Università degli studi di Milano-Bicocca
Scuola di Economia e Statistica

Corso di Laurea in
Scienze Statistiche ed Economiche

analisi delle metodologie per la sentiment analysis: il caso covid-19

relatore: Prof. Matteo Borrotti

tesi di laurea di:
Elena Signori
matricola n. 843017

anno accademico 2020/2021

A mia nonna Caterina

Indice

Introduzione 1

1 La Sentiment Analysis 3
- 1.1 Livelli di Sentiment Analysis . . . 5
- 1.2 Metodologie possibili . . . 6
  - 1.2.1 Fine-Grained . . . 6
  - 1.2.2 Aspect-Based . . . 6
  - 1.2.3 Emotion Detection . . . 7
- 1.3 Approcci utilizzabili . . . 8
  - 1.3.1 Rule-based . . . 8
  - 1.3.2 Machine learning . . . 10
  - 1.3.3 Hybrid . . . 15
- 1.4 Criticità . . . 16

2 Emotion Detection Sentiment Analysis 19
- 2.1 Metodi di clustering . . . 19
  - 2.1.1 Algoritmi di clustering gerarchico . . . 20
  - 2.1.2 Algoritmi di clustering partizionale . . . 22
- 2.2 Lexicons . . . 26

3 Applicazione Sentiment Analysis 31
- 3.1 Materiali e Metodi . . . 33
  - 3.1.1 Web scraping e descrizione dei datasets . . . 33
  - 3.1.2 Pre-processing . . . 34
  - 3.1.3 Metodologia e lexicon scelti . . . 35
  - 3.1.4 Metodi di clustering applicati . . . 36
- 3.2 Risultati . . . 36
  - 3.2.1 Analisi preliminari . . . 36
  - 3.2.2 Emotion Detection . . . 38

Conclusione 47

Bibliografia 49

Introduzione

L'interesse e la curiosità personale per la *Sentiment Analysis* e per la varietà dei suoi ambiti di applicazione, sono stati i fattori determinanti per la scelta del presente argomento.

Negli ultimi decenni si è assistito ad una progressiva diffusione di *Internet* e delle relative connessioni che ne conseguono. Queste hanno permesso la nascita di *siti web*, *microblogs* e soprattutto *social networks*, favorendo una crescita esponenziale del numero di utenti in rete. Questi portali consentono il continuo scambio di informazioni tra le persone, rendendo il web una fonte inesauribile di conoscenza che suscita interesse in molti ambiti come quello economico, politico e sportivo. L'elaborazione delle informazioni ricavate dal web consente, infatti, di comprendere le opinioni degli utenti, in modo da poter valutare quasi in tempo reale la risposta della popolazione o del pubblico in seguito a determinate scelte oppure relativamente ad un determinato prodotto o servizio.

Nella maggior parte dei casi, le informazioni ricavabili dal web sono di natura testuale e questo le rende particolarmente insidiose da elaborare, dal momento che una macchina non è in grado di comprendere il linguaggio umano allo stesso modo di una persona. Per trovare una soluzione che permettesse l'analisi dei dati raccolti, si è sviluppato un ramo dell'*Intelligenza Artificiale* particolarmente degno di interesse: la *Sentiment Analysis*. Essa consente l'elaborazione del liguaggio umano e l'analisi testuale, per identificare opinioni soggettive attraverso varie metodologie e approcci, utilizzabili a seconda del contesto di interesse e in continua evoluzione.

Nell'ambito dei social networks, la metodologia dell'*Emotion Detection* risulta più adatta delle altre, in quanto ha come obiettivo quello di cogliere le emozioni relativamente ad un determinato argomento. Per la sua applicazione ci si avvale di strumenti specifici, come i *lexicons*, particolari dizionari nei quali ogni termine ha associato il sentimento che esprime.

L'attuale situazione pandemica legata al COVID-19 e, in particolare, lo sviluppo dei vaccini contro tale virus, hanno generato attraverso i social networks una considerevole quantità di informazioni in merito all'opinione degli utenti. Lo scopo del presente elaborato, è quello di proporre una possibile applicazione di tecniche di *Sentiment Analysis*, nello specifico di *Emotion Detection*, ad una serie di informazioni estratte dal social network Twitter relative ai vaccini anti COVID-19, con l'obiettivo di cogliere la posizione degli utenti rispetto a tale argomento.

Dalle analisi effettuate è emerso che l'opinione degli utenti si è evoluta con il passare del tempo, a seconda dei risultati ottenuti dalla ricerca e dalle notizie diffuse dai mass media. Nel complesso, si delinea la prevalenza di un sentimento di fiducia nei confronti della scienza e di aspettativa rispetto all'ottenimento di risultati postivi e di miglioramento della situazione pandemica. D'altra parte risulta presente anche un certo timore per le nuove tecnologie impiegate nella formulazione dei vaccini e per la loro possibile inefficacia.

La trattazione, prima di esporre i risultati conseguiti, verterà sull'esposizione teorica delle tecniche impiegate nell'applicazione. Il primo capitolo, dopo una breve introduzione al *Natural Language Processing* e alla *Sentiment Analysis*, contiene la descrizione delle principali metodologie presenti in letteratura e degli approcci utilizzabili; inoltre, riporta le criticità più frequentemente riscontrabili nell'applicazione della *Sentiment Analysis* proponendo una possibile soluzione per ognuna. Il secondo capitolo è dedicato alla metodologia di *Emotion Detection* e all'illustazione degli strumenti più impiegati per la sua messa in pratica, quali lexicons e metodi di clustering. Nel terzo ed ultimo capitolo, viene proposta l'applicazione delle tecniche di *Sentiment Analysis* ad alcuni datasets ricavati da Twitter. Nel dettaglio, vengono descritti i datasets utilizzati e le varie metodologie impiegate per lo svolgimento delle analisi, seguite dall'esposizione dei risultati conseguiti.

In conclusione all'elaborato, verrà proposta una valutazione complessiva dei risultati, ottenuti attraverso le tecniche di *Sentiment Analysis* scelte per lo svolgimento dell'analisi di interesse.

.

Capitolo 1

La Sentiment Analysis

La continua evoluzione del mondo di *Internet* e la sua progressiva diffusione hanno portato, negli ultimi decenni, un numero sempre maggiore di utenti sul web. Nel corso del tempo si è creata una rete sempre più fitta di connessioni che ha portato in seguito alla nascita di *siti web*, *microblogs* e *social networks* – come Twitter, Tumblr, Facebook, Instagram –, oggi in continuo aumento, i quali offrono innumerevoli potenzialità sia dal punto di vista informativo sia relazionale.

Sempre più spesso gli utenti scelgono spontaneamente di esprimere le proprie opinioni in merito ai più vari argomenti, spaziando da prodotti e servizi, al mondo politico, a scelte di tipo etico. In genere, quando bisogna prendere una decisione è molto comune chiedere il parere della famiglia o degli amici più stretti; grazie al web è ora possibile trovare centinaia di opinioni, qualsiasi sia il problema di nostro interesse, e avere quindi un numero maggiore di punti di vista.

La necessità di condividere il proprio pensiero deriva proprio dalla consapevolezza che alcuni potrebbero trovare utile o interessante venire a conoscenza del parere altrui per compiere le proprie scelte. Un ulteriore motivo che spinge molte persone ad esporsi sul web, in particolare relativamente a questioni etiche e morali, è la convinzione che la propria opinione sia quella più giusta e la volontà di influenzare il pensiero altrui. A tal proposito possiamo osservare numerosi esempi nell'ambito politico o medico, soprattutto se si tratta di questioni di carattere etico, come l'immigrazione o l'eutanasia. Inoltre, alla condivisione spontanea del proprio parere si contrappone la richiesta esplicita dell'opinione degli altri utenti, spesso interrogata attraverso sondaggi sui social networks anche in merito alle più banali questioni, come la scelta di un vestito.

Questo continuo scambio di opinioni genera una quantità considerevole di informazione, che può essere analizzata per trarre conclusioni utili circa il pensiero generale condiviso dagli utenti, per cogliere il sentimento diffuso e la sua intensità. L'analisi di queste informazioni genera conoscenza che può essere sfruttata in particolare dalle organizzazioni per indirizzare le proprie decisioni in modo da soddisfare il più possibile le aspettative e le necessità dei propri clienti. Questo è il motivo per cui oggi l'immagine, intesa sia come aspetto estetico sia come reputazione, assume un ruolo centrale. Infatti, sia persone fisiche che organizzazioni, sono sempre più sottoposte all'opinione e al giudizio altrui che non trovano limiti né ostacoli nell'essere espressi, proprio grazie al web.

Tuttavia, le informazioni disponibili sono di tipo non strutturato, il che le rende più complesse da elaborare e analizzare, rispetto a quelle di tipo strutturato. Infatti, l'*informazione strutturata* si presenta in una forma ordinata e organizzata, solitamente distinta in campi per ciascuno dei quali è definito uno specifico formato; questo rende i dati facilmente trattabili. Al contrario, l'*informazione non strutturata* si presenta in forma libera, spesso di tipo testuale o multimediale e questo rende la sua elaborazione più difficoltosa.

Il *Natural Language Processing* (NLP) è un ramo dell'intelligenza artificiale che permette all'elaboratore di comprendere, manipolare ed interagire con il linguaggio umano, espresso sia in forma orale sia scritta, attraverso alcune tecniche computazionali (Liddy, 2001). La *Sentiment Analysis* (SA) – nota anche come *Opinion Mining* – rientra tra le tecniche proprie del NLP e ha l'obiettivo di analizzare le opinioni, le valutazioni, gli atteggiamenti e le emozioni delle persone nei confronti di entità, individui, problemi, eventi, argomenti e loro attributi (Liu, 2011). Come già detto, le nostre convinzioni e le percezioni della realtà, così come le scelte che facciamo, sono considerevolmente condizionate da come gli altri vedono e valutano il mondo. Per questa ragione, quando dobbiamo prendere una decisione spesso cerchiamo le opinioni di altri.

Liu (2011) fornisce la seguente definizione di *opinione* specificando gli elementi principali che la compongono:

- **Entity:** è un prodotto, servizio, persona, evento, organizzazione o argomento a cui si riferisce il pensiero espresso;
- **Aspect:** gli aspetti di un'entità sono i suoi attributi e le sue componenti;
- **Opinion holder:** è la persona o l'organizzazione che esprime l'opinione;

- **Opinion:** è l'opinione in sé che viene espressa e può essere positiva, negativa o neutra oppure può essere espressa con diversi livelli di intensità. Inoltre, si possono distinguere due tipi di opinioni:
  - **Direct opinions:** sono espresse relativamente alle entità e ai loro aspetti in modo diretto;
  - **Indirect opinions:** le opinioni sulle entità sono espresse in base ai loro effetti su altre entità.

Da quanto appena esposto risulta chiaro che la *Sentiment Analysis* si concentra su frasi soggettive e non su quelle oggettive, che non devono essere confuse con opinioni soggettive indirette.

È bene sottolineare che la *Sentiment Analysis* può essere applicata a testi scritti, segnali audio e a video. L'ambito in cui è maggiormente impiegata è quello relativo a documenti scritti in quanto risulta molto più semplice. Infatti, sono disponibili un numero elevato di tecniche utilizzabili. Per quanto riguarda i segnali audio, la *Sentiment Analysis* si sta evolvendo in modo significativo data la sempre maggiore presenza di assistenti vocali nei dispositivi tecnologici -– ad esempio Siri, Alexa o Google Assistant – che offrono diverse potenzialità. Infine, il ramo relativo ai video risulta meno sviluppato degli altri, in quanto più complesso e perché necessita di tecnologie di intelligenza artificiale avanzate, anche se si stanno facendo molti progressi (Vartanova, 2019).

1.1 Livelli di Sentiment Analysis

La *Sentiment Analysis* può essere svolta a diversi livelli, in base all'obiettivo dell'analisi e all'interesse del ricercatore (Mishra & Jha, 2012).

Document-level: a questo livello viene analizzato l'intero testo disponibile per identificare il sentimento generale espresso dall'*opinion holder*, con l'obiettivo di cogliere la positività, negatività o neutralità del sentimento. Per ottenere un risultato soddisfacente, è necessario che il documento riguardi una sola entità e sia stato scritto da un unico autore.

Sentence-level: viene considerata la singola frase, di cui viene classificato il sentimento. Se la frase è oggettiva viene tralasciata in quanto, come già spiegato precedentemente, la *Sentiment Analysis* si concentra su opinioni soggettive.

Feature-level: è il livello con maggiore granularità e considera le opinioni espresse in merito al singolo aspetto di un'entità. Quindi, una singola frase potrebbe corrispondere ad una sola opinione oppure potrebbe contenere al suo interno più opinioni relative a diversi aspetti della stessa entità.

1.2 Metodologie possibili

Come già detto all'inizio di questo capitolo, l'obiettivo principale della *Sentiment Analysis* è il riconoscimento del sentimento espresso per poterlo categorizzare come positivo, negativo o neutro per determinare la natura dell'opinione espressa complessivamente in un testo. Tuttavia, esistono tecniche di vario tipo.

1.2.1 Fine-Grained

Nella maggior parte dei casi, una volta individuato il sentimento questo viene classificato in modo binario come positivo o negativo – eventualmente anche neutro. Con la metodologia *Fine-Grained* viene utilizzata una scala più fine rispetto a quella binaria, infatti, al sentimento viene assegnata una tra cinque possibili modalità che possono essere ad esempio:

1. Fortemente negativo
2. Negativo
3. Neutro
4. Positivo
5. Fortemente Positivo

Questa metodologia è di facile comprensione e utilizzo in quanto ricorda il sistema usato per le recensioni online. In rete sono molto frequenti i metodi di valutazione a cinque livelli per le recensioni di prodotti e servizi, il più comune è organizzato sulla base di cinque stelle.

1.2.2 Aspect-Based

Nella *Aspect-Based Sentiment Analysis* (ABSA) – nota anche come *Feature-Based* – l'obiettivo è quello di identificare il singolo aspetto dell'entità a cui si riferisce l'autore del testo considerato e, in base all'opinione espressa, attribuire una modalità al sentimento.

Si tratta di una metodologia utile soprattutto per le aziende che offrono prodotti o servizi, perché è possibile ricavare informazioni più dettagliate sull'opinione dei clienti. Infatti, può accadere che un cliente sia soddisfatto di una determinata caratteristica di un prodotto, mentre altre non siano state di suo gradimento. Grazie alla ABSA, analizzando le recensioni relative al prodotto in questione, l'azienda può migliorare quelle componenti che non soddisfano le esigenze del cliente.

Ad esempio, per le aziende sviluppatrici di *app* o di *software*, questa metodologia consente una maggiore efficienza nel risolvere eventuali problemi. Infatti, grazie alle segnalazioni degli utenti si può capire se una determinata area dell'*app* presenta dei difetti ed è quindi possibile intervenire tempestivamente, senza che gli sviluppatori debbano preoccuparsi di leggere un gran numero di comunicazioni ricevute.

Sicuramente, è molto più semplice classificare un testo come positivo o negativo, rispetto alla maggiore specificità fornita dalla ABSA. Tuttavia, questa tecnica seppur più complessa, permette di approfondire le necessità dei clienti e comprendere le loro aspettative, perché vengono colti tutti gli elementi contenuti nel testo, senza limitarsi a considerarlo nel complesso.

1.2.3 Emotion Detection

Oltre ad attribuire alle frasi del testo di interesse una modalità delle cinque disponibili della *Fine-Grained Sentiment Analysis* oppure approfondire attraverso la *Aspect-Based Sentiment Analysis*, definendo l'opinione come positiva o negativa, è possibile scendere ancora più nel dettaglio.

In molti casi, non è sufficiente limitarsi alla sola classificazione binaria, infatti, anche se è vero che l'essere umano prova sentimenti positivi o negativi, d'altra parte è noto che le emozioni che si possono provare sono varie. Per questo motivo, l'*Emotion-Detection Sentiment Analysis* si pone l'obiettivo di identificare il più specificatamente possibile l'emozione dell'autore (*opinion holder*) in base a quanto scritto nel testo – ad esempio paura, gioia, tristezza, rabbia, sorpresa ecc.

Questa metodologia risulta più complessa rispetto alle precedenti, perché il riconoscimento delle emozioni è difficile da rendere automatico ed eseguibile da un algoritmo. Ciò è dovuto al fatto che mentre, ad esempio, la *gioia* è chiaramente positiva e la *tristezza* negativa, se si considera la *sorpresa* questa può essere

in alcuni casi intesa positivamente in altri negativamente, quindi l'algoritmo necessiterà di più elementi per poterla identificare.

Un altro fattore che causa difficoltà nell'applicazione di questa metodologia è il fatto che alcune espressioni si evolvono o ne nascono di nuove, in particolare tra i più giovani, i quali diffondono l'uso di parole alle quali viene attribuito un significato diverso da quello conosciuto abitualmente.

1.3 Approcci utilizzabili

La *Sentiment Analysis* viene svolta in modo automatico attraverso gli elaboratori i quali devono essere istruiti attraverso degli algoritmi. L'algoritmo riceve in ingresso un testo e, in base alla metodologia applicata – si veda Sezione 1.2 – restituisce una classificazione del testo, come positivo o negativo, oppure le emozioni identificate.

Per gli algoritmi impiegati si osservano due approcci principali, i quali possono essere combinati in un approccio ibrido.

1.3.1 Rule-based

L'approccio *Rule-based* consiste principalmente in tre fasi:

1. Apprendimento delle regole di estrazione delle *feature* (oggetti o aspetti);
2. Estrazione delle frasi di opinione;
3. Identificazione dell'orientamento (*polarità*) dell'opinione.

La *prima fase* serve per imparare le regole di estrazione che saranno necessarie per poter eseguire le fasi successive. Come primo step, vengono individuate le frasi che esprimono un'opinione in merito all'oggetto di interesse (ad esempio: un prodotto, un film o un libro) le quali verranno usate, in seguito, per costruire le regole. Poi, le frasi individuate vengono scomposte e convertite in set di parole. Questo passaggio avviene attraverso alcune tecniche molto usate nella *Sentiment Analysis*:

Part Of Speech tagging: il POS tagging prevede che per ogni parola venga identificato il suo ruolo grammaticale nella frase, a seconda del contesto e del suo significato. Ci sono, infatti, diversi termini la cui categoria lessicale (verbo, aggettivo, avverbio, preposizione, nome, ecc.) può variare a seconda della posizione che occupano nella frase.

Stemming: si tratta di un processo che permette di ricondurre ogni parola alla propria radice, in modo da semplificare le operazioni di interrogazione del dataset.

Meaningful Words Selection: con questa procedura vengono mantenuti solo i termini che sono significativi per l'analisi a livello semantico; vengono quindi eliminate le cosiddette *stopwords*: parole poco importanti come le preposizioni, gli articoli, i pronomi e le congiunzioni.

Dopo aver svolto questi passaggi, come ultimo step si esegue l'apprendimento delle regole di estrazione per ogni *feature*, cioè per ogni oggetto di interesse. Ogni regola viene costruita usando le parole più importanti, che sono state selezionate al secondo step attraverso la *Meaningful Word Selection*. Quindi, per ogni *feature* si avrà a disposizione un dizionario contenente le parole più usate per riferirsi a quell'oggetto. La regola verrà formulata nel modo seguente: se una parola, di una nuova frase di opinione, appartiene al dizionario relativo ad un determinato oggetto, allora è molto probabile che la nuova frase si riferisca a quell'oggetto.

Nella *seconda fase* vengono estratte le frasi di opinione per ogni *feature* sulla base delle regole create in precedenza. Il modo più semplice per stabilire se una nuova frase è relativa all'oggetto considerato, è quello di osservare se le parole del dizionario corrispondente sono presenti al suo interno. Tuttavia, bisogna tenere in considerazione che, in alcuni casi, più regole possono essere valide per la medesima frase.

La *terza fase* ha come obiettivo quello di stabilire l'orientamento dell'opinione di ogni frase estratta nella fase precedente. Per questa procedura possono essere applicati diversi metodi, tra i quali troviamo anche l'uso dei dizionari *lexicons* – che verranno approfonditi nel capitolo successivo – o dizionari creati manualmente e specifici per l'analisi di interesse (Yang & Shih, 2012).

Ad esempio, nell'articolo di Romanyshyn (2013), per il riconoscimento del sentimento è stato precedentemente creato un dizionario contenente parole alle quali è associata sia la natura del sentimento – positivo o negativo – sia la rispettiva emozione. Successivamente, sono stati individuati i singoli termini contenuti nella frase in esame, ai quali è stata assegnata la modalità del sentimento e l'emozione sulla base di quanto contenuto nel dizionario. Qualora una parola non fosse presente nel dizionario, in genere le viene attribuito sentimento "neutro" ed emozione "nessuna". Dopo aver assegnato ad ogni parola le rispettive informazioni, si utilizza un algoritmo che deve valutare la frase nel suo complesso

per poter stabilire la natura del contenuto. Per svolgere questo compito è stato istruito con una serie di regole, che permettono di stabilire quali combinazioni di parole sono da considerare come opinione positiva e quali come negativa. Alcuni esempi, presi dall'articolo citato, possono essere:

- INTENSIFIER + POSITIVE → VERY POSITIVE
- INTENSIFIER + NEGATIVE → VERY NEGATIVE
- POSITIVE + NEUTRAL → POSITIVE
- NEGATIVE + NEUTRAL → NEGATIVE

1.3.2 Machine learning

Il secondo approccio che può essere utilizzato e quello del *Machine Learning*, una branca dell'*Intelligenza Artificiale* (AI). Una delle domande che ha portato alla nascita di questa tecnologia è la seguente:

«Come si può costruire un sistema informatico che migliori automaticamente attraverso l'esperienza?»

Questo quesito ha trovato risposta nel *Machine Learning*, che ha permesso l'*apprendimento automatico* dell'elaboratore attraverso l'*esperienza*.

Secondo Mitchell et al. (1997), infatti, "un programma apprende dall'esperienza (E) con riferimento ad alcune classi di compiti (T) e con misurazione della performance (P), se le sue performance nel compito T, come misurato da P, migliorano con l'esperienza E". Ciò significa che se il programma migliora la propria efficienza nel svolgere una determinata attività, attraverso la ripetitività della stessa – vale a dire con l'esperienza – , allora si può dire che ha appreso.

Questo avviene senza che sia necessario utilizzare linguaggi di programmazione per definire a priori le azioni del programma, a seconda degli scenari che si potrebbero verificare. Se si considera, ad esempio, una partita di scacchi, non è necessario fornire all'elaboratore una lunga serie di comandi, per comunicargli quale mossa compiere in seguito a quella dell'avversario. È sufficiente che esso svolga una serie di partite contro sé stesso ed impari dagli errori commessi, avendo come obiettivo quello di realizzare il maggior numero di vincite possibili. L'elaboratore riceverà quindi poche istruzioni, poiché basterà definire i compiti che devono essere svolti dal programma – giocare a scacchi – , il criterio di misurazione delle performance – il numero di vincite – e l'esperienza attraverso

la quale verrà compiuto l'apprendimento – le partite di scacchi contro sé stesso (Mitchell et al., 1997).

Nel caso della SA, tale approccio risulta più automatico rispetto al *Rule-based*. Infatti, l'algoritmo prevede che venga istruito un classificatore che si occupa di assegnare ad ogni testo o frase il sentimento o l'emozione più appropriati. I passaggi che caratterizzano l'algoritmo di *Machine Learning* sono i seguenti (Thakkar & Patel, 2015):

Data collection: vengono raccolti una serie di testi, se necessario da diverse fonti, i quali serviranno per istruire il classifcatore.

Pre-processing: si effettua una prima pulizia dei dati in modo da renderli più facilmente trattabili. Ad esempio, vengono rimossi i segni di punteggiatura, lettere maiuscole, eventuali emoticons ecc.

Training data: i testi raccolti nella *data collection* vengono classificati manualmente, assegnando a ciascuno di essi il sentimento adatto. Questo dataset è di fondamentale importanza perché viene utilizzato per istruire il classificatore, quindi deve essere il più corretto e rigoroso possibile.

Classification : il classificatore, una volta istruito utilizzando le informazioni contenute nel *training set*, riceve in input nuovi testi, sconosciuti e non classificati manualmente. Il suo scopo è quello di assegnare ai nuovi testi il sentimento adatto sulla base di quanto appreso precedentemente.

L'approccio di *Machine Learning* permette di ottenere dei buoni risultati anche quando il dizionario diventa di grandi dimensioni. Tuttavia, rispetto al *rule-based* risulta più complesso data la necessità di scegliere il classificatore più adeguato e avere a disposizione dei *training data*, senza dimenticare che i risultati ottenuti vanno interpretati correttamente. D'altra parte, anche se richiede più tempo rispetto al metodo precedente, l'accuratezza dei risultati è maggiore (Thakkar & Patel, 2015).

La scelta del tipo di classificatore viene fatta in base ai dati e al contesto in cui viene applicato il *Machine Learning*. Inoltre, è bene sottolineare che la procedura può essere di tipo *supervisionato* o *non supervisionato*. Nel primo caso, è previsto l'uso di un *training set*, come illustrato precedentemente, che viene impiegato per l'istruzione del classificatore. Nel caso, invece, dell'approccio *non supervisionato*, non si dispone di questo dataset ma vengono utilizzati altri metodi, che hanno lo scopo di identificare una qualche struttura nei dati forniti in input.

Di seguito sono riportati alcuni dei classificatori più utilizzati nel *Machine Learning*.

Naive Bayes

Scegliendo un *Naive Bayes Classifier* si cerca di assegnare all'elemento di interesse, che può essere un documento o una frase (si veda Sezione 1.1) una classe. Nel contesto preso in considerazione, le classi possibili sono quella *positiva* e quella *negativa*, le quali vengono assegnate sulla base della probabilità che risulta maggiore in base alla formula seguente:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

dove x rappresenta l'elemento di interesse e c la classe.

La probabilità di interesse ($P(c|x)$), cioè la probabilità di appartenenza alla classe c, può essere ottenuta in modo più semplice attraverso la formula:

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \ldots \times P(x_n|c) \times P(c)$$

dove X è il documento o la frase di interesse, composta da $x_1, x_2, \ldots, x_n$ termini. Quindi, la probabilità di ogni singolo termine contribuisce in modo indipendente alla probabilità a posteriori di appartenenza alla classe c. Tuttavia, tale indipendenza è rara da verificare (Romero Llombart, 2017).

Random Forest

Con il *Random Forest* vengono istruiti una serie di *alberi decisionali*. Un albero decisionale è una struttura di dati che viene letta dall'alto verso il basso ed è costituita da una serie di elementi (Figura 1.1):

- **Nodi:** sono gli elementi che contengono le informazioni;
- **Archi:** sono le connessioni che legano i nodi;
- **Radice:** è il nodo principale, da cui ha origine l'*albero*;
- **Foglie:** sono i nodi terminali, che non presentano ulteriori connessioni.

Gli *alberi decisionali* vengono utilizzati per effettuare delle classificazioni sulla base di alcune condizioni che vengono scelte secondo il criterio del *guadagno di informazione*. Questo significa che l'albero decisionale includerà una certa

condizione, contenuta in uno dei nodi, solo se questa si rivela utile a dividere i testi o frasi nelle classi.

L'obiettivo è quello di riuscire a dividere gli oggetti nelle classi con il minor numero di condizioni possibili. Per valutare l'utilità di una condizione in termini di guadagno di informazione, generalmente si utilizza una metrica chiamata *impurità*, la quale risulta nulla nel caso in cui una condizione attribuisce tutti gli oggetti ad una sola classe. In questo caso significa che ci si trova in presenza di un *nodo puro*, che costituisce quindi una *foglia* dell'albero (Casadei, 2019b).

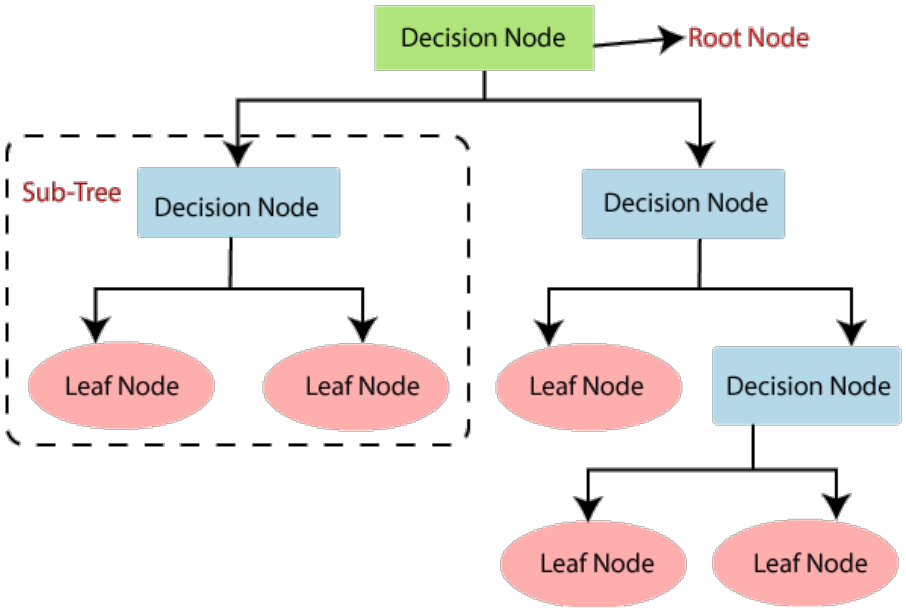

Figura 1.1: Schema di un albero decisionale

Support Vector Machines

Il classificatore basato sui *Support Vector Machine* (SVM) è un metodo secondo il quale ogni set di termini, che compongono un testo una frase, rappresenta un punto – un *vettore* – in un *iperspazio* e il SVM cerca di separare tali punti attraverso un *iperpiano* (Figura 1.2). Quest'ultimo deve essere identificato in modo che sia massimizzata la distanza tra l'iperpiano stesso e ogni vettore (*margine*).

L'individuazione dell'iperpiano migliore non è sempre semplice, soprattuto quando vi è un elevato numero di classi. Infatti, le migliori performance si ottengono nel caso di classificazione binaria, per questo in alcuni casi si permette la misclassificazione di alcuni elementi per migliorare la performance complessiva (Romero Llombart, 2017).

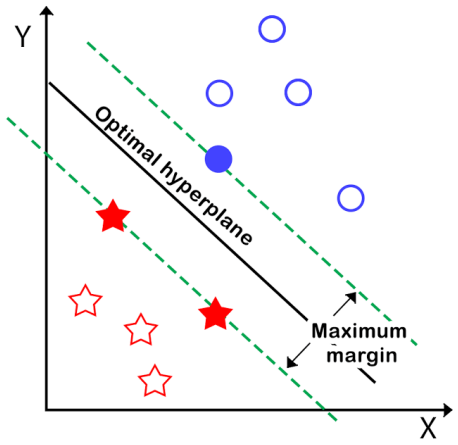

Figura 1.2: Schema di un Support Vector Machine

Reti Neurali

Le *Reti Neurali* – note anche come *Reti Neurali Artificiali* – sono dei sistemi informatici che simulano le elaborazioni compiute dalle *reti neurali biologiche*, costituite da un gran numero di cellule nervose, chiamate *neuroni*, e collegate tra loro attraverso miliardi di connessioni.

Le singole cellule nervose ricevono diversi tipi di informazioni a seconda del proprio ruolo. Infatti, si distinguono unità di ingresso (*input*) che ricevono informazioni dall'ambiente, unità di uscita (*output*) che emettono risposte nell'ambiente e unità nascoste (*hidden*) che comunicano solo con le altre cellule interne alla rete. Ogni unità si attiva quando il segnale che riceve supera una certa soglia e, a quel punto, emette un segnale che viene trasmesso nella rete attraverso le altre unità a cui è connessa.

Le *Reti Neurali* supervisionate vengono istruite fornendo esempi di ingressi con le rispettive uscite e, attraverso queste informazioni, la rete impara ad inferire la relazione che le lega. L'algoritmo apprende usando tali esempi per modificare i pesi delle connessioni e altri parametri della rete, in modo da minimizzare l'errore di previsione dell *output*, noto a priori. In seguito, verranno forniti in *input* dati sconosciuti e, se l'apprendimento è stato compiuto con successo, l'algoritmo dovrebbe aver imparato a riconoscere le relazioni che legano le informazioni in ingresso a quelle di uscita e dovrebbe essere in grado di compiere la previsione.

Tuttavia, esistono anche *Reti Neurali* non supervisionate, il cui algoritmo viene addestrato sulla base di dati per i quali non è nota l'informazione in uscita, cioè sono note solo le variabil in ingresso. In questo caso, i pesi e altri parametri

vengono modificati solo sulla base dei dati disponibili, cercando di raggruppare le informazioni ricevute in *input* formando dei *clusters* rappresentativi di tali informazioni (Gallo & Bioagromed, 2007).

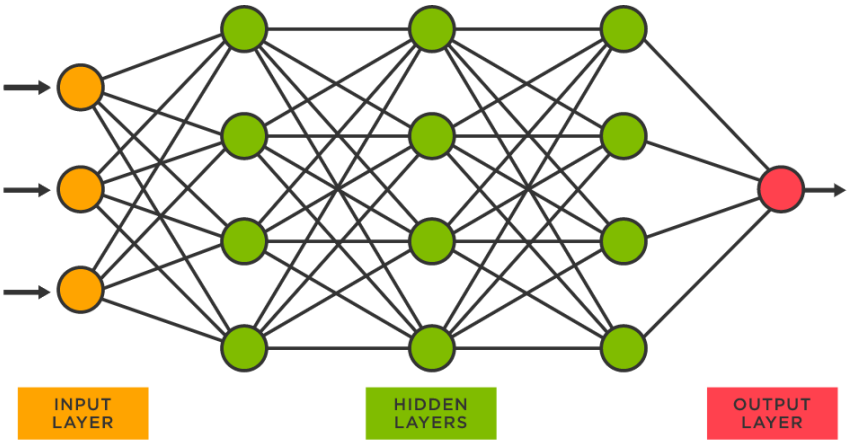

Figura 1.3: Schema di una Rete Neurale

I classificatori illustrati sono, generalmente, usati in modo *supervisionato*, fatta eccezione per le *Reti Neurali* che possono essere sia *supervisionate* sia *non supervisionate*. Altri classificatori *non supervisionati* che vengono impiegati, sono i metodi di *clustering* che saranno approfonditi nel capitolo successivo.

1.3.3 Hybrid

Per sfruttare i vantaggi dei due approcci visti fino ad ora, è stato pensato un ulteriore approccio *hybrid* che possa sfruttare la velocità del *Rule-based approach* e l'accuratezza del *Machine Learning*.

Ad esempio, Liu et al. (2004) hanno proposto una soluzione che permetta all'utente di impiegare un minore sforzo per arrivare ad istruire il classificatore. Infatti, la fase che richiede più tempo è la creazione del *training set*, che prevede l'assegnazione manuale di un'etichetta ad ogni testo che compone questo dataset. Per semplificare e velocizzare questo processo, con il loro approccio viene chiesto all'utente di fornire alcune parole significative per ogni classe, le quali serviranno per attribuire una delle classi disponibili ai testi che andranno a formare il *training set*. In questo modo, l'utente non deve assegnare un'etichetta ad ogni testo, ma individuare per ciascuna di esse alcune parole in modo che possano essere attribuite ai singoli testi in automatico. Questo porta ad una significativa riduzione di tempo, in particolare quando si dispone di un numero elevato di documenti.

1.4 Criticità

Quanto esposto nei paragrafi precedenti mostra i numerosi vantaggi e utilità della *Sentiment Analysis*, ai quali si contrappongono però alcune criticità che devono essere prese in considerazione. Oltre alle problematiche date dalla scelta dell'approccio (si veda Sezione 1.3) numerose difficoltà si incontrano proprio nella determinazione della polarità del sentimento e nel riconoscimento dell'emozione.

Tra le problematiche più frequenti troviamo:

- Ironia e sarcasmo;
- Negazioni;
- Ambiguità di alcune parole;
- Multipolarità.

A cominciare dall'*ironia* e dal *sarcasmo*, questi costituiscono un ostacolo non di poca importanza per la *Sentiment Analysis*. Infatti, a meno che gli algoritmi non siano studiati per riconoscerli, cosa assai complessa e rara, è difficile che vengano individuati ed interpretati correttamente. Il problema è dato non solo dal fatto che vengono usate parole positive per esprimere opinioni negative e viceversa, ma anche dal fatto che non ci sono termini "specifici" o limitati, quindi è difficile istruire un classificatore. Per trattare tale problematica, si possono utilizzare approcci differenti per l'identificazione automatica tra i quali troviamo: metodi *Rule-Based*, metodi statistici, algoritmi di *Machine Learning* e di *Deep Learning*. In particolare, il *Deep Learning* sta riscontrando sempre maggiore popolarità, soprattutto per quanto riguarda le *Deep Neural Networks*: reti neurali ancora più complesse di quelle illustrate nella Sezione 1.3.2.

A loro volta anche le *negazioni* sono insidiose da trattare, in quanto bisogna determinare se la negazione si riferisce all'intera frase oppure solo ad una o più parole. Ad esempio, nell'espressione «*Il documentario non era interessante*» la negazione si riferisce solo alla parola che la segue (cioè "interessante"); nel caso invece della frase «*Non chiamerei questo film una commedia*» l'effetto del "non" permane fino alla fine della frase. In aggiunta, esistono diverse tipologie oltre alla negazione per eccellenza composta utilizzando il "non". Infatti, possono essere utilizzati anche prefissi o suffissi (a seconda della lingua) – ad esempio completo-incompleto – oppure la negazione può essere implicita, nel senso che la frase in sé esprime un concetto negativo. L'approccio più semplice che viene

utilizzato per trattare le negazioni, consiste nell'attribuire polarità negativa a tutte le parole che seguono una negazione, fino al segno di punteggiatura successivo.

Per *ambiguità* di alcune parole, si intende il fatto che alcuni termini cambiano significato a seconda del contesto, quindi la polarità – positiva o negativa – non può essere definita a priori, considerando individualmente la parola, ma bisogna tenere conto dell'intera frase. Per risolvere il problema, si può attribuire la polarità tenendo conto del contesto in cui viene impiegato il termine, quindi non a priori, oppure si sceglie di assegnare alla parola ambigua una polarità neutra.

Infine, la *multipolarità* si manifesta qualora in un testo vengano trattate più entità o più aspetti per i quali l'autore ha opinioni differenti. Quindi, se si considera il testo nella sua interezza – si veda Sezione 1.1 – si ottiene una polarità complessiva che trascura quella relativa ai singoli aspetti, fornendo un'interpretazione non propriamente corretta dell'opinione dell'*opinion holder*. Per questo motivo, è importante estrarre tutte le entità o gli aspetti della singola frase, con la rispettiva polarità e sentimento, e calcolare la polarità complessiva solo se necessario (Eremyan, 2018).

Capitolo 2

Emotion Detection Sentiment Analysis

Come descritto nella Sezione 1.2.3 l'*Emotion Detection* si concentra sull'identificazione del sentimento espresso nel testo, non solo in termini di positività o negatività, ma soprattutto di emozione.

Per poter raggiungere tale obiettivo vengono utilizzati due importanti strumenti, i quali consentono di classificare i singoli testi: i *metodi di clustering* e i *lexicons*. I primi permettono di raggruppare in modo non supervisionato i testi oggetto di interesse, in gruppi che siano più omogenei possibile al loro interno. Si tratta di una *classificazione non supervisionata* che si distingue da quella supervisionata per l'assenza di un training set, cioè un dataset impiegato per istruire un classificatore. I *lexicons* sono, invece, dei dizionari che vengono impiegati per attribuire un'emozione ad ogni singolo termine presente nel testo.

2.1 Metodi di clustering

I *metodi di clustering* sono tecniche di apprendimento automatico che vengono utilizzati in vari ambiti di analisi, tra cui la *Sentiment Analysis*. Sono metodi di tipo *non supervisionato*, poiché non prevedono l'uso di dati già etichettati per istruire l'algoritmo. Tali tecniche hanno come obiettivo quello di raggruppare gli oggetti in classi che siano il più omogenee possibile. Ogni classe costituisce un *cluster*, cioè un insieme di oggetti simili tra loro, ma che presentano dissimilarità rispetto agli elementi appartenenti alle altre classi (Data Skills, 2015). Infatti, per ottenere dei buoni *clusters* si cerca di minimizzare la varianza all'interno dei gruppi, in modo che gli oggetti siano simili tra loro, mentre si vuole massimizzare la varianza tra i gruppi, in modo che due elementi appartenenti a due classi diverse siano il più possibile diversi tra loro. In genere, un algoritmo di clustering riceve in input una serie di elementi e restituisce come output un certo numero di

clusters. Il raggruppamento degli oggetti nelle classi, viene eseguito dall'algoritmo in base ad una *misura di similarità* (o *dissimilarità*), scelta dal ricercatore, che serve per calcolare la *distanza* a livello di similitudine tra gli elementi disponibili. Di seguito sono riportate le più usate.

- Distanza euclidea:
$$D(d_i, d_j) = \sqrt{\sum_{k=1}^{n} (d_{ik} - d_{jk})^2}$$
- Distanza coseno:
$$D(d_i, d_j) = \frac{\sum_{k=1}^{n} d_{ik} d_{jk}}{\sqrt{\sum_{k=1}^{n} d_{ik}^2 \sum_{k=1}^{n} d_{jk}^2}}$$
- Distanza di Manhattan:
$$D(d_i, d_j) = \sum_{k=1}^{n} \left| d_{ik} - d_{jk} \right|$$

dove d_i e d_j rappresentano due documenti o frasi, mentre d_{ik} e d_{jk} rappresentano i singoli termini che li compongono.

Gli algoritmi di clustering si dividono i due tipologie principali che vengono descritte nelle sezioni seguenti.

2.1.1 Algoritmi di clustering gerarchico

Gli *algoritmi gerarchici* producono una gerarchia di clusters che può essere rappresentata attraverso un grafico chiamato *dendogramma* (Figura 2.1). In questo tipo di grafico, la similarità tra due oggetti è rappresentata dall'altezza del più basso *nodo* interno che condividono. Quindi, ad esempio, i due oggetti di sinistra sono più simili tra loro rispetto a quanto lo siano i due oggetti di destra.

La gerarchia rappresentata può essere costruita attraverso due approcci differenti. L'approccio *bottom-up* (o agglomerativo) procede dal basso verso l'alto per unire clusters più piccoli, a partire da elementi singoli, in gruppi di dimensioni maggiori, unendo i gruppi che si trovano più vicini tra loro. Al contrario, l'approccio *top-down* (o divisivo) considera inizialmente un unico cluster che comprende tutti gli elementi disponibili, che in seguito viene diviso progressivamente in clusters di dimensioni inferiori, fino ad arrivare ad un numero di gruppi pari al numero degli oggetti. In entrambi gli approcci, per

individuare il numero di clusters bisogna scegliere a quale livello eseguire un taglio del grafico, cioè a quale distanza. Nella Figura 2.1, la linea rossa rappresenta il livello di taglio ed identifica tre clusters, poiché interseca tre linee del dendogramma.

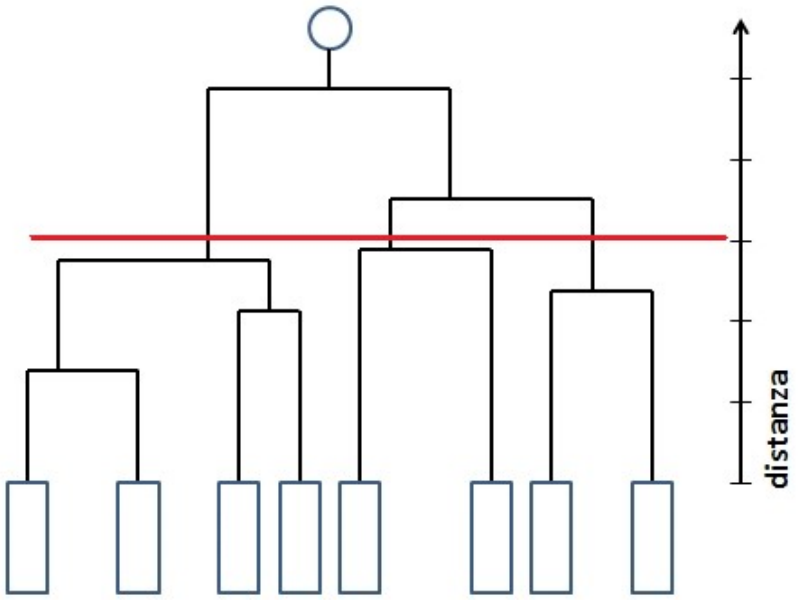

Figura 2.1: Esempio di grafico dendogramma

Calcolare la distanza tra due singoli elementi può essere un'operazione semplice, per la quale si possono impiegare le misure di distanza elencate precedentemente. Questo calcolo si complica nel momento in cui la distanza deve essere calcolata tra un oggetto e un cluster oppure tra due clusters. In questi casi, infatti, data la presenza di più elementi, è necessario definire cosa si intenda per distanza e soprattutto tra quali oggetti vada calcolata. In tale senso, si hanno a disposizione diverse possibilità di cui vengono esposte le quattro più comuni (Figura 2.2).

Single linkage: noto anche come tecnica del *vicino più vicino*, è il metodo più semplice. La distanza tra due clusters è definita come la minore distanza tra una coppia di elementi, dove la coppia si intende formata da un solo elemento per gruppo. In altre parole, si considera la distanza tra i due oggetti più vicini, appartenenti a gruppi diversi.

Complete linkage: la distanza tra due gruppi viene calcolata in modo opposto rispetto al precedente. Infatti, si considera la distanza maggiore tra una

coppia di oggetti appartenenti a clusters diversi, cioè quella tra gli elementi più lontani.

Average linkage: secondo questo metodo, la distanza tra due gruppi è determinata dalla media delle distanze tra tutte le possibili coppie di oggetti appartenenti ai due clusters.

Ward's linkage: Ward (1963) ha proposto una procedura di clustering che permette di formare le partizioni in modo che sia minimizzata la perdita di informazioni. Come criterio per determinare la perdita, viene presa in considerazione la devianza intra-gruppo. Ad ogni passo dell'algoritmo, vengono quindi uniti i due clusters dalla cui unione deriva il minor aumento di tale devianza (FrontlineSolvers, 2012).

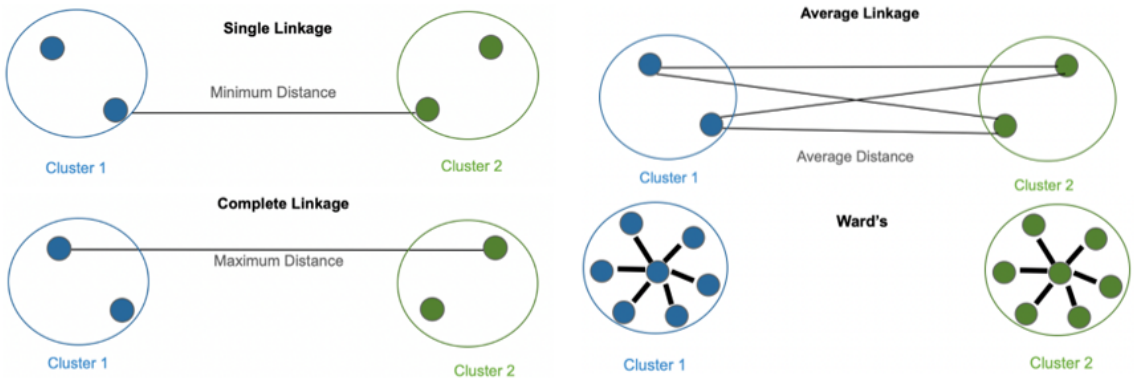

Figura 2.2: Rappresentazione dei possibili linkage

2.1.2 Algoritmi di clustering partizionale

Gli *algoritmi partizionali* distribuiscono le osservazioni nei vari clusters sulla base di regole precise. Inoltre, poiché l'output dell'algoritmo è un solo set di clusters, all'utente viene richiesto di dichiarare il numero desiderato di clusters prima di eseguire l'elaborazione.

Uno dei metodi partizionali più usati è l'algoritmo *K-means*, che permette di partizionare gli elementi in K gruppi con l'obiettivo di minimizzare la varianza intra-gruppo. Si tratta di un algoritmo semplice, ma molto performante, basato sui *centroidi*. Un *centroide* è un punto, appartenente allo spazio degli oggetti disponibili, che media le distanze di tutti gli elementi appartenenti ad un certo cluster. In altri termini, si può dire che esso costituisce una sorta di *baricentro* del

cluster e non è uno degli elementi che costituiscono le osservazioni – anche se in alcuni casi può capitare che coincidano.

L'algoritmo *K-means* è una procedura di tipo iterativo che segue le fasi seguenti:

1. Dato che, inizialmente, non si conoscono i gruppi del dataset in studio, il ricercatore deve scegliere il numero K di clusters che intende identificare (in genere è il primo step per tutti i metodi partizionali).

2. I centroidi, appartenenti allo spazio delle osservazioni, vengono identificati in modo casuale ed è importante che siano sufficientemente distanti tra loro, altrimenti l'algoritmo potrebbe non giungere a convergenza. La scelta iniziale dei centroidi è fondamentale perché da essa dipende la convergenza o non covergenza dell'algoritmo, ma anche la velocità con cui questa viene raggiunta (in termini di numero di iterazioni).

3. Viene calcolata la distanza di ogni punto del dataset da ogni centroide e ciascun elemento viene assegnato al cluster relativo al centroide più vicino. In questo modo si identificano le composizioni dei clusters.

4. Si ridefinisce il centroide di ogni gruppo, in base alla nuova composizione individuata, calcolando la media delle posizioni occupate dagli oggetti del cluster.

5. Si itera dallo step 3 fino al raggiungimento della convergenza. La convergenza può essere raggiunta quando non si osservano variazioni nelle posizioni dei centroidi oppure quando è stato raggiunto il numero massimo di iterazioni, qualora sia stato fissato.

La scelta del numero di clusters può mettere in difficoltà il ricercatore, poiché non sempre si riesce a cogliere dai dati il numero ottimale. Per facilitare questa scelta è stato pensato l'*elbow method* – *metodo del gomito* – che permette di testare un diverso numero di clusters per poi individuare, in modo oggettivo, quale sia il più adatto. Nello specifico, si itera l'algoritmo *K-means* per diversi valori di K e, per ognuno di essi, si calcola la somma delle distanze al quadrato tra ogni centroide e gli elementi del cluster corrispondente. I valori ottenuti vengono rappresentati in un grafico bidimensionale, dove sull'asse X sono riportati i valori di K, mentre sull'asse Y le distanze (Figura 2.3). Il grafico viene letto da destra verso sinistra, fino ad individuare il valore di K in corrispondenza del quale

l'andamento del grafico subisce una variazione significativa, cioè si nota un aumento repentino dei valori delle distanze. Nel grafico riportato, ad esempio, si osserva che procedendo dal valore 9 al valore 3 di K l'andamento è crescente in modo quasi lineare. A partire dal valore 3, invece, si nota che la curva cresce più velocemente. Quindi si può identificare in K=3 il *"gomito"* del grafico e, di conseguenza il numero ottimale di clusters sarà pari a tre (Casadei, 2019a).

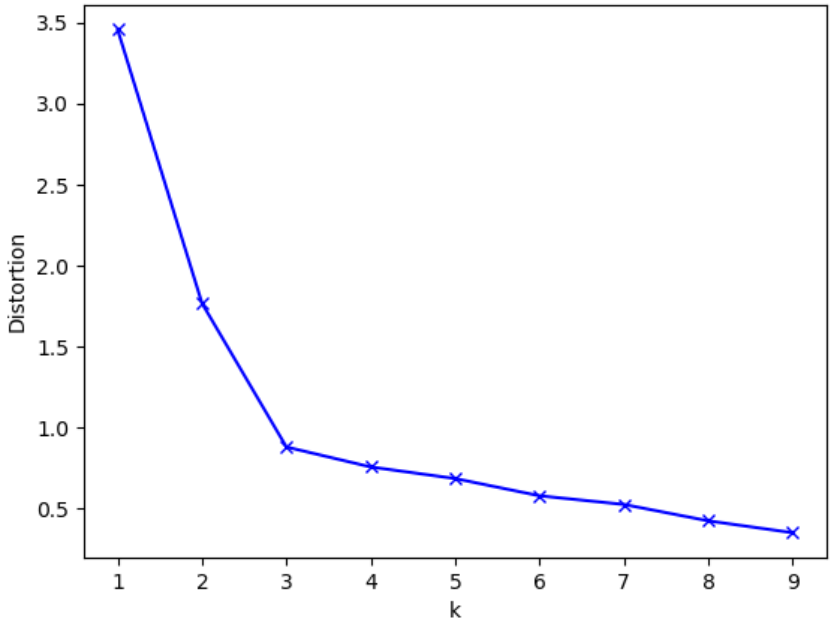

Figura 2.3: Valore di K ottimale secondo l'*elbow method*

In alternativa all'algoritmo *K-means* si può utilizzare l'algoritmo *K-medoids*, in cui i centroidi sono sostituiti dai *medoidi*. Un *medoide* è un'osservazione del dataset in esame, per la quale la dissimilarità media con tutte le altre osservazioni appartenenti al cluster è minima. Quindi, è l'elemento più simile a tutti gli altri appartenenti al proprio cluster. Da questa definizione si osserva che questo algoritmo è strettamente legato a quello precedente, infatti, le fasi da compiere sono le stesse. Verranno quindi scelti casualmente i K medoidi e verranno assegnati gli elementi ai clusters. Poi, all'interno di ogni gruppo, si calcoleranno le distanze di ogni possibile medoide da tutti gli altri punti per individuare il nuovo medoide. Infine, si riassegneranno tutte le osservazioni in base ai nuovi medoidi e si ripeterà la procedura fino al raggiungimento della convergenza.

Il vantaggio dell'algoritmo *K-medoids* risiede proprio nella definizione di medoide. Trattandosi di un'osservazione del dataset, a differenza di quanto accade per la determinazione dei centroidi che richiede il calcolo di una media,

l'individuazione dei medoidi non è influenzata dall'eventuale presenza di outliers e questo rende il metodo più robusto. Inoltre, nell'algoritmo *K-means* si predilige l'uso della distanza Euclidea, mentre per il *K-medoids* possono essere scelte altre metriche di distanza, come la distanza coseno o altre distanze (Analystics India Mag, 2021).

Per valutare la bontà della tecnica di clustering applicata si possono usare diverse metriche, tra cui una delle più robuste ed usate è la *silhouette*. Con questo metodo, viene cacolato un *coefficiente di silhouette* per ogni osservazione del dataset che misura quanto un elemento è simile al proprio cluster in confronto ad altri clusters. Il valore di questa metrica è compreso nell'intervallo $[-1, 1]$, dove un valore alto indica che quell'oggetto è ben classificato nel proprio cluster mentre non è simile agli elementi degli altri. Al contrario, un valore basso indica che l'elemento non si inserisce bene nel proprio cluster e potrebbe essere stato misclassificato. Infine, un valore prossimo allo 0 indica che il punto si trova al confine della banda di decisione e potrebbe essere inserito in entrambi i clusters, senza che venga commesso un errore significativo. Quindi, se la maggior parte delle osservazioni del dataset presentano un valore alto, allora la configurazione del clustering è appropriata e ben eseguita. Se, invece, le osservazioni hanno in gran parte valori negativi, potrebbe essere stato scelto un numero troppo elevato o troppo basso di clusters.

Per calcolare il *coefficiente di silhouette* di un'osservazione *i-esima* si procede nel modo seguente:

1. Si calcola la distanza media dell'osservazione *i-esima* da tutte le altre osservazioni presenti nel clustering in cui è inserita. Indichiamo tale distanza con $a(i)$.

2. Si calcola la distanza media dell'osservazione *i-esima* da tutte le altre osservazioni presenti nel cluster più vicino a quello a cui essa appartiene. Indichiamo tale distanza con $b(i)$.

3. Si calcola il *coefficiente di silhouette*, indicato con $s(i)$, per l'osservazione *i-esima* utilizzando la formula seguente:

$$s(i) = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}}$$

Dopo aver calcolato il *coefficiente di silhouette* per ogni osservazione, si calcola la loro media per ottenere il punteggio complessivo di silhouette.

A livello grafico, si possono rappresentare i coefficienti delle singole osservazioni, raggruppati in base al cluster di appartenenza, e osservare se i loro punteggi sono inferiori alla silhouette complessiva, cioè quella media. In caso affermativo significa che il clustering non è soddisfacente e che il numero di gruppi deve essere variato (Figura 2.4). Inoltre, è possibile individuare il numero di clusters ottimale rappresentando l'andamento del punteggio totale della silhouette, in modo simile a quanto riportato per l'*elbow method*. Tale grafico si può ottenere riportando sull'asse X i possibili valori di K e sull'asse Y i valori della silhouette ottenuti in base al numero di clusters. Il numero di gruppi ottimale si individua in corrispondenza del più alto valore di silhouette all'interno del grafico (Kumar, 2020).

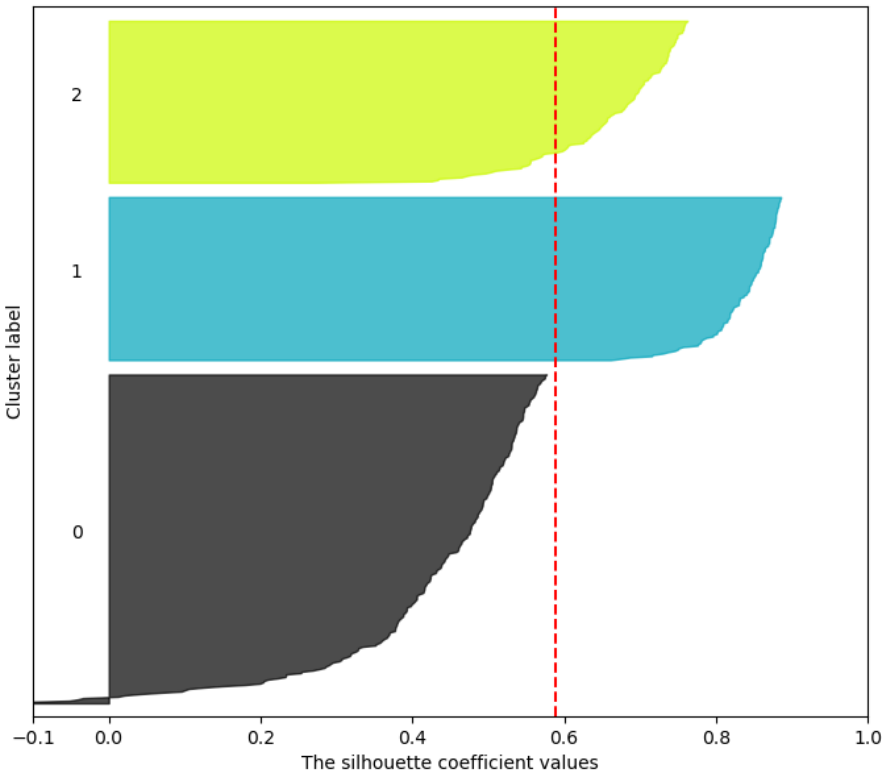

Figura 2.4: Esempio di rappresentazione dei coefficienti di silhouette

2.2 Lexicons

Gli esseri umani, nell'approcciarsi ad un testo, impiegano le proprie capacità di comprensione per capire l'intento emotivo delle parole, per dedurre se una determinata sezione del testo è positiva o negativa, oppure caratterizzata da

qualche emozione più specifica. Nella *Sentiment Analysis* possiamo usare alcuni strumenti per avvicinarci al contenuto emotivo del testo, attraverso tecniche di programmazione. Un modo possibile è quello di considerare un testo come una combinazione di singoli termini. Il sentimento dell'intero documento sarà, quindi, dato dalla somma del contenuto emotivo di ogni parola (Silge & Robinson, 2017).

Uno strumento molto usato, seppur non l'unico disponibile, è costituito dai *lexicons*. Si tratta di dizionari contenenti informazioni sia a livello semantico sia grammaticale, relative a singoli termini (*uni-grammi*) o insiemi di parole (*n-grammi*). La differenza rispetto ad un dizionario tradizionale è che un lexicon, in genere, non presenta la descrizione del termine, la pronuncia e altri elementi, come i sinonimi e i contrari (Guthrie et al., 1996).

In base all'interesse del ricercatore, esistono diversi tipi di lexicons. Dal momento che la polarità di un termine dipende, in molti casi, dal contesto in cui questo è inserito, si possono utilizzare *domain-specific lexicons*, che forniscono buone performance in quanto specifici per il contesto di interesse. Tuttavia, potrebbe non essere già disponibile un lexicon appropriato, quindi si rende necessario generarlo manualmente. Questa procedura non è semplice, soprattutto se si disponde di un solo *corpus* per il contesto considerato. Per questo esistono *general-purpose lexicons* (chiamati anche *general-domain*), cioè dizionari all'interno dei quali la polarità delle parole non è specifica per un contesto in particolare, ma è relativa al significato più generico possibile del termine (Muhammad et al., 2020).

Prendiamo in considerazione alcuni tra i *general-purpose lexicons* più usati nella *Sentiment Analysis*: AFINN, Bing e NRC. Questi lexicons sono tutti basati su *uni-grammi* in lingua inglese e sono stati creati usando il *crowdsourcing* o da uno degli autori. In seguito, sono stati validati utilizzando una combinazione di *crowdsourcing*, recensioni o dati di Twitter (Silge & Robinson, 2017).

- Il **lexicon AFINN** assegna alle parole un punteggio compreso tra -5 e 5, dove un numero negativo indica un sentimento negativo e un punteggio positivo indica polarità positiva. È il dizionario maggiormente utilizzato e, attualmente, conta più di tremila termini.

- Il **lexicon Bing** compie una classificazione binaria dei termini, con modalità positiva o negativa per la polarità. A differenza dell'AFINN non è quindi possibile distinguere tra diversi livelli di polarità.

- Il **lexicon NRC** categorizza il sentimento in modo binario ("si"/"no") per diverse modalità possibili: positivo, negativo, rabbia, aspettativa, disgusto, paura, gioia, tristezza, sorpresa e fiducia.

Nell'articolo di Kim (2018) vengono evidenziati alcuni limiti di questi lexicons. La ricerca è stata concentrata sulle parole utilizzate da Charles Dickens nelle sue opere, piuttosto che considerare le relazioni tra le parole all'interno delle frasi in quanto, qualora vengano impiegati strumenti e tecniche di text mining per analizzare testi letterari, risulta più semplice. L'analisi di confronto dei tre lexicon appena illustrati ha evidenziato alcuni limiti.

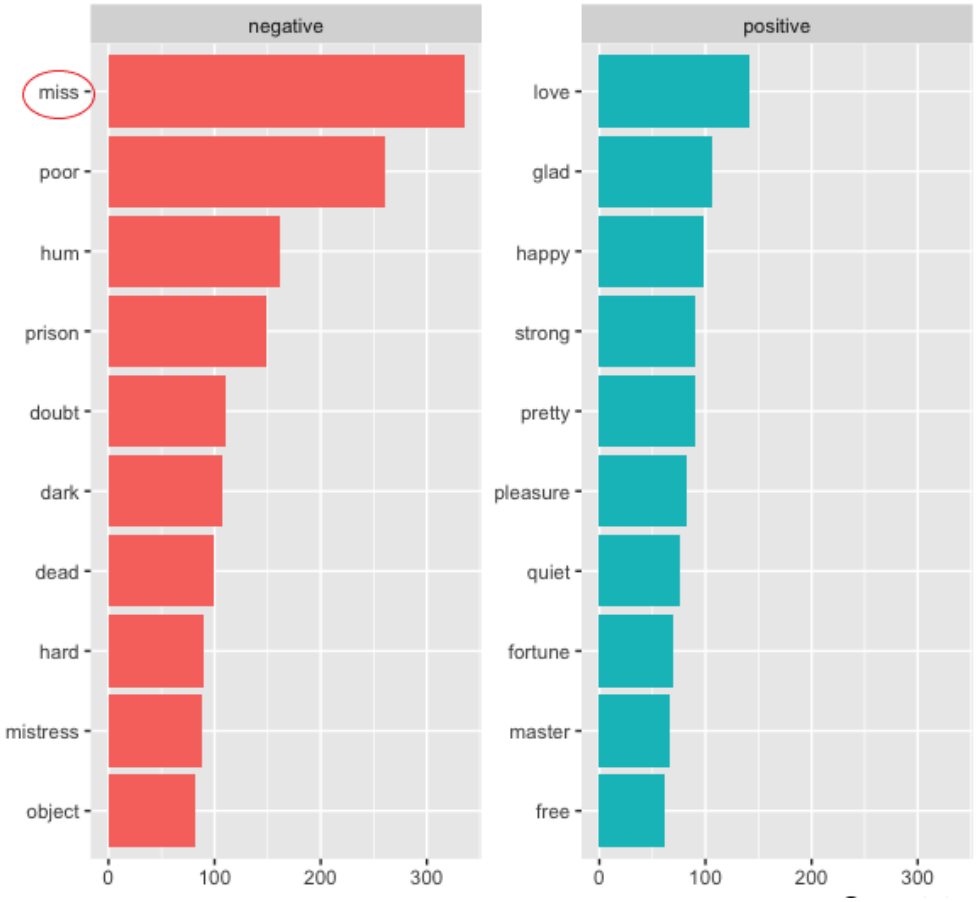

Figura 2.5: Classificazione secondo i lexicon AFINN e Bing

Ad esempio, i lexicons *AFINN* e *Bing* percepiscono la parola *miss* come negativa, considerandola come un verbo. Tuttavia, nella lingua inglese tale termine viene usato come titolo femminile per riferirsi ad una donna non sposata e, poiché molto frequente nei testi letterari, la sua classificazione come negativo porta ad una *Sentiment Analysis* non corretta (Figura 2.5). Osservando la figura

riportata, appare evidente che in questo modo si avrebbe nel complesso un maggior numero di termini negativi, a causa dell'elevato uso di questa parola, con conseguente misclassificazione. Ciò accade perché i lexicons *AFINN* e *Bing* non riconoscono la classificazione di genere. Il lexicon *NRC*, invece, non classifica il termine *miss* come negativo, compiendo una migliore classificazione.

Per quanto riguarda il lexicon *NRC* è bene sottolineare che al suo interno, molti termini compaiono più di una volta, in quanto ognuno può appartenere a più di una delle classi possibili. Questo porta, inevitabilmente, ad alcune difficoltà nella *Sentiment Analysis*. Ad esempio, sempre considerando le opere di Dickens, il termine *mother* viene inserito nelle categorie *gioia*, *tristezza* e *negativo*. In questo caso, la classificazione più corretta viene compiuta dagli altri due lexicons, *AFINN* e *Bing*, perché non considerano questa parola, in quanto non si tratta di un termine emotivo (Kim, 2018).

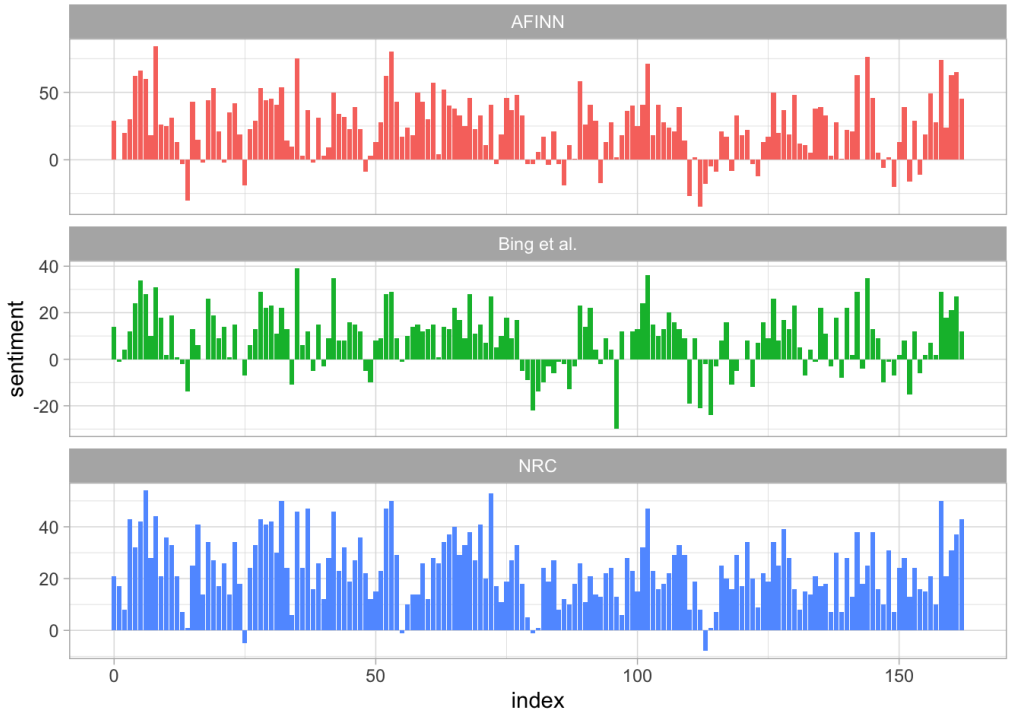

Figura 2.6: Confronto tra i tre lexicons applicati a *Orgoglio e Pregiudizio*

Un ulteriore confronto tra i tre lexicons è stato eseguito da Silge & Robinson (2017), applicandoli ad *Orgoglio e Pregiudizio* di Jane Austen. Il testo è stato suddiviso in sezioni e per ognuna di esse è stato calcolato il sentimento complessivo, in termini di positività e negatività. Come si può notare dalla Figura 2.6, si osservano alcune differenze. L'andamento del sentimento sembra essere simile nel corso del

romanzo, se si osservano i picchi e i cali, anche se risulta sensibilmente diverso in termini di valori assoluti. Infatti, il lexicon *AFINN* assegna valori più alti in termini assoluti, con molti più positivi, mentre il lexicon *Bing* assegna valori più bassi e sembrano esserci gruppi di sezioni positive (o negative) di maggiori dimensioni – come si può notare dalle sequenze di colonne contigue sopra (o sotto) lo zero. Il lexicon *NRC*, invece, si distingue in modo più evidente dagli altri due. Infatti, il testo nel complesso è classificato positivamente, anche se si può comunque cogliere un andamento simile di picchi e cali relativi.

Differenze simili a quelle osservate per l'opera *Orgoglio e Pregiudizio* possono essere osservate anche applicando i lexicons ad altre opere letterarie. In ogni caso, risulta che il lexicon *NRC* assegna un sentimento più positivo rispetto agli altri, il lexicon *AFINN* presenta varianza maggiore, mentre il lexicon *Bing* individua tratti di testo più lunghi, che siano simili tra loro in termini di sentimento. In generale, la tendenza del sentimento in un testo letterario presenta un andamento simile indipendentemente dal lexicon scelto (Silge & Robinson, 2017).

Capitolo 3

Applicazione Sentiment Analysis

La situazione pandemica attuale è dovuta ad un nuovo ceppo della famiglia dei coronavirus che, a dicembre 2019 in Cina, ha compiuto il cosiddetto passaggio di specie. È infatti noto che i virus appartenenti a questa famiglia circolano generalmente tra gli animali, ma è possibile che in determinate circostanze possano essere trasmessi all'uomo. In passato è già accaduto, come nel caso della SARS (SARS-CoV-1) – trasmessa all'uomo dallo zibetto alla fine del 2002 in Cina – che ha causato circa 8 000 casi in 33 paesi in soli otto mesi. Il nuovo coronavirus, denominato "COVID-19" (SARS-CoV-2), appartiene alla stessa famiglia di virus della SARS. I due virus causano in coloro che contraggono la malattia una sintomatologia simile, che interessa il sistema respiratorio e viene definita come Sindrome Respiratoria Acuta Grave (Ministero della Salute, 2020).

La diffusione ha avuto origine a partire dalla città di Wuhan (Cina), dove nel mese di dicembre del 2019 sono stati registrati alcuni casi di polmonite di eziologia sconosciuta, comunicati all'Organizzazione Mondiale della Sanità (OMS) il 31 dicembre. Nei giorni successivi, è stata isolata nei pazienti ricoverati la sequenza genetica del nuovo coronavirus, condivisa in seguito con altri paesi per lo sviluppo di test diagnostici specifici. Il COVID-19 ha mostrato da subito una maggiore velocità di trasmissione e quindi una maggiore diffusione, rispetto alla SARS, tanto che l'11 marzo 2020 il direttore generale dell'OMS ha dichiarato che si era giunti allo stato di pandemia globale.

L'elevata velocità di trasmissione e il considerevole numero di morti – ad oggi più di quattro milioni nel mondo – hanno reso chiara la necessità di formulare un vaccino nel minor tempo possibile. A tale scopo è stato creato l'Access COVID-19 Tool Accelerator (ACT-Accelerator), una collaborazione globale lanciata dall'OMS

e dai suoi partner per accelerare lo sviluppo, la produzione e l'accesso equo a test, trattamenti e vaccini per il COVID-19 (WHO, 2020).

Grazie al coordinamento permesso dall'ACT-Accelerator e alla condivisione di informazioni tra gli stati, è stato possibile velocizzare la ricerca e lo sviluppo dei vaccini che ha portato tra la fine del 2020 e l'inizio del 2021 all'approvazione, da parte di vari enti, dei primi vaccini contro il COVID-19: Pfizer/BioNThec, Moderna e AstraZeneca.

Le rapide tempistiche di sviluppo dei vaccini hanno generato perplessità ed incertezza nella popolazione. Una volta giunto il momento dell'entrata in commercio dei vaccini e dell'inizio delle vaccinazioni, in molti si sono domandati se fossero sicuri e se fossero stati testati nel modo adeguato. I dubbi e i timori sono sorti anche a causa del fatto che i vaccini Pfizer/BioNTech e Moderna utilizzano una nuova tecnologia a RNA messaggero (mRNA), che permette di inoculare una sequenza genetica con le istruzioni per produrre l'antigene.

Ulteriori preoccupazioni si sono diffuse quando hanno iniziato a presentarsi alcune controversie in merito all'uso del vaccino AstraZeneca. Nei Paesi ad alto reddito, inzialmente era stato approvato per tutte le fasce d'età a partire dai 18 anni, anche se in alcuni Stati veniva somministrato solo a coloro che avevano meno di 55-60 anni, poiché mancavano dati relativi alla popolazione più anziana. In seguito, è stata estesa l'inoculazione anche alle fasce d'età più alte, per poi essere limitata solo agli anziani dai 65 anni in su. Inoltre, nel corso dei mesi si sono verificate sospesioni dell'uso di questo vaccino in diversi Paesi, a causa di alcuni eventi trombotici verificatisi, in particolare, in donne giovani. Nonostante eventi avversi di questo tipo si siano verificati anche in seguito all'uso di altri vaccini, si è diffuso un certo timore da parte della popolazione nei confronti del vaccino AstraZeneca, a cui hanno contribuito i mass media evidenziando gli eventi accaduti.

La poca informazione sulle nuove tecnologie e la paura di eventuali eventi avversi, hanno diffuso la convinzione che i vaccini siano troppo sperimentali e che non siano stati testati a sufficienza. A questi pensieri si contrappone l'opinione di coloro che hanno fiducia nella scienza e vedono nei vaccini l'unica soluzione possibile per risolvere, almeno in parte, la situazione. Queste idee, contrastanti tra loro, sono facilmente osservabili in particolare sui *social networks* dove milioni di persone di tutto il mondo hanno la possibilità di condividere il proprio parere.

Nel presente capitolo viene proposta un'applicazione di *Sentiment Analysis*, svolta analizzando il social network *Twitter*, da cui è stato estratto un considere-

vole numero di *post* relativi ai vaccini contro il COVID-19, per cercare di cogliere il sentimento complessivo diffuso tra gli utenti.

3.1 Materiali e Metodi

3.1.1 Web scraping e descrizione dei datasets

Per lo svolgimento dell'analisi appena presentata, è stato preso in considerazione uno dei social network più utilizzati: Twitter. Si tratta di un sito di *microblogging* che permette agli utenti di postare dei messaggi noti come *tweet* che possono includere immagini, video e collegamenti a siti web tramite URL. Inoltre, le persone possono entrare in contatto tra loro seguendo altri utenti per visualizzare i contenuti da essi pubblicati nella propria *homepage*. Molte attività su Twitter implicano l'uso di *hashtag* (composizioni di termini precedute dal simbolo #) che permettono di aggregare i tweet relativi allo stesso argomento (WebWise, 2018).

Le informazioni impiegate nella presente analisi sono raccolte in quattro diversi datasets e sono costituite dai tweet contenenti hashtag relativi ai vaccini. Nello specifico, sono stati considerati i seguenti:

- #covidvaccine
- #pfizer
- #moderna
- #astrazeneca

Il primo hashtag è più generico dei successivi, in quanto si è ritenuto utile avere uno sguardo complessivo sull'opinione in merito ai vaccini e, in seguito, approfondire il parere relativo ad alcuni dei vaccini più utilizzati. Per quanto riguarda la lingua è stata scelta quella inglese, perché permette di raccogliere informazioni provenienti da tutto il mondo.

Il *web scraping* dei dati – l'estrazione di informazioni non strutturate da siti web – è stato eseguito utilizzando il linguaggio di programmazione `Python` che offre il pacchetto `snscrape`, con il quale si possono scaricare informazioni da alcuni dei social networks più utilizzati quali: Facebook, Twitter, Instagram, Reddit e Telegram.

I quattro dataset scaricati contengono variabili associate ai tweet tra cui: data e ora, contenuto del tweet, URL del tweet, informazioni relative all'utente,

numero di like, numero di commenti, numero di re-tweet, lingua, eventuali URL contenuti nel tweet, hashtag, utenti menzionati, media contenuti nel post e relativo URL. Visto l'obiettivo dell'analisi, è stata effettuata una prima pulizia dei dati utilizzando Excel e sono state mantenute le informazioni relative a: *Data e Ora, Contenuto* e *Lingua*. La variabile relativa alla lingua è stata utilizzata per filtrare e selezionare solo i tweet in lingua inglese e, poi, è stata eliminata in quanto non necessaria ai fini dell'analisi. In seguito, i datasets sono stati caricati in R dove la variabile *Data e Ora* è stata divisa in due diverse variabili, di cui è stata mantenuta solo l'informazione sulla data di pubblicazione del tweet.

Come risultato, sono stati ottenuti quattro dataset costituiti da due variabili – *Data* e *Contenuto* – , per il periodo compreso tra il 1 dicembre 2020 e il 24 giugno 2021, contenenti il seguente numero di osservazioni:

- #covidvaccine: 497 644 osservazioni
- #pfizer: 110 820 osservazioni
- #moderna: 60 068 osservazioni
- #astrazeneca: 120 685 osservazioni

3.1.2 Pre-processing

Una volta ottenuti i datasets definitivi, è stato svolto il *pre-processing* dei dati: una fase fondamentale nella *Sentiment Analysis* necessaria per pulire e organizzare al meglio i dati, in modo che siano più facilmente trattabili ai fini dei propri obiettivi. Lo scopo è quello di ottenere dei dati in una forma *tidy*, cioè ordinata, in quanto le informazioni raccolte sul web sono in genere *non strutturate* (se ne è parlato all'inizio del Capitolo 1). Nel caso in esame, trattando dati di natura testuale, è stato creato un *corpus*, vale a dire una raccolta di testi costituiti dai singoli tweet, contraddistinti da un identificativo numerico e corredati da metadati relativi alla data.

Successivamente alla creazione del *corpus*, sono state eseguite ulteriori operazioni di pulizia dei dati per renderli più facilmente trattabili, tra cui: conversione dei caratteri da maiuscoli a minuscoli, rimozione degli URL, rimozione dei Tag degli utenti (riconosciuti dal simbolo @), rimozione degli Hashtag (riconosciuti dal simbolo #), rimozione di simboli particolari (es. ™ ‡ ‰ ¤), rimozione di numeri e punteggiatura (es. " , . ; ? !). In seguito a queste operazioni, è stata eseguita la rimozione delle *stopwords*, cioè di tutti quei termini che ricorrono

frequentemente nel testo, ma che non sono significativi a livello semantico. Fanno parte di questa categoria di parole, ad esempio, gli articoli, le congiunzioni, le preposizioni e anche alcuni verbi (es. *have, be, get*). A questi si aggiungono ulteriori termini specifici per il contesto in esame; in questo caso sono stati rimossi i nomi di alcuni stati, come ad esempio *India*, in quanto molto ricorrenti all'interno dei tweet.

Terminate le fasi di pulizia, per ogni dataset sono state salvate due copie del corpus. Sulla prima è stato eseguito lo *stemming* dei tweet, una procedura che permette di ricondurre i termini alla propria radice in modo che, ad esempio, la forma singolare e plurale della stessa parola oppure le coniugazioni dello stesso verbo vengano considerate come uno stesso termine. Questa copia del corpus è stata in seguito utilizzata per l'applicazione del clustering, che risulta più efficiente utilizzando termini riportati alla propria radice. La seconda copia è stata, invece, mantenuta invariata ed è stata usata per svolgere un'analisi generale dei sentimenti, in quanto sono state impiegate funzioni che utilizzano dizionari contenenti diverse forme dei termini e non solo la loro forma base.

Come ultima fase di preparazione dei dati, utilizzando il corpus sul quale è stato eseguito lo stemming, è stata costruita la *Term-Document Matrix* (TDM). Si tratta di una matrice che presenta sulle colonne gli identificativi dei documenti che compongono il corpus – in questo caso dei tweet – e sulle righe le singole parole contenute nei documenti. All'interno della matrice si trovano le *term frequency*, cioè le frequenze di apparizione di ogni termine nei documenti; in questo caso sono state considerate le frequenze assolute.

3.1.3 Metodologia e lexicon scelti

Dato l'obiettivo dell'analisi di studiare il sentimento e l'opinione degli utenti di Twitter in merito ai vaccini contro il COVID-19, è stato scelto di utilizzare la metodologia dell'*Emotion Detection*, in quanto più adatta al contesto. Infatti, le altre metodologie – si veda Sezione 1.2 – non sono adeguate in questo caso, poiché la *Fine-Grained* è particolarmente indicata nel caso in cui si trattino recensioni basate su un voto, espresso attraverso una scala con cinque modalità, mentre la *Aspect-Based* è utile qualora si voglia comprendere l'opinione degli utenti in merito ai diversi aspetti di un prodotto.

Avendo scelto di utilizzare la metodologia *Emotion Detection* la scelta del *lexicon* è ricaduta sul NRC poiché, a differenza dei lexicons AFINN e Bing, non si

limita ad identificare la polarità dei termini – positiva o negativa – ma riconosce anche le possibili emozioni espresse dal termine considerato (si veda Sezione 2.2). La scelta del lexicon si è rivelata importante soprattutto per le prime fasi dell'analisi dove è stato osservato il sentimento espresso nel complesso, senza fare distinzione tra i vari tweet disponibili, ma considerandoli come un unico testo composto da molte frasi. In questo modo è stato possibile identificare l'emozione espressa da ogni frase, che è stata poi aggregata per permettere di avere in modo chiaro e immediato un'idea sull'opinione più diffusa in merito ai vaccini.

Infine, è opportuno precisare che per l'utilizzo del lexicon, che ha permesso di estrarre il sentimento di ogni termine, sono state utilizzate le versioni dei quattro dataset sulle quali non era stato eseguito lo stemming in precedenza, in quanto i lexicons contengono tutte le forme dei termini.

3.1.4 Metodi di clustering applicati

Per l'applicazione dei metodi di clustering sono stati utilizzate le copie dei dataset su cui è stato eseguito lo stemming dei termini. Tra i metodi esposti nella Sezione 2.1 è stato applicato l'algoritmo *K-means*, in quanto più adatto, data l'alta dimensionalità dei dati, sia perché un algoritmo gerarchico non avrebbe fornito risultati facilmente comprensibili se rappresentati graficamente.

Essendo in una situazione *non supervisionata*, non è possibile conoscere il reale numero di clusters, poiché non sono note a priori le etichette dei vari tweet, per questo sono stati testati diversi valori di K. Per individuare il numero ottimale di clusters, si è fatto ricorso all'*elbow-method*, implementando una funzione creata appositamente e rappresentando i valori della somma delle distanze al quadrato (si veda Sezione 2.1.2). Una volta identificato il numero di gruppi più opportuno, è stato eseguito l'algoritmo e i risultati sono stati rappresentati graficamente. Inoltre, per poter fornire un'interpretazione migliore, è stato impiegato nuovamente il lexicon NRC sui singoli cluster per avere una visione sulla presenza di emozioni all'interno dei gruppi.

3.2 Risultati

3.2.1 Analisi preliminari

Dopo aver scaricato i dati di interesse e aver eseguito le fasi di pulizia, come descritto nelle sezioni 3.1.1 e 3.1.2 di questo capitolo, sono state svolte alcune analisi preliminari.

Per quanto riguarda il dataset relativo a *#covidvaccine*, dopo aver costruito la *Term-Document Matrix*, si è osservato che quest'ultima era costituita da quasi cento mila termini e presentava una *sparsità* del 100% (una matrice si dice *sparsa* se la maggior parte dei suoi elementi è pari a zero). In questo caso, la percentuale 100% indica che sono stati mantenuti tutti i termini, compresi quelli che compaiono in un solo documento e presentano quindi valore zero per tutti gli altri, da qui deriva l'elevato numero di termini inclusi.

Per ridurre la *sparsità* e quindi anche il numero di termini, è stata impostata una soglia di sparsità pari a 0.996, che rappresenta la frequenza relativa con cui i termini compaiono nei documenti. In questo modo, le parole che presentano una sparsità maggiore di 0.996, cioè quelle che sono contenute in meno dello 0.4% dei documenti, vengono escluse dalla matrice. Il risultato ottenuto è una TDM che comprende circa 500 termini. Le stesse operazioni sono state svolte anche per gli altri tre dataset, utilizzando la stessa soglia per la *sparsità*, che ha permesso di ottenere delle TDM di circa 500 termini in ogni caso.

Una volta ottenuta la matrice definitiva, sono state rappresentate le frequenze dei termini contenuti nella TDM, per poter meglio osservare quali fossero i più frequenti. Per la rappresentazione grafica sono stati scelti sia un barplot sia un *wordcloud* (Figura 3.1).

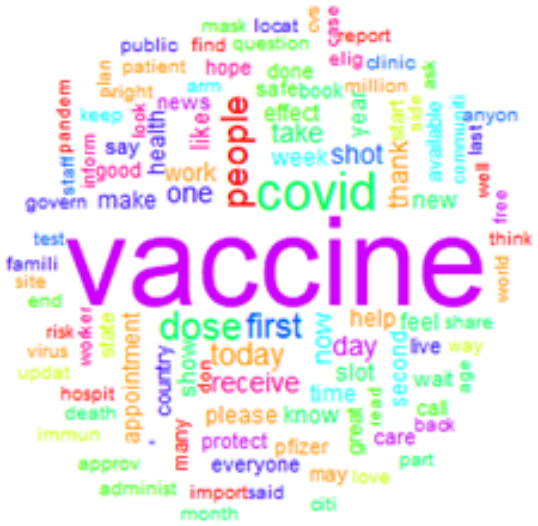

Figura 3.1: Wordcloud dei termini più frequenti nel dataset #covidvaccine

Analizzando i risultati per il dataset *#covidvaccine*, come ci si potrebbe aspettare, si osserva che i due termini più frequenti sono *vaccine* e *covid*, ma si possono trovare anche altre parole strettamente legate alla somministrazione

dei vaccini, come *shot, first, second, dose, receive, appointment, arm*. Inoltre, già da questa prima rappresentazione si notano termini che permettono di cogliere alcune emozioni provate dagli utenti, che possono essere in parte positive come per le parole *hope, great, safe* e *free*, ma ci sono anche termini che sembrano espirimere preoccupazione, come *question, test, death, side* ed *effect*.

Confrontando i termini più frequenti nei quattro dataset si osservano molte similitudini e poche differenze. Queste ultime, sono particolarmente evidenti per il dataset #*astrazeneca*, che tra tutti i vaccini è stato quello interessato da diverse controversie, come è stato spiegato precedentemente. Infatti, tra i termini più frequenti si nota che molte delle parole positive che sono presenti negli altri casi, qui mancano. Ne sono un esempio le parole *hope, great, free* e perfino il termine *protect*, presenti per gli altri due vaccini. D'altra parte, si evidenziano alcune espressioni negative, che risultano assenti tra i termini frequenti degli altri dataset, quali *blood* e *die*, legati probabilmente agli eventi trombotici che si sono verificati in alcuni vaccinati, e il termine *suspend* relativo alla sospensione di questo vaccino in alcuni Paesi.

3.2.2 Emotion Detection

Successivamente alle analisi iniziali, è stata svolta una classificazione delle emozioni su tutti e quattro i dataset. Utilizzando il lexicon NRC sono stati attribuiti i singoli tweet ai sentimenti in essi contenuti, poi per ogni sentimento sono state sommate le frequenze dei documenti ad esso attribuiti e i risultati ottenuti sono stati rappresentati graficamente.

Per quanto riguarda il dataset relativo a #*covidvaccine* (Figura 3.2) si osserva che i due sentimenti più diffusi sono la *fiducia* e l'*aspettativa*. Questo indica che, nel complesso, gli utenti di Twitter hanno un'opinione dei vaccini contro il COVID-19 che li porta ad avere fiducia nella scienza e si aspettano che, con il loro impiego, si possa ottenere almeno un miglioramento della situazione pandemica. D'altra parte, la terza emozione più frequente è la *paura* a causa delle nuove tecnologie utilizzate e del rapido sviluppo dei vaccini.

Per comprendere meglio questi risultati, sono state calcolate anche le proporzioni con cui le varie emozioni si presentano nel dataset, che hanno permesso di osservare che la *fiducia* è presente in quasi un quarto dei tweet considerati, mentre l'*aspettativa* supera la *paura* del 3% circa.

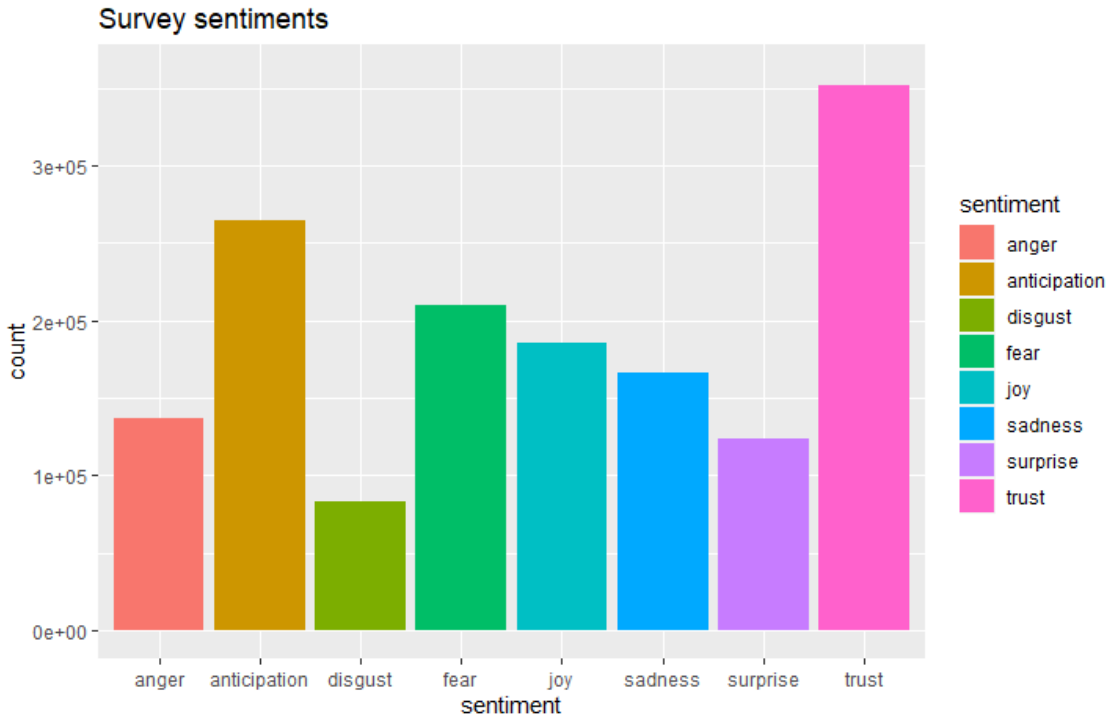

Figura 3.2: Frequenze assolute dei sentimenti per il dataset #covidvaccine

Ripetendo le medesime analisi per gli altri dataset, sono stati ottenuti risultati analoghi (Tabella 3.1), anche se emergono alcune evidenze in contrasto con quanto osservato per i termini più frequenti. Apparentemente, sembrava che per il vaccino Astrazeneca ci fosse un sentimento tendente al negativo e alla sfiducia. Si nota, invece, che il sentimento di *fiducia* risulta pari a quello espresso in favore del vaccino Pfizer/BioNTech, anche se si osserva la maggiore presenza, seppur in modo lieve, di *paura* e *rabbia* nei confronti del vaccino Astrazeneca. Per quanto riguarda il vaccino Moderna, è presente un livello più alto di *tristezza* e un minore livello di *fiducia*. Infine, osservando il sentimento di *sorpresa* si nota che risulta considerevolmente presente per i vaccini Pfizer/BioNTech e Moderna rispetto al valore relativo ad Astrazeneca; ciò può essere dovuto allo scetticismo iniziale dovuto alla nuova tecnologia ad mRNA impiegata dai primi due vaccini, rispetto al metodo tradizionale del terzo, che si è rivelata in seguito molto efficace, avendo ottenuto buoni risultati a livello clinico.

Considerando che i risultati esposti finora mostrano le emozioni in modo statico, si è deciso di svolgere anche un'analisi che tenga conto dell'andamento temporale del sentimento, per poter osservare le variazioni che si sono verificate in seguito ai vari eventi che hanno interessato i vaccini. Per tale scopo, sono

Tabella 3.1: Proporzioni delle emozioni

Emotion	Pfizer	Moderna	Astrazeneca
trust	0.21	0.19	0.21
anticipation	0.16	0.15	0.16
fear	0.15	0.15	0.16
joy	0.10	0.10	0.10
sadness	0.12	0.13	0.12
anger	0.11	0.11	0.12
surprise	0.10	0.11	0.08
disgust	0.05	0.05	0.06

state considerate solo le modalità *positività* e *negatività* fornite dal lexicon NRC. Per il singolo tweet è stata determinata la polarità complessiva del sentimento calcolando la differenza tra il livello di positività e di negatività e, in seguito, i risultati ottenuti sono stati raggruppati per data, utilizzando un'apposita funzione.

Osservando il grafico per l'hashtag *#covidvaccine* non si nota nessuna particolare oscillazione, infatti, i picchi e i cali presenti nella polarità complessiva del sentimento si alternano in modo quasi regolare.

Confrontando i grafici relativi ai singoli vaccini si notano andamenti differenti (Figura 3.3). Il grafico relativo al vaccino PFizer/BioNTech evidenzia un andamento regolare, iniziato con valori più elevati e successivamente rimasto costante, ad eccezione di alcuni picchi in corrispondenza delle date seguenti:

- **01/12/2020:** la Gran Bretagna approva il vaccino Pfizer/BioNTech;
- **08/12/2020:** una signora di 90 anni è la prima donna nel mondo a ricevere la somministrazione del vaccino Pfizer/BioNTech;
- **31/03/2021:** sono resi noti i risultati clinici degli studi di Fase3 eseguiti su adolescenti tra i 12 e 15 anni, che hanno mostrato un'efficacia del 100% del vaccino Pfizer/BioNTech.

Per il vaccino Moderna risulta subito evidente che la polarità complessiva si mantiene su valori inferiori rispetto agli altri due vaccini e presenta un andamento più oscillatorio con pochi picchi significativamente evidenti. Questi ultimi sono in particolare due e sono relativi ai seguenti eventi:

- **19/12/2020:** approvazione della Food and Drug Administration (FDA) del vaccino Moderna per tutti gli individui maggiorenni;

- **25/05/2021:** sono stati rilasciati i risultati degli studi di Fase2 sugli adolescenti, con risultati positivi.

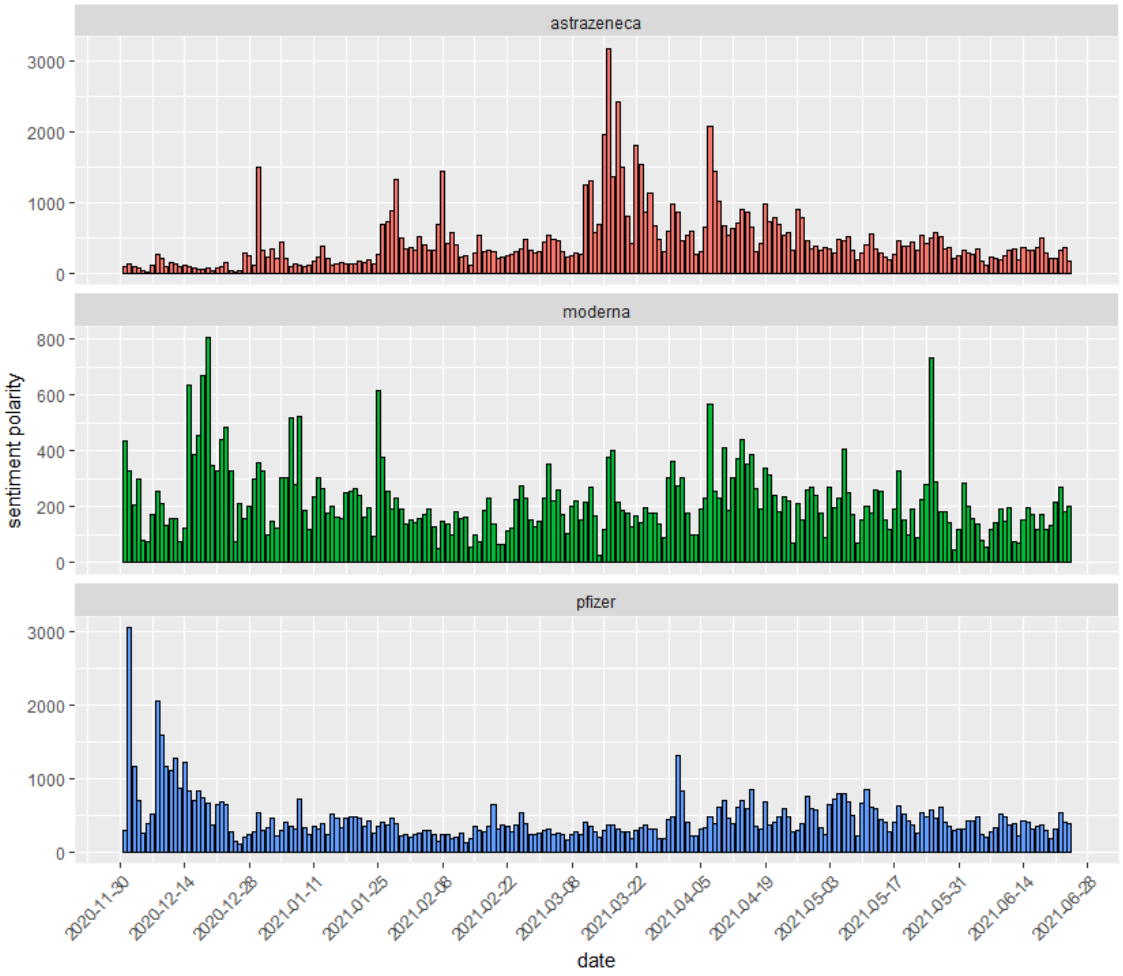

Figura 3.3: Andamento temporale della polarità complessiva del sentimento

Infine, il sentimento relativo al vaccino Astrazeneca presenta inizialmente valori bassi, che sono aumentati nella parte centrale del periodo di osservazione, per poi tornare a decrescere. I picchi che si evidenziano e che presentano valori notevolmente alti, sono relativi alle seguenti date e rispettivi avvenimenti:

- **30/12/2020:** la Gran Bretagna approva il vaccino Astrazeneca;
- **29/01/2021:** l'Agenzia Europea del Farmaco (EMA) approva il vaccino Astrazeneca;
- **21/02/2021:** vengono modificate le disposizioni di somministrazione del vaccino Astrazeneca, limitandone l'inoculazione a coloro che hanno meno di 65 anni;

- **16/03/2021:** viene sospeso l'uso del vaccino Astrazeneca;
- **07/04/2021:** l'EMA annuncia la scoperta di una possibile correlazione tra la somministrazione del vaccino Astrazeneca e gli eventi trombotici in alcuni individui vaccinati.

Nelle rappresentazioni appena analizzate, i picchi indicano che il sentimento di positività sovrasta quello di negatività e, in altri termini, significa che gli utenti si mostrano positivi e quindi in accordo con quanto accaduto in corrispondenza di quelle date.

Come ultima fase dell'analisi, si è deciso di applicare l'algoritmo di clustering *K-means* per capire se fosse possible identificare dei gruppi di utenti che avessero una visione differente in merito ai vaccini.

A partire dal dataset relativo all'hashtag #*covidvaccine*, la rappresentazione dell'*elbow method* non presenta un "gomito" evidente e cresce in modo quasi lineare fino a K= 2, dove si osserva una una variazione più significativa. Per questo si è scelto di eseguire l'algoritmo con un numero di clusters pari a due. Rappresentando graficamente il risultato ottenuto, i due gruppi non si distinguono in modo netto, ma risultano in parte sovrapposti, quindi ne è stato analizzato il contenuto per poter fornire un'interpretazione migliore. Ne sono risultati due clusters molto simili in termini di sentimento, anche se tra i termini più frequenti si osservano lievi differenze, poiché un gruppo di utenti utilizza parole più positive rispetto all'altro.

Per quanto riguarda, invece, i dataset specifici per i singoli vaccini, i risultati sono più interessanti. Nel caso del dataset relativo al vaccino Pfizer/BioNTech è stato individuato un numero di cluster ottimale pari a tre. Eseguendo il clustering, come è avvenuto per il dataset precedente, si ottengono gruppi in parte sovrapposti e non ben distinti (Figura 3.4). Tuttavia, esplorando i tweet contenuti al loro interno, si identificano tre tipi diversi di utenti che hanno posizioni diverse in merito alla questione del vaccino.

Un primo gruppo è composto da utenti che non sono positivi né molto felici rispetto alla situazione, infatti, escludendo il sentimento di *fiducia*, che risulta sempre il più presente, prevale la *paura* (16%). Altre emozioni che si manifestano in proporzione considerevole sono la *tristezza* (13%) e la *rabbia* (12%), evidenziando che questo gruppo di utenti ha fiducia nel fatto che veranno prese le decisioni opportune in merito al vaccino, ma non è favorevole alle procedure di ricerca e di sviluppo troppo sperimentali, né alla somministrazione del vaccino. Il

secondo cluster identificato, presenta invece una posizione più neutra rispetto alla questione, con emozioni positive e negative quasi equilibrate. Infatti, lasciando nuovamente da parte la *fiducia*, si osserva che *aspettativa* e *paura* si manifestano circa nella stessa misura (16% e 15%), così come accade anche per *gioia* e *tristezza* (11% e 12%). Questo gruppo di individui si mostra quasi in una posizione di indecisione rispetto al vaccino Pfizer/BioNTech. Infine, il terzo ed ultimo cluster mostra utenti più positivi e ottimisti, fiduciosi nel raggiungimento di buoni risultati (*fiducia* 19%) e con più *aspettativa* (18%). Si mostrano, inoltre, più felici per la possibile fine della pandemia (*gioia* 12%) e meno spaventati e tristi (*paura* 13% e *tristezza* 11%).

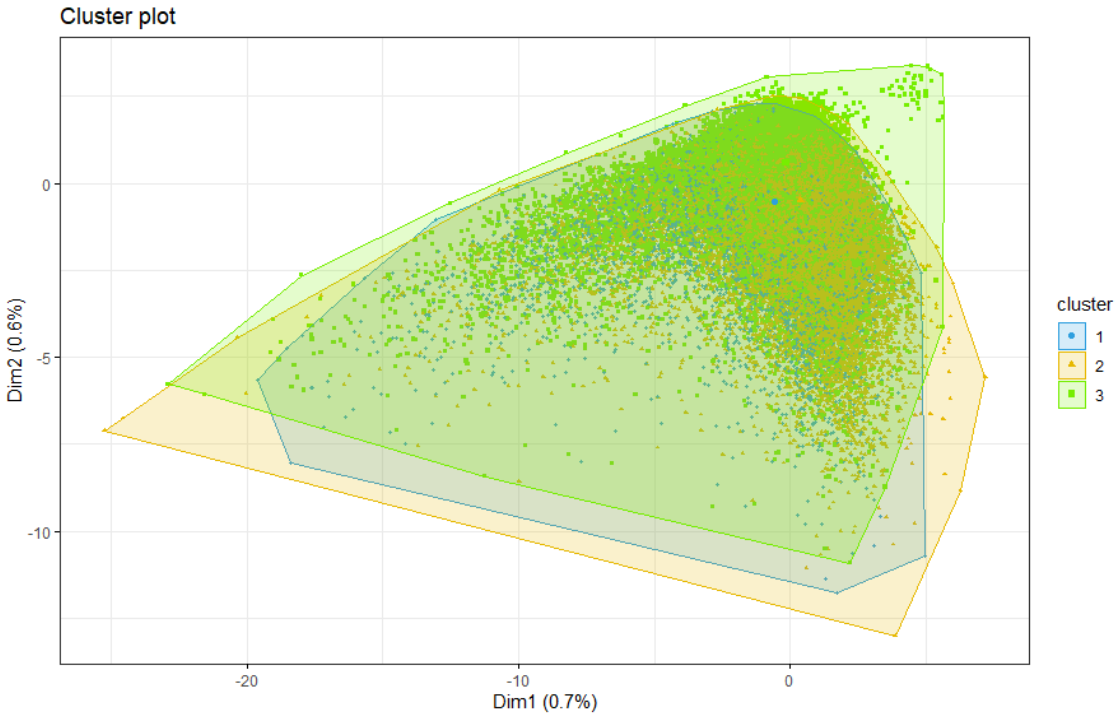

Figura 3.4: Clustering per il dataset #pfizer

Per l'esecuzione dell'algoritmo *K-means* sul dataset riferito all'hashtag #*moderna*, è stato scelto K=2 poiché il grafico dell'*elbow method* ha mostrato un andamento crescente lineare, mentre in corrispondenza del valore 2 si è osservata una variazione più evidente. Anche in questo caso, i due gruppi sembrano essere sovrapposti e poco distinti anche se, dall'analisi del contenuto dei tweet, emergono due posizioni diverse tra gli utenti (Figura 3.5). Il primo cluster delinea una posizione contraria e timorosa nei confronti del presente vaccino. Infatti, sono presenti in proporzioni considerevoli *paura* (15%), *tristezza* (14%) e

rabbia (12%), che si presentano con almeno due punti percentuali in meno nel secondo cluster, dove prevalgono invece *fiducia* (20%), *aspettativa* (17%) e *gioia* (12%). Un'ulteriore aspetto che può essere preso in considerazione riguarda il numero di tweet appartenenti ad ogni gruppo, in quanto risulta che il cluster più numeroso è quello con una posizione maggiormente negativa nei confronti del vaccino Moderna.

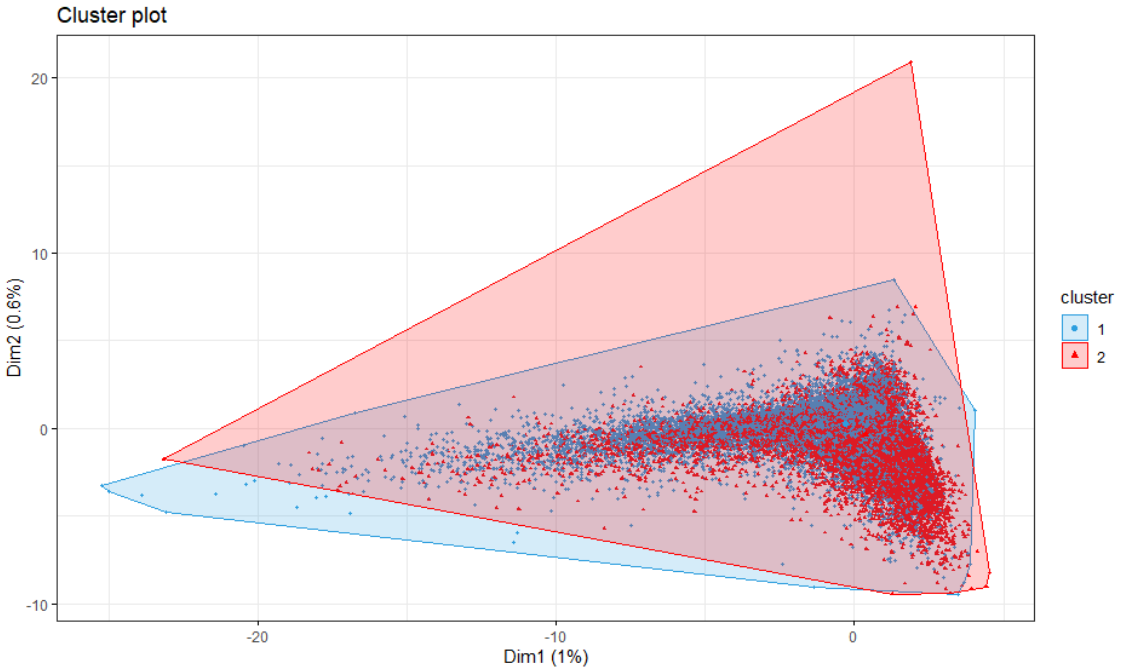

Figura 3.5: Clustering per il dataset #moderna

Infine, per quanto riguarda il vaccino Astrazeneca, è stato individuato un numero di clusters ottimale pari a due, seppur con alcune difficoltà dal momento che il grafico del *metodo del gomito* ha mostrato un andamento crescente simile a quello osservato nel caso del vaccino Moderna. Anche in questa situazione, è stato necessario approfondire la composizione dei gruppi in quanto, come è avvenuto per i due datasets precedenti, la rappresentazione grafica dei gruppi risultava sovrapposta (Figura 3.6). Tuttavia, osservando il contenuto dei clusters e analizzandone il sentimento, è risultato che in entrambi si denota un'opinione complessivamente negativa, caratterizzata da una considerevole presenza di *paura* (15% in un gruppo e 16% nell'altro) e *tristezza* (11% nel primo e 12% nel secondo). Le differenze principali si osservano in relazione al sentimento di *rabbia* che risulta presente in misura minore nel secondo gruppo, compensato da livelli maggiori di *aspettativa* e *fiducia*. Nel complesso si può quindi dire che, anche se i

due gruppi hanno un'opinione prevalentemente pessimistica nei confronti del vaccino Astrazeneca, in uno dei due sono presenti anche emozioni positive in una proporzione maggiore.

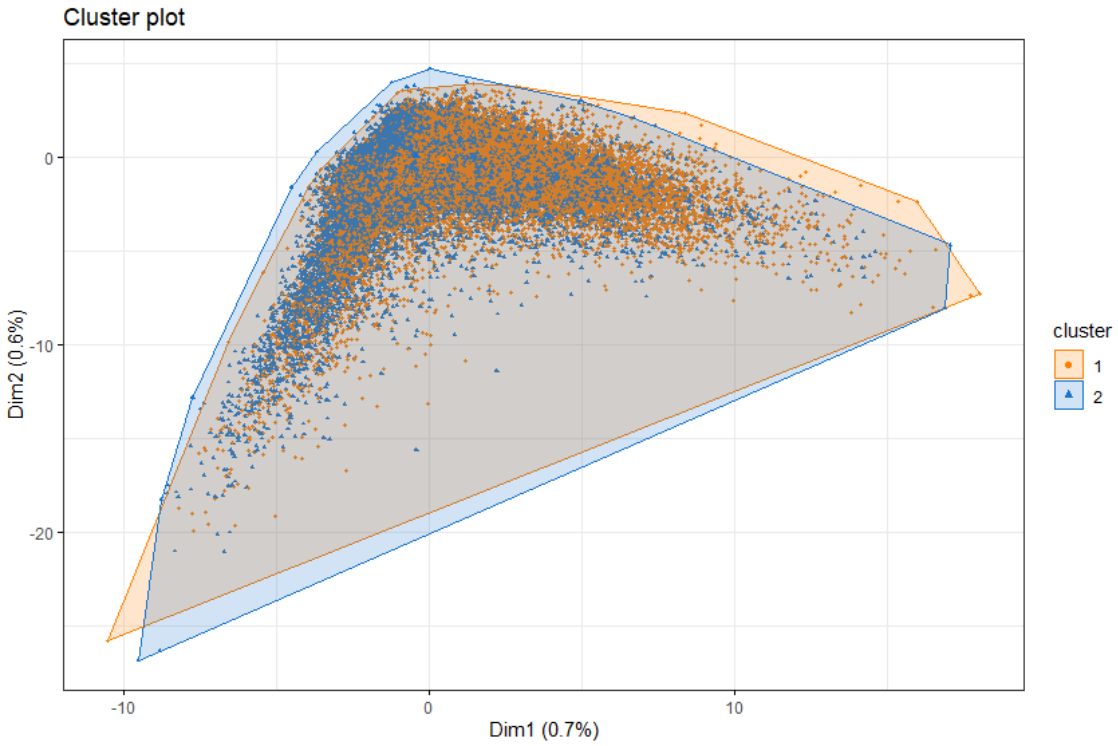

Figura 3.6: Clustering per il dataset #astrazeneca

Conclusione

L'obiettivo della presente tesi di laurea era quello di mostrare l'utilità della *Sentiment Analysis* e riportare un esempio di una sua applicazione relativamente ad un argomento attuale come quello del COVID-19.

Dopo una prima descrizione teorica di tale tecnica, è stata proposta l'applicazione della metodologia di *Emotion Detection* ad alcuni dataset contenenti una serie di tweet, esclusivamente in lingua inglese, relativi ai vaccini contro il COVID-19. Dopo aver scaricato i dati attraverso l'utilizzo del linguaggio di programmazione `Python`, questi sono stati caricati in `RStudio` dove è stato possibile applicare gli strumenti propri dell'*Emotion Detection*, quali *lexicons* e metodi di *clustering*.

Attraverso l'utilizzo del lexicon NRC, è stato possibile analizzare le emozioni espresse dagli utenti, sia da un punto di vista globale sia nello specifico per i singoli vaccini. Le analisi hanno evidenziato la complessiva positività degli utenti in riferimento all'argomento, data la predominanza del sentimento di *fiducia* seguito da quello di *aspettativa*, seppur con una nota di *paura*. Le elaborazioni hanno anche mostrato come, nel corso dei mesi considerati, le notizie abbiano influenzato ed indirizzato le opinioni delle persone. Si è osservato, infatti come in corrispondenza del rilascio di risultati di studi clinici o di dichiarazioni riguardanti i vaccini, gli utenti abbiano mostrato maggiore positività e si siano trovati in accordo con le decisioni prese.

Con l'applicazione dell'algoritmo di clustering *K-means* si è cercato di individuare gruppi di tweet che avessero un contenuto omogeneo in termini di sentimento. Nonostante da un punto di vista grafico non si siano delineati gruppi nettamente distinti, analizzando i tweet contenuti in ogni cluster è stato possibile cogliere alcune differenze tra i gruppi. Risultati soddisfacenti si sono ottenuti principalmente per i dataset specifici per i singoli vaccini, identificando gruppi di utenti con una visione simile, in certi casi più ottimistica mentre in altri più pessimistica.

Il lavoro svolto ha mostrato le pontenzialità della *Sentiment Analysis* per il raggiungimento degli obiettivi prefissati. Risultati ancor più soddisfacenti ed esaustivi si sarebbero potuti ottenere attraverso tecniche più avanzate e l'utilizzo di elaboratori più efficienti, data l'elevata quantità di dati trattata che ha reso difficoltose alcune analisi.

Bibliografia

Analystics India Mag (2021). Comprehensive guide to k-medoids clustering algorithm.

Casadei, C. (2019a). Apprendimento non supervisionato – clustering k-means.

Casadei, C. (2019b). Classificatori non lineari – alberi decisionali e foreste casuali.

Data Skills (2015). Tecniche di clustering.

Eremyan, R. (2018). Four pitfalls of sentiment analysis accuracy.

FrontlineSolvers (2012). Hierarchical clustering.

Gallo, C. & Bioagromed, C. d. R. I. (2007). *Reti Neurali Artificiali: Teoria ed Applicazioni Finanziarie. Dipartimento di Scienze Economiche, Matematiche e Statistiche, Universita'di Foggia, Quaderni DSEMS (gen. 2007)* .

Guthrie, L., Pustejovsky, J., Wilks, Y. & Slator, B. M. (1996). *The role of lexicons in natural language processing. Communications of the ACM* **39**, 63–72.

Kim, H. (2018). *Limits of the Bing, AFINN and NRC Lexicons with the Tidytext Package in R* .

Kumar, S. (2020). Silhouette method — better than elbow method to find optimal clusters.

Liddy, E. D. (2001). *Natural language processing* .

Liu, B. (2011). Opinion mining and sentiment analysis. In *Web Data Mining*. Springer, pp. 459–526.

Liu, B., Li, X., Lee, W. S. & Yu, P. S. (2004). Text classification by labeling words. In *AAAI*, vol. 4.

Ministero della Salute (2020). Nuovo coronavirus.

Mishra, N. & Jha, C. (2012). *Classification of opinion mining techniques. International Journal of Computer Applications* **56**.

Mitchell, T. M. et al. (1997). *Machine learning* .

Muhammad, S. H., Brazdil, P. & Jorge, A. (2020). *Incremental Approach for Automatic Generation of Domain-Specific Sentiment Lexicon. Advances in Information Retrieval* **12036**, 619.

Romanyshyn, M. (2013). *Rule-based sentiment analysis of ukrainian reviews. International Journal of Artificial Intelligence & Applications* **4**, 103.

Romero Llombart, Ò. (2017). *Using machine learning techniques for sentiment analysis* .

Silge, J. & Robinson, D. (2017). Welcome to text mining with r.

Thakkar, H. & Patel, D. (2015). *Approaches for sentiment analysis on twitter: A state-of-art study. arXiv preprint arXiv:1512.01043* .

Vartanova, V. (2019). Sentiment and emotion analysis for beginners: types and challenges.

WebWise (2018). Explained: what is twitter?

WHO (2020). Coronavirus disease (covid-19) pandemic.

Yang, C.-S. & Shih, H.-P. (2012). *A rule-based approach for effective sentiment analysis* .