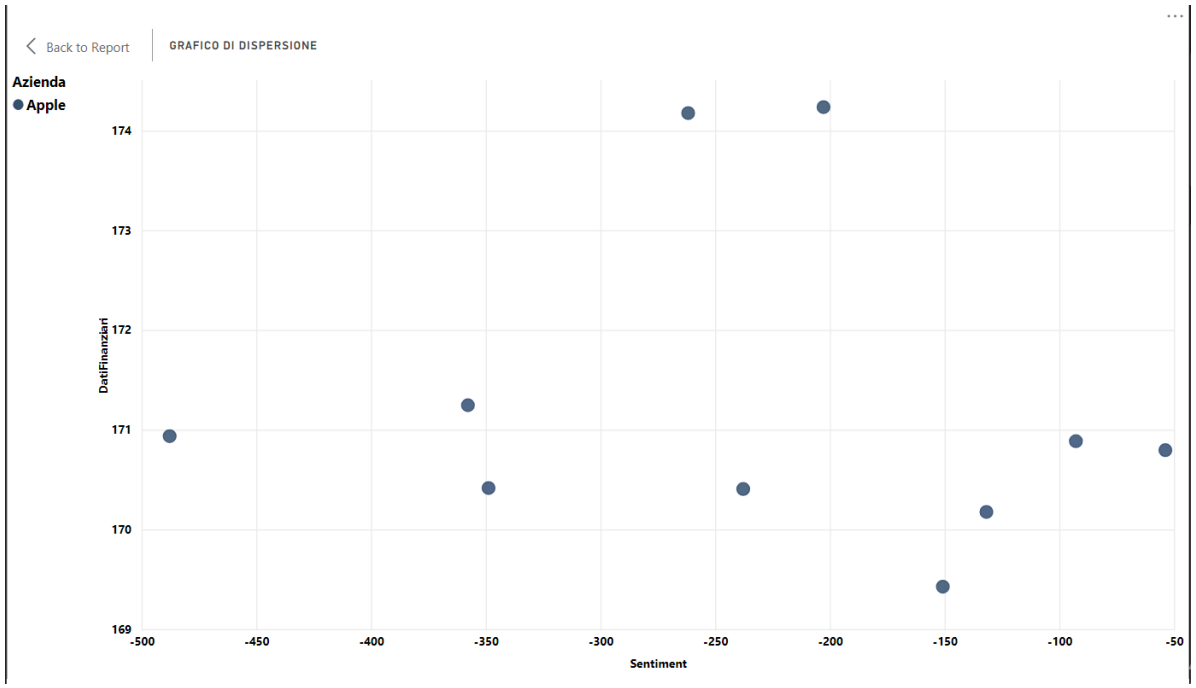


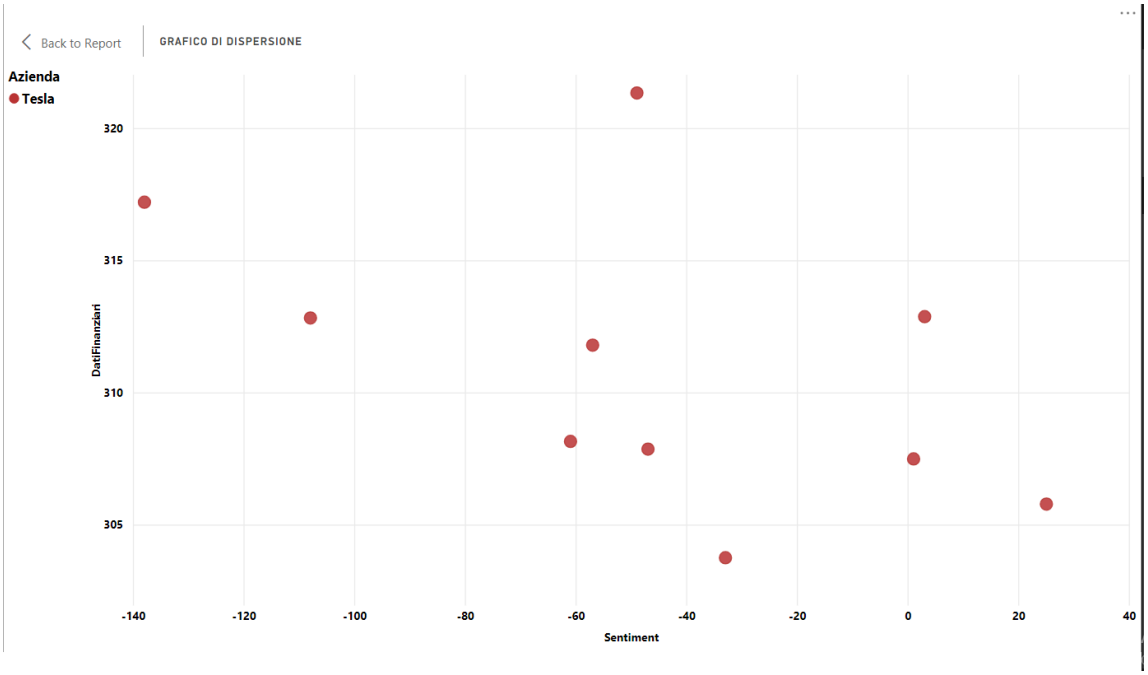
Figura 41: Grafico di Dispersione Apple



- Invece per quanto riguarda Tesla abbiamo un valore di -0.46, questo valore comporta che le variabili valutate sono inversamente correlate

moderatamente. Il grafico di dispersione effettivamente mostra la tendenza verso una correlazione negativa:

Figura 42: Grafico di Dispersione Tesla



Si nota nel grafico la tendenza verso una correlazione negativa. Sicuramente aumentando il numero dei dati si potrà delineare meglio il tipo di relazione che intercorre fra queste due variabili.

Alla luce dei dati ottenuti, la situazione appare incerta. Se da un lato apparentemente il nostro sistema ci conduce a ipotizzare che tra il Sentiment e l'andamento azionario non ci sia una relazione lineare (situazione Amazon e Apple). D'altro lato nel caso Tesla, ci si potrebbe sbilanciare a definire una correlazione negativa fra variabili.

Bisogna anche considerare che il risultato di non correlazione fra variabili, predetto dal nostro sistema, non per forza di cose è un risultato sbagliato e negativo. Ricordiamo che sono molti i fattori a influenzare l'andamento di un titolo borsa: l'andamento del mercato, il comportamento degli investitori o il lancio di nuovi prodotti, per citarne qualcuno. Non c'è da stupirsi se la percezione che gli utenti di Twitter hanno sulle aziende si discosti totalmente dall'andamento del titolo azionari.

In secondo luogo, bisogna anche considerare che il campionamento effettuato, ossia l'estrazione dei dati, è avvenuta in maniera del tutto

casuale, ossia gli unici fattori sotto il nostro controllo erano il periodo di riferimento e le keyword su cui basare la ricerca. Abbiamo pochissime informazioni sulle persone dietro ai tweet. Informazioni quali il sesso o la nazionalità[176] o il titolo di studio, sono fuori dalla nostra portata. Il campionamento dei dati quindi è stato fatto senza un metodo, senza considerare molto fattori importanti. Questo comporta che vengono considerati anche tweet che possono essere fuorvianti rispetto allo scopo prefissato. Una soluzione per evitare questo problema è quello di estrarre solo tweet che provengono da pagine di twitter specializzate in materie economiche e finanziarie, quali per esempio giornali o riviste del settore. In questo modo potremmo avere dei risultati più in linea con i nostri fini, proprio perché si effettua un campionamento mirato.

In conclusione, possiamo fare queste osservazioni, circa la difficoltà di determinare il tipo di relazione che intercorre tra Sentiment e Indice Azionario:

1. Innanzitutto, uno dei motivi per cui non è stato possibile determinare il tipo di relazione che intercorre tra queste due variabili, è la scarsità di dati a disposizione. Perché, pur disponendo di molti dati relativi al Sentiment, vi era la necessità di aggregarli in un valore giornaliero così da avere la stessa dimensione temporale degli indici azionari. Inoltre, bisogna considerare che i dati degli indici azionari si riferiscono ai solo giorni feriali, di conseguenza per i giorni festivi si è dovuto adottare la media dei valori.
2. In secondo luogo, al fine di migliorare la classificazione e quindi il sentiment generale rilevato, si consiglia di ampliare la base di conoscenza del proprio dataset di traning. Ma nel nostro caso anche di effettuare manualmente il tagging[177] dei tweet di traning e test. In questo modo si riducono sicuramente i tweet non classificabili, e migliora anche l'accuratezza del modello.

[176] La geolocalizzazione ci permette di conoscere la posiziona della persona che ha fatto in determinato tweet, non la nazionalità. Non è detto che scrive un tweet a New York sia effettivamente di nazionalità americana.

[177] Non è stato possibile effettuare un tagging manuale nel progetto di tesi per un questione di tempo. Andare a valutare il sentiment di migliaia di tweet al fine di creare una base di conoscenza per il classificatore, avrebbe comportato una mole di lavoro non gestibile.

Conclusioni

A conclusione del progetto di tesi cerchiamo di dare una risposta alle domande formulate all'inizio dell'elaborato. Nell'introduzione ci eravamo domandati se effettivamente i Big Data creassero valore per le aziende. Alla luce del valore svolto possiamo affermare che avendo bene a mente cosa cercare, la disponibilità di grosse mole di dati permette di effettuare analisi e di estrarre informazioni fondamentali. Certamente affrontare analisi di questo tipo produce dei rischi sia di costo (bisogna stare attenti che i costi di progetto non lievitino in maniera esponenziale), sia di tempo, perché implementare un sistema di BI non è immediato e, in base al progetto, vengono coinvolte molte persone e figure professionali all'interno di una azienda.

D'altro canto, è evidente che effettuare queste analisi apre molte opportunità. Pensiamo al nostro progetto, nello specifico ai dati geolocalizzati quanto possono essere sfruttati sia da aziende private, per ricerche di marketing e di branding, sia da aziende pubbliche (pensiamo ai comuni e alle regioni che vogliano, alla luce di iniziative proposte, capire l'opinione dei propri cittadini).

Riprendiamo un concetto esposto all'inizio di questa tesi per concludere. Compito di un sistema di BI (e in generale del sistema informativo) è quello di estrarre nuova conoscenza dai dati, al fine di aiutare i manager a prendere decisioni migliori e in un lasso di tempo ridotto. I Big Data permettono di effettuare nuove tipologie di analisi su nuovi dati, concetti come volume, varietà e velocità dei dati non devono oggi spaventare le aziende. Con la giusta preparazione e pianificazione sarà possibile anche prevedere cosa accadrà domani.

Possibili sviluppi futuri di questo lavoro di tesi includono:

- impiego e test di diversi tipi di classificatori, oltre al *Naive Bayes* utilizzato nella nostra implementazione, al fine di aumentare le prestazioni del sistema di classificazione. In fase di classificazione abbiamo utilizzato il modello bayesiano perché ben si adatta a trattare grosse mole di dati, essendo uno dei sistemi di classificazione più efficienti. Tuttavia, essendo un metodo supervisionato presuppone l'uso di un traning set (precostituito o creandolo attraverso il tagging manuale), questa è la fase in cui si generano la maggior

parte degli errori. Una soluzione a questo limite è quello di adottare il modello bayesiano in combinazione con sistemi non supervisionati. Per esempio, si potrebbe prima effettuare una cluster analysis per determinare i raggruppamenti presenti nei dati, e successivamente in base ai raggruppamenti predetti applicare il metodo bayesiano;

- adozione di un sistema di pre-classificazione al fine di discriminare i messaggi oggettivi, che non presentano sentimenti o opinioni ma esprimono solamente fatti, dai messaggi soggettivi che realmente esprimono un orientamento personale (*sentiment*);
- utilizzo di tecniche basate sulla *sentiment consistency* per cercare di rilevare ironia e sarcasmo all'interno dei *tweets.*

Bibliografia e Sitografia

1. Eliot, T. S. (1936). Collected poems, 1909-1935. London, Faber.
2. VERCELLIS, Carlo. Business intelligence. Modelli matematici e sistemi per le decisioni. McGraw-Hill, 2006.
3. CARAMIELLO, Carlo. L'indagine prospettiva nel campo aziendale. Colombo Cursi, 1966.
4. Vicari, S., & Troilo, G. (2011). Creatività organizzativa e generazione di conoscenza: il contributo della teoria dei sistemi cognitivi. Sinergie Italian Journal of Management, (61-62), 189-211.
5. MICELLI, S., & CORO, G. (2006). I nuovi distretti produttivi: innovazione internazionalizzazione e competitività dei territori.
6. G. Volpato, Economia e Gestione delle Imprese, Carocci Editore, 2006.
7. Martinez, M. (2004). Organizzazione, informazioni, tecnologie(pp. 1-352). Il mulino.
8. Spada, A. (2010). La razionalità in Herbert Simon e le ambiguità tra behavioral economics ed economica cognitiva. Department of Economics and Statistics "S. Cognetti De Martiis" and CESMEP, Working Paper, (6).
9. Grillo, M., & Silva, F. (1989). Impresa, concorrenza, organizzazione (pp. 1-630). Carocci.
10. A. Rezzani, "Data Skills Understandig the World", dal sito: www.dataskills.it
11. Clerici, A., & De Pra, M. (2012). Informatica e web. EGEA spa.
12. www.gartner.com
13. Rezzani, A. (2012). Business intelligence: processi, metodi, utilizzo in azienda. Apogeo.
14. Inmon, W. H. (2005). Building the data warehouse. John wiley & sons.
15. www. databasemaster.it
16. A. Rezzani, Big Data Analytics, Maggioli Editore 2017
17. Big data Interim report nell'ambito dell'indagine conoscitiva di cui alla delibera n. 217/17/CONS
18. DatiOpen.it
19. Wikipidea.org.
20. www.osservatori.net

21. Kimball, R., & Ross, M. (2011). The data warehouse toolkit: the complete guide to dimensional modeling. John Wiley & Sons.

22. http://www.beantech.it

23. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.

24. Holsheimer, M., & Siebes, A. (1994). Data mining: The search for knowledge in databases. Centrum voor Wiskunde en Informatica.

25. Linoff, G. S., & Berry, M. J. (2011). Data mining techniques: for marketing, sales, and customer relationship management. John Wiley & Sons.

26. Younis, E. M. (2015). Sentiment analysis and text mining for social media microblogs using open source tools: an empirical study. International Journal of Computer Applications, 112(5).

27. Hotho, A., Nürnberger, A., & Paaß, G. (2005, May). A brief survey of text mining. In Ldv Forum (Vol. 20, No. 1, pp. 19-62).

28. Porcu, V. (2016). Guida al text mining e alla sentiment analysis con R: Impara l'analisi dei testi con le tecniche di machine learning in R. Valentina Porcu.

29. LIU, Bing. Sentiment analysis: Mining opinions, sentiments, and emotions. Cambridge University Press, 2015

30. Ceron, A., Curini, L., & Iacus, S. M. (2014). Social Media e Sentiment Analysis: L'evoluzione dei fenomeni sociali attraverso la Rete (Vol. 9). Springer Science & Business Media.

31. https://about.twitter.com

32. https://www.digitalic.it

33. https://twitter.com/it/privacy

34. https://help.twitter.com

35. http://thinknook.com/twitter-sentiment-analysis-training-corpus-dataset-2012-09-22