Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea in Informatica per il Management

SENTIMENT ANALYSIS: PREVISIONE CON TECNICHE DI INTELLIGENZA ARTIFICIALE

Relatrice:
Chiar.ma Prof.
ELENA L. PICCOLOMINI

Presentata da:
MATTEO TRAMONTANO

III Sessione
Anno Accademico 2018/2019

Alla mia famiglia ...

Introduzione

In statistica e informatica, la locuzione inglese **big data**, o in italiano megadati, indica genericamente una raccolta di dati informativi così estesa in termini di volume, velocità e varietà da richiedere tecnologie e metodi analitici specifici per l'estrazione di valore o conoscenza.
Il termine è utilizzato in riferimento alla capacità (propria della scienza dei dati) di analizzare ovvero estrapolare e mettere in relazione un'enorme mole di dati eterogenei, strutturati e non strutturati (grazie a sofisticati metodi statistici e informatici di elaborazione), allo scopo di scoprire i legami tra fenomeni diversi (ad esempio correlazioni) e prevedere quelli futuri [1].

Il termine **Web 2.0** indica genericamente la seconda fase di sviluppo e diffusione di Internet, caratterizzata da un forte incremento dell'interazione tra sito e utente: maggiore partecipazione dei fruitori, che spesso diventano anche autori (blog, chat, forum, wiki); più efficiente condivisione delle informazioni, che possono essere più facilmente recuperate e scambiate con strumenti peer to peer o con sistemi di diffusione di contenuti multimediali come Youtube; affermazione dei social network. Nuovi linguaggi di programmazione consentono un rapido e costante aggiornamento dei siti web anche per chi non possieda una preparazione tecnica specifica. Il fenomeno è ancora in fortissima evoluzione [2].

L'evoluzione strettamente interconnessa tra l'aumento esponenziale di dati (appunto i Big Data) e lo sviluppo del Web 2.0 ha reso possibile lo sviluppo

di tecnologie atte all'interpretazione di questi dati per i più svariati utilizzi. Diventa quindi necessario uno strumento capace di poter interpretare gli infiniti testi presenti sulla rete, che siano, in base al fine ultimo, scritti su blog, giornali on-line, social network o in qualsiasi altra piattaforma che consenta agli utenti di esprimere la propria opinione.
Questo strumento prende il nome di **Sentiment Analysys**, si tratta di un campo dell'elaborazione del linguaggio naturale che si occupa di costruire sistemi per identificazione ed estrazione di opinioni dal testo. Si basa sui principali metodi di linguistica computazionale e di analisi testuale [3].
Nel corso degli ultimi anni questo strumento è stato molto utilizzato nei più svariati settori.
Lo scopo di questa tesi è di spiegare la realizzazione e l'applicazione della Sentiment Analysis, focalizzandosi in modo particolare sul punto di vista algoritmico di questa pratica. Questa, infatti, può essere svolta con numerosi metodi differenti e può avere infiniti utilizzi che variano da quelli valutativi a quelli predittivi di qualsiasi campo che sia commerciale, sanitario, politico, finanziario ecc.
L'interesse tra tutti questi settori è stato posto, con fini di previsione, sull'ambito economico-finanziario. In particolare è stato scelto il campo delle criptovalute, ancora più nello specifico il Bitcoin, la più famosa tra la criptovalute, la quale negli ultimi 10 anni sta rivoluzionando il mercato mondiale.
La tesi è quindi strutturata come segue:

- **Capitolo 1:** Questo capitolo ha lo scopo di mostrare e fare comprendere al lettore l'enorme risorsa che rappresentano i dati presenti in rete (grazie alla diffusione esponenziale dei Social Media), nonchè le differenti classificazioni di essi in base ai vari parametri. Non viene trattato in questo capitolo, ne nel resto della tesi, un altro argomento molto importante ovvero la protezione di questi.

- **Capitolo 2:** In questo capitolo si entra nel dettaglio del funzionamento della Sentiment Analysis. Vengono spiegati i principale modelli e metodi alla base di questa pratica, per poi entrare nel dettaglio solo di

alcuni di essi. Lo scopo è quindi quello di mostrare la moltitudine di potenzialità che questo strumento offre

- **Capitolo 3:** Nel terzo ed ultimo capitolo viene mostrato un caso studio, in particolare viene effettuata l'analisi del sentiment su un set di Tweet (in un periodo di tempo ben definito) riguardanti il Bitcoin, con lo scopo di predirne l'andamento. I dati previsti vengono poi confrontati con gli effettivi dati storici e vengono tratte le conclusioni.

Indice

Elenco delle figure

Elenco delle tabelle

Capitolo 1

Il ruolo della rete

Il pensiero altrui è da sempre il punto focale di chi ha il compito di prendere decisioni, o in alternativa, di chi ha già preso una decisione e vuole sapere gli effetti di tale decisione [5]. Ciò è applicabile qualunque sia il soggetto in questione delle frase sopra, che sia un venditore, un ricercatore o un politico. Scopo di questo capitolo è la spiegazione del fenomeno dell'esplosione Social Media al fine di mostrare l'enorme quantità di dati che questi sono in grado di fornire, la loro classificazione ed i loro possibili utilizzi.

1.1 La rete sociale

Con il concetto di rete sociale si intende qualunque struttura, formata da un insieme di persone o organizzazioni di persone e le loro interazioni. Per gli esseri umani i legami vanno dalla conoscenza casuale, ai rapporti di lavoro, ai vincoli familiari o anche rapporti commerciali.
Dal momento che la rete sociale si trova su una piattaforma virtuale si parla di social media: piattaforme virtuali che consento agli utenti presenti su esse di generare e condividere contenuti, i quali rimangono sulla piattaforma, spesso in maniera permanente.
Particolarmente rilevanti sono quei social media che esprimono al loro interno una comunicazione bilaterale, ovvero dove vi è sia la produzione di contenuti,

sia la produzione di relazioni.
Al fine di studiare la validità dei dati, e soprattutto il modo con cui andranno analizzati possono essere utilizzati in primis due diversi indicatori:

- indicatori statici
- indicatori dinamici

1.1.1 Indicatori statici

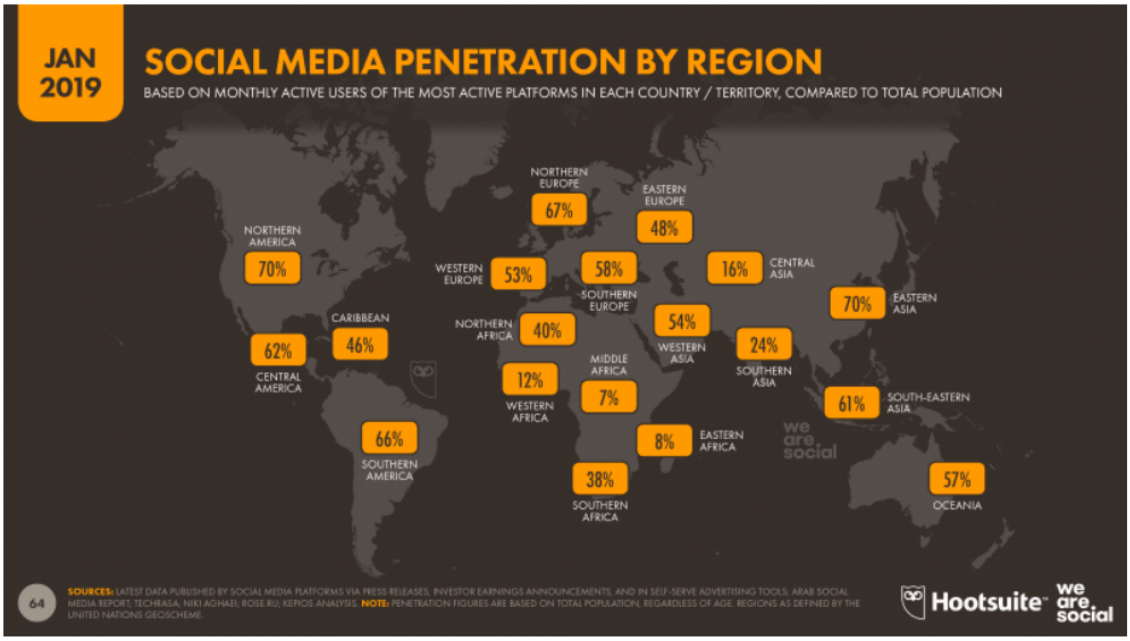

Figura 1.1: Social media penetration by region [4]

Gli indicatori statici, sono quegli indicatori che fotografano una situazione in un determinato momento, come ad esempio il numero di utenti di internet, il numero degli utenti presenti sui social media...
Particolare attenzione nel campo dell'analisi va attribuiti agli utenti, in quanto rappresentano non semplicemente dei dati ma delle persone e quindi incoerenza tra loro.

In primo luogo vanno distinti gli utenti attivi da quelli presenti, il mezzo che essi utilizzano per connettersi o la fascia di età.
Nei capitoli successivi verrà mostrato come queste distinzioni risultano fondamentali ai fini della Sentiment Analysis.
La Figura 1.1 mostra un esempio di indicatori statici ovvero il tasso di penetrazione dei social media suddiviso per continenti nel Gennaio 2019.
(fonte: StudioSamo)

1.1.2 Indicatori dinamici

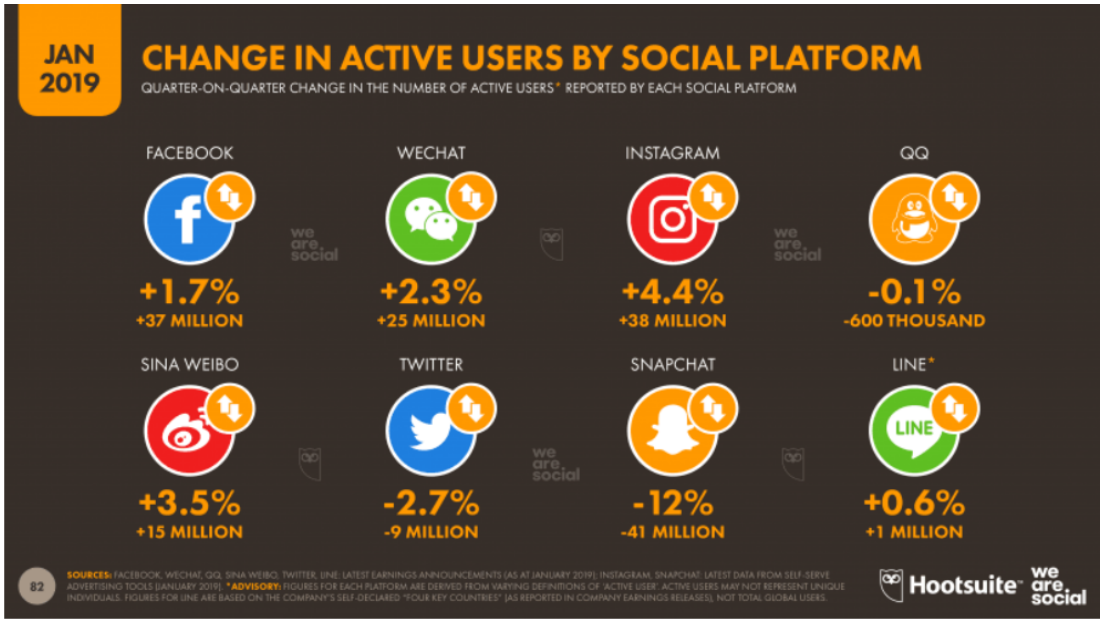

Figura 1.2: Change in active users by social platform [4]

Gli indicatori dinamici sono quegli indicatori che si riferiscono a tassi di crescita nel tempo. Questi possono essere calcolati su base mensile, trimestrale, o annuale. Tuttavia per confrontarli è necessario utilizzare lo stesso arco temporale in modo da non invalidare l'analisi. Tuttavia se i dati presentano granularità diversa (per esempio venendo calcolati su base mensile piuttosto che annuale) è comunque possibile effettuare una rimodulazione

per renderli omogenei. L'immagine sopra mostra l'avanzamento degli utenti attivi nel tempo dei vari social media, si tratta quindi di indicatori dinamici. (fonte: StudioSamo)

1.1.3 Top-Down VS Bottom-Up

Per via della loro enorme diffusione i social media sono stati quindi studiati a lungo in ambito informatico, commerciale, psicologico e altri.
In questi studi si sono distinti principalmente due approcci:

- approccio top-down
- approccio bottom-up

L'approccio top down

L'approccio top-down cerca di capire in che maniera ed in che misura la comunicazione sui social media sia in grado di influire sulle scelte, sui pensieri e sui comportamenti dei suoi utenti.
Questo approccio è visibile in innumerevoli campi.
Nel campo del giornalismo, per esempio, nonostante il rischio della "fake news" i social media al giorno d'oggi rappresentano la prima fonte di informazione per la popolazione e, al contrario di quanto di possa pensare, sono i media tradizionali che spesso cercano notizie all'interno dei social media.
Un altro campo di efficacia di questo approccio è l'ambito commerciale, i social media infatti, visto la loro facilità d'accesso e il loro costo ridotto, diventando un importante strumento di "brand management", consentendo inoltre il confronto diretto con compratori e potenziali.
Anche in campo politico questo approccio è da prendere in considerazione in quanto è stato riscontrato che i candidati che mantengono una relazione diretta e continua con i loro follower attraverso i social media hanno un effetto significativo sui voti riscontrati alle elezioni.
In particolare una approccio sempre più comune in campo politico è quello

di trasmettere messaggi personalizzati per ogni gruppo di elettori.

L'approccio bottom-up

Ancora più interessante è l'approccio bottom-up che vede i social media come strumento di aggregazione da cui estrarre informazioni in modo da comprendere i fenomeni più complessi.
Ciò si traduce nella possibilità di effettuare previsioni, identificando dinamiche che si stanno verificando in tempo reale.
Rientrano in questa categoria anche i sempre più numerosi studi che utilizzano i l'analisi testuale per fare delle vere e proprie previsioni.
Esistono innumerevoli progetti che si occupano di questo, ovvero di effettuare previsioni sulla base delle ricerche effettuate in rete come il programma OSI (Open Source Indicators) che raccogliendo i dati che circolano in rete monitora la diffusione di idee su essa cercando di fare previsioni sui cambiamenti d'umore della popolazione nei diversi luoghi del mondo. Questo perché è stata provata una stretta correlazione tra queste variazioni d'umore e vari fenomeni sociali come crisi, rivolte e persino catastrofi naturali.
Altro esempio è il programma Recorded Future, un programma collaborativo tra Google e CIA.
Questi programmi analizzando ricerche in rete, siti web, blog e social media sono in grado di eseguire previsioni *nowcasting* e *forecasting*.
Essi si sono dimostrati capaci di eseguire le più svariate previsioni, dall'ambito economico in cui dimostrano come il volume dei commenti pubblicati su forum specializzati sia un indicatore della volatilità dei mercati azionari, ma anche l'andamento della disoccupazione, così come quello del mercato immobiliare. Dal punto di vista del marketing dimostrano anche come sia possibile prevedere l'andamento delle vendite che possono essere di un prodotto di moda cosí come l'incasso di un film al botteghino.
Caso molto interessante è l'area delle scienze mediche, anche in questo campo l'analisi ha spesso portato a previsioni veritiere. Analizzando le ricerche

eseguite, sulla base di particolari parole chiave queste ricerche si sono rivelate utili per prevedere lo scoppio di epidemie, anche se sembrano sovrastimare la diffusione del virus. Un esempio recente di questo tipo di analisi è stato il caso "***BlueDot***".

Bottom up: il caso Blue Dot

Il progetto Blue Dot nasce dal dottor Kamran Khan, medico specializzato in malattie infettive. Start up fondata in Canada nel 2014, questa si occupa appunto della previsione di malattie, virus, epidemie o quant'altro attraverso l'uso opportuno di tecniche di intelligenza artificiale.
Il 31 dicembre scorso BlueDot aveva già informato le autorità canadesi di quanto stava accadendo in Cina con il Coronavirus, anticipando quello che sarebbe successo ed anche le città in cui il virus si sarebbe diffuso, nell'ordine esatto: Bangkok, Seoul, Taipei, Tokyo.
BlueDot ha infatti sviluppato una tecnologia di elaborazione del linguaggio naturale e di machine learning per aggregare notizie, dati di compagnie aeree e segnalazioni di malattie epidemiche negli animali, in modo da tracciare un quadro evolutivo delle epidemie in corso che sia più veloce della diffusione della malattia e che permetta di prevenirla: si tratta di una piattaforma basata sull'AI in grado di elaborare miliardi di dati. Tra questi, anche quelli provenienti dai social network, il monitoraggio della vendita di biglietti aerei e degli altri mezzi di trasporto per prevedere l'apertura di nuovi focolai [6][7].

Capitolo 2

Dentro la Sentiment Analysis

In questo capitolo si entrerà nello specifico del processo della Sentiment Analysis. Verranno descritti i principali metodi per effettuare l'estrapolazione delle opinioni dai testi compresa l'aggregazione di queste e la valutazione dei risultati. Tuttavia prima di entrare nel vivo delle diverse tecniche esistenti per effettuare la Sentiment Analysis è necessario fornire alcune definizioni preliminari ed effettuare alcune operazioni.

2.1 Definizioni preliminari

L'analisi del sentiment o Sentiment Analysis (nota anche come opinion mining) è un campo dell'elaborazione del linguaggio naturale (Natural Language Processing) che si occupa di costruire sistemi per l'identificazione ed estrazione di opinioni dal testo.
Essa si basa sui principali metodi di linguistica computazionale e di analisi testuale.
Come sopra introdotto questa è utilizzata in molteplici settori, dalla politica ai mercati azionari, dal marketing alla comunicazione, dall'ambito sportivo a quello delle scienze mediche e naturali, dall'analisi dei social media alla valutazione delle preferenze del consumatore.
Scopo di chiunque effetti questa analisi è dunque quello di estrarre l'opinione

e di conseguenza il sentimento di una frase, un testo o un intero documento scritto in linguaggio naturale, ovvero in linguaggio umano e non macchina. La Sentiment Analysis pone la sua attenzione sulla polarità dell'emozione (Positiva, Negativa o Neutra) per questo motivo ci si riferisce a Sentimento o Opinione come se fossero equivalenti. Tuttavia al fine di rendere possibile questa conversione sono state formalizzate diverse definizioni:

Definizione: Entità

Viene chiamata entità qualsiasi prodotto, persona, evento, organizzazione o problema riconducibile ad una coppia [8]

$$(C, T) \tag{2.1}$$

In cui;

- C rappresenta un insieme di componenti di quell'entità
- T rappresenta un insieme di attributi del componente C di quell'entità

Esempio:

> Una macchina fotografica rappresenta l'entità, sua volta questa macchina fotografica avrà diverse componenti come batteria, telecamera, memoria e così via. Ed infine ogni componente come ad esempio la batteria avrà poi i suoi attributi come capacità di memoria, velocità di memoria e così via.

Definizione: Opinione

In questa definizione formale, l'opinione viene descritta come una quintupla [8]

$$(e_i, a_{ij}, O_{ijkl}, h_k, t_l) \tag{2.2}$$

- e_i rappresenta l'entità in questione

- a_{ij} rappresenta l'aspetto j-esimo dell'entità e_i
- O_{ijkl} rappresenta l'orientamento dell'opinione
- h_k rappresenta l' "opinion holder"
- t_l rappresenta il tempo, ovvero quando l'opinione è stata espressa

Partendo quindi da queste definizioni si passa da testo destrutturato a dati strutturati, consentendo di ricavare informazioni più accurate per il nostro scopo. Utilizzando il termine entità come oggetto a cui si riferisce li testo.
Un esempio:

> L'obbiettivo della telecamera è grandioso
> Mario Rossi
> 27/01/2020

In questo esempio

- e_i è la telecemera
- a_{ij} è l'obbiettvo
- O_{ijkl} positivo
- h_k Mario Rossi
- t_l 27/01/2020

Come si può notare questa definizione da quindi la possibilità di passare da testo non strutturato a testo strutturato consentendo una futura estrazione dell'opinione.
Avendo quindi un insieme di opinioni raggruppate in quintuple, estratta l'entità e l'attributo di riferimento si riesce ad avere un idea strutturata del sentiment relativo a quell'attributo di quell'entità.
D'ora in poi, sfruttando le definizioni sopra definite, si utilizzer'a il termine entit'a per riferirsi all'oggetto su cui 'e espressa l'opinione.

2.2 Il pre-processing

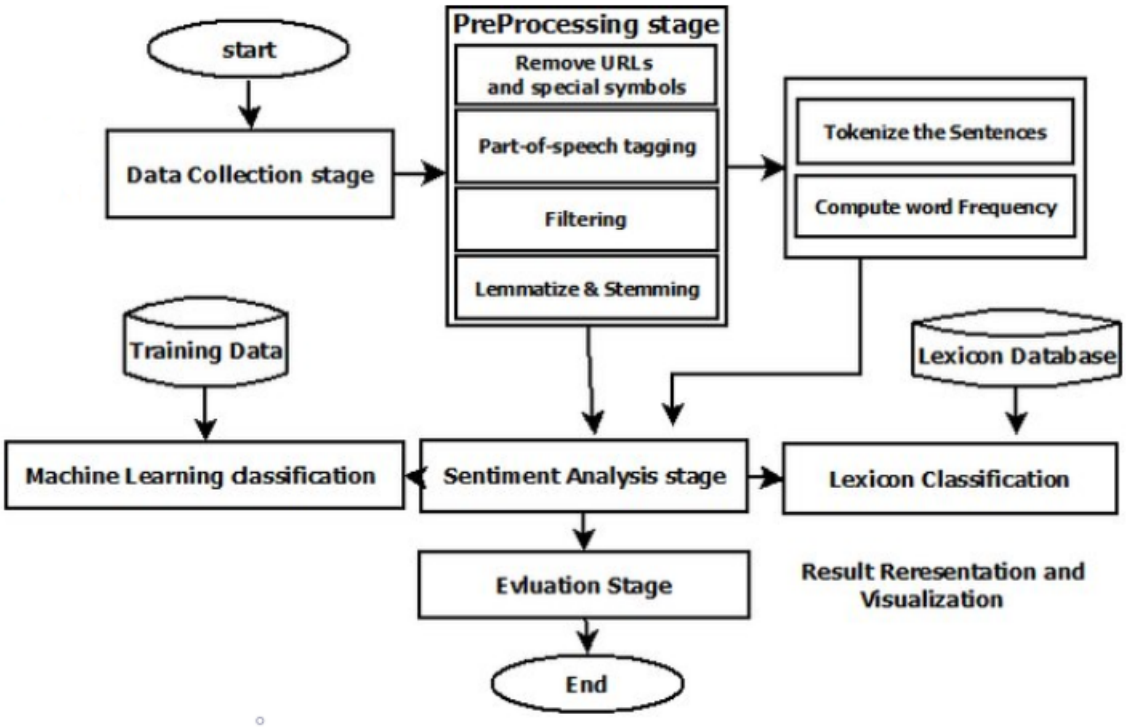

Figura 2.1: Sentiment Analysis process

Una caratteristica del processo di generazione della Sentiment Analysis è il fatto che il sentimento non viene determinato da poche informazioni individuali ma bensí dall'aggregazione di tutti i sentimenti rilevati.
Obbiettivo primario è quello di identificare la presenza di emozioni all'interno del testo.

Text pre-processing

Come si nota dalla Figura 2.1, la prima fase dell'analisi è il *"Text pre-processing"*, ovvero la preparazione del testo.

La preparazione dei dati viene eseguita per eliminare dati incompleti, dati rumorosi e incoerenti.
La fase di pre-processing è fondamentale per ottenere dati adeguati da fornire agli algoritmi successivi [9].

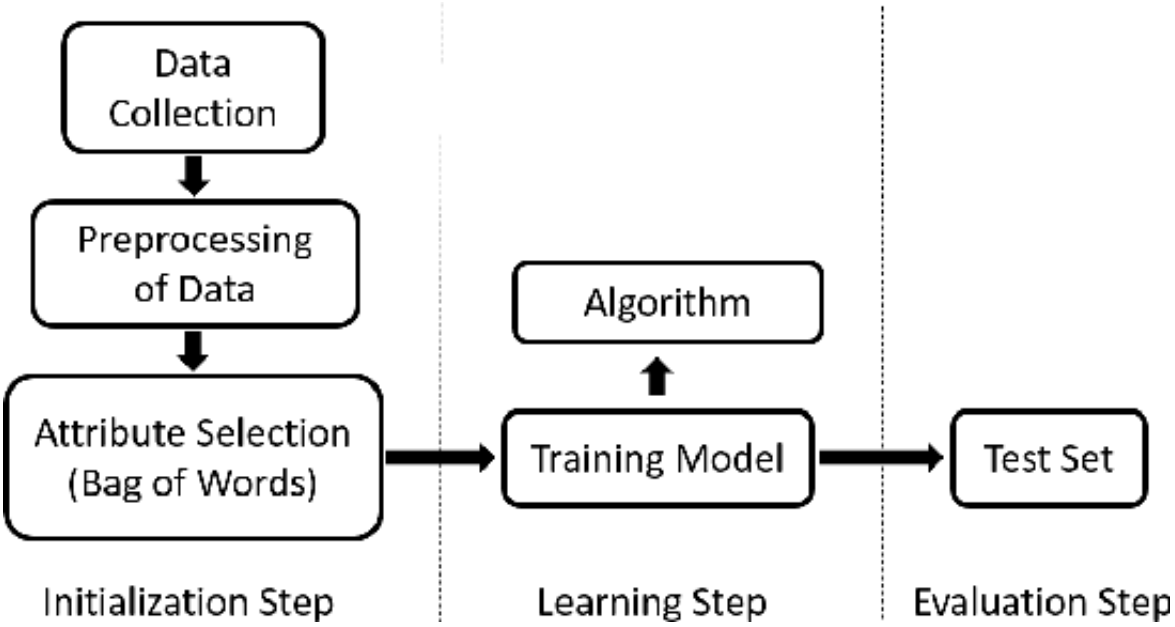

Figura 2.2: Sentiment Analysis Pre-processing

2.2.1 Le principali operazioni:

Le operazioni di Pre-Processing sono necessarie per eseguire qualsiasi funzionalità di data mining. La pre-elaborazione dei dati prevede le seguenti attività:

- **Rimozione delle "stop words"**
  Le *stop words* sono quelle parole che non hanno particolare significato per l'analisi testuale, ma bensì sono utilizzate nel lignuaggio naturale.

Queste variano in base alla lingua di riferimento, nel linguaggio italiano ne sono un esempio le preposizioni (al, quel, di, degli ecc.).

- **Rimozione degli URLs**
  In generale gli URL non vengono considerati nell'analisi in quanto non forniscono indicazioni riguardo al sentimento.

- **Filtering**
  Ovvero l'eliminazione di lettere ripetute (esempio: *"bellissimoooo"*) che le persone usano per aggiungere intensità di espressione. Tuttavia questi tipi di parole vengono considerate alla loro forma "normale". L'eliminazione segue "regola della doppia" secondo la quale una lettera non può ripetersi più di due volte consecutive nella stessa parola.

- **Eliminazione delle domande**
  Le cosiddette "question words", come ad esemio quando, come, perché vengono eliminate, in quanto non contribuiscono alla polarità del sentiment.

- **Rimozione di caratteri speciali**
  Devono essere rimossi tutti i caratteri speciali del tipo:

  ```
  [, {, (, \, ",
  ```

  I quali oltre a non essere indicativi per la polarità rischiano di indurre a erronee traduzioni nella Sentiment Analysis.

Lo scopo è quindi la riduzione del testo in un dato quantitativo tale da potere essere trattato da un modello statistico.
I testi si possono distinguere in base alle loro caratteristiche, ad esempio ci sono dei modelli più adatti a testi brevi ed altri adatti a testi più lunghi. Lo scopo, a prescindere dal tipo di testo, è quello di ottenere una forma simile ad una matrice di dati, eliminando l'informazione relativa all'ordine con cui le parole appaiono nel testo. Si parla quindi di *"bag of words"* cioè dell'insieme di termini senza tenere conto del loro ordine [8][9][10].

2.2.2 Lo stemming

Da qui infatti si può ridurre il testo ad un insieme ridotto di termini detti stilemi (stem). Con stilemi si intende una singola parola (unigram), oppure, se si ritiene importante l'ordine di una coppia di parole (bigram) (ad esempio potrei volere distinguere il termine "bianca casa" da "Casa Bianca"), e così via fino ad arrivare agli n-gram [8].
Questa fase di trasformazione dai testi in stilemi è detta fase di stemming, e può essere realizzata per qualsiasi lingua utilizzando gli appositi strumenti.
Gli stilemi non devono essere necessariamente termini interi, in genere si preferisce la radice fondamentale. Esempio la radice famig. per indicare tutti i termini famiglia, famiglie, famigliare ecc.
Supponiamo di avere i seguenti testi

> testo 1: "il nucleare conviene poichè è economico"
> testo 2: "il nucleare produce scorie"
> testo 3: "il nucleare mi fa paura per le scorie, le radiazioni e l'inquinamento

Analizzando ogni frase e trasformando i termini rilevanti, otteniamo i seguenti steam:

- s1: "nucleare"
- s2: "paura"
- s3: "radiazioni"
- s4: "inquinamento"
- s5: "scorie"
- s6: "inquinamento"

Post	Categoria	s1	s2	s3	s4	s5	s6
$testo_1$	positiva	1	0	0	0	0	1
$testo_2$	neutra	1	0	0	0	1	0
$testo_3$	negativa	1	1	1	1	1	0

Tabella 2.1: Matrice di stemming

La Tabella 2.1 mostra un esempio semplificato di matrice di stemming. Questa rappresenta il punto di partenza di ogni analisi.
E' facile pensare che la matrice di stemming arrivi velocemente a contenere un numero elevato di colonne, ovvero che gli steam siano molto numerosi, tuttavia analisi empiriche dimostrano che le colonne della matrice ovvero gli steam, tipicamente non siano mai più di 500. Quello che invece si presenta come sfida computazionale è il numero di righe della matrice, ovvero il numero di testi da analizzare che può superare facilmente i diversi milioni per ciascuna analisi.

Le criticità dello stemming

Sebbene lo stemming si una tecnica ampiamente utile e utilizzata presenta comunque delle criticità.
Una prima criticità è nella natura intrinseca dello stemming, ovvero il fatto che **lo stemming dipende dalla lingua.**
Quando un documento multilinguistico utilizza terminologie straniere all'interno del corpus (es. terminologie inglesi) lo stemming potrebbe non essere più efficace. In questi casi è necessario disporre di un doppio algoritmo di stem, uno per ciascuna lingua.
Nota: Non è sempre facile riconoscere l'origine di un termine. Ad esempio, la parola "file" appartiene sia al vocabolario italiano che inglese ma con significati diversi (archivio in inglese, coda in italiano)[11].

Una seconda criticità, più problematica sta nella natura invece dei termini, possiamo infatti ritrovare:

1. Termini con stessa radice ma significati diversi
2. Termini con stesso significato ma radice diversa
3. Termini composti da altri termini

Rientrano nel **caso 1** quei termini che pur avendo la stessa radice hanno significati totalmente diversi. Come l'esempio mostrato nella figura 2.3.

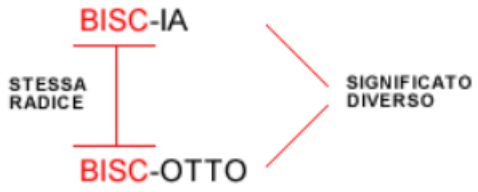

Figura 2.3: Limits of stemming

Rientrano nel **caso 2** quei termini che avendo radice diversa, hanno lo stesso significato. Come l'esempio mostrato nella figura 2.4. Il caso 1 ed il

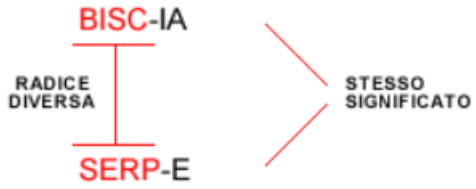

Figura 2.4: Limits of stemming (2)

caso 2 dimostrano come lo stemming sebbene ancora molto utilizzato presenti dei limiti, la cui natura sia intrinseca del linguaggio analizzato.

Rientrano nel **caso 3** i termini composto, che una volta scomposti perdono il loro significato originale, acquisendone un altro. Come l'esempio mostrato nella figura 2.5.

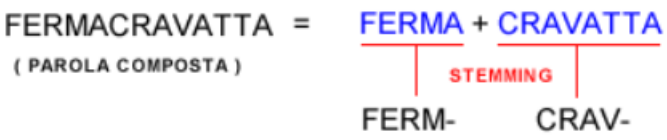

Figura 2.5: Limits of stemming (3)

Per via di queste criticità alle tecniche di stemming vengono associate tecniche di **lemming** .

2.2.3 Il lemming

Definizione di lemma:
Il lemma è la parola intesa come radice morfologica che per convenzione rappresenta tutte le forme di una flessione.[11]

Ad esempio, il lemma delle forme verbali (sono, sei, è, siamo, siete, sono) è il verbo all'infinito (essere).
Il lemma consente di migliorare il processo di matching perché evita i limiti della selezione per radice. Nel caso dei lemmi, la parte iniziale delle parole appartenenti allo stesso insieme può anche differire.

Voce	**Radice**	**Lemma**
Vado	Va-	Andare
Vai	Va-	Andare
Andiamo	And-	Andare
Andate	And-	Andare
Vanno	Va-	Andare

Tabella 2.2: Differenza tra stemming e lemming

Come si nota dalla tabella 2.2, utilizzare tecniche di lemming aiuta a superare le criticità dello stemming.

2.3 Gli approcci della Sentiment Analysis

Terminato il *Text Pre-Processing*, che consiste in tutte quelle operazioni di preparazione del testo, si passa all'applicazione vera e propria della Sentiment Analysis. Essa può essere classificata come segue [12].

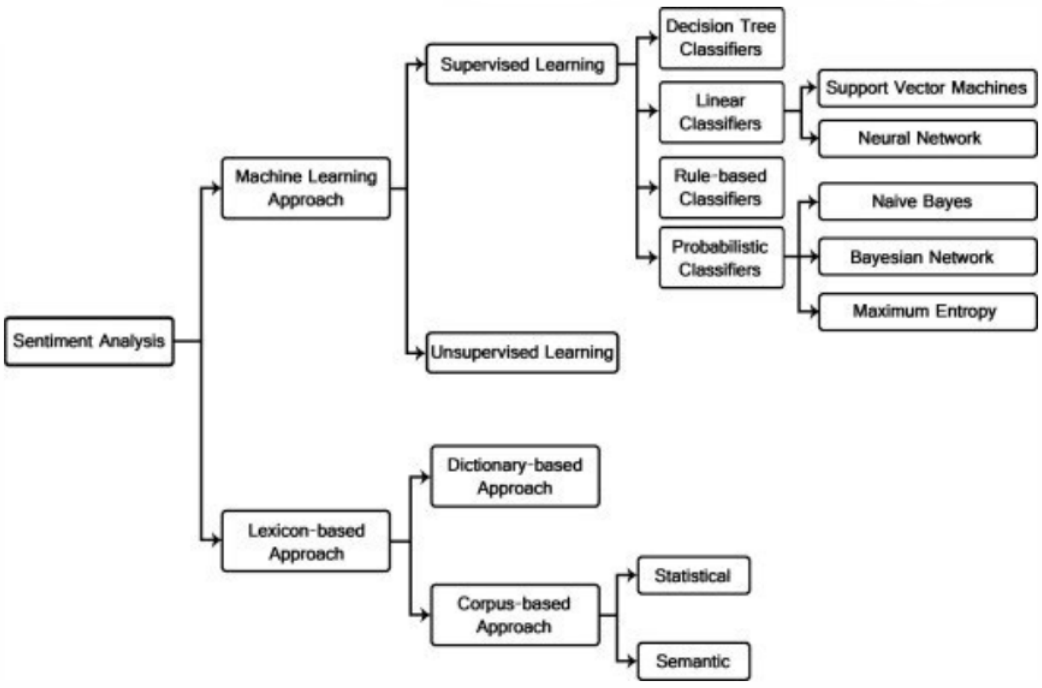

Figura 2.6: Sentiment Analysis Ramifications

Come si nota dalla figura 1.7, la Sentiment Analysis si basa fondamentalmente su due macro approcci, dai quali poi si estendono più sotto approcci.

Questi due macro approcci sono:

- L'approccio basato sul lessico
- L'approccio basato sul machine learnign

In linea generale, **l'approccio lessicale** prevede l'utilizzo di un dizionario con informazioni riguardanti la polarità dei termini o delle frasi.
La polarità complessiva del testo viene poi determinata in base alla polarità dei termini che lo compongono.

Per quanto riguarda invece **l' approccio machine learning**, questo è basata su algoritmi che vengono addestrati a predire la polarità di un testo non noto a priori.

Saranno mostrati ora i principali algoritmi e funzionamenti di entrambi gli approcci.

2.4 Approcci basati sul lessico

Questo approccio è basato su lessici automaticamente o manualmente costruiti.
Come si può notare dalla figura 1.8, l'approccio basato sul lessico si divide in due macro famiglie

- Corpus-based Approach
- Dictionary-based Approach

2.4.1 Approccio basato sui dizionari

L'approccio *Dictionary-Based* é basato sullo sfruttamento di un lessico composto da una lista di termini affiancati da un valore di polarità che ne

indica la connotazione positiva, negativa o neutrale.
Vengono stilate delle liste a partire dalle Opinion Word le quali vengono integrate attraverso l'uso di database, contenenti informazioni sintattico-lessicali, sinonimi e contrari. Tale processo innesca un ciclo che termina nel momento in cui non vengono più trovate nuove parole. Alla fine dell'iterazione viene fatto un controllo manuale per valutare eventuali errori. Prevede indatti l'utilizzo di un dizionario con informazioni riguardanti la polarità di parole o frasi. La polarità del testo viene determinata in base alla polarità dei termini da cui esso è composto.
Questo approccio ha come pro di non necessitare di alcun adattamento. Mentre vede come **contro** il fatto di essere basato sulla coerenza del lessico. Il principale svantaggio di questo tipo di approccio è infatti quello di non riuscire ad ottenere opinion words riguardanti specifici domini o contesti.

Per venire incontro a questa esigenza sono esistono infatti diversi *"vocabolari del sentiment"*, tra i più noti troviamo:

- WordNet
- SentiWordNet (estensione di WordNet)
- WordNetAffect (estensione di WordNet)
- SenticNet

Sotto ne sono mostrati alcuni

WordNet

WordNet è un database semantico-lessicale per la lingua inglese elaborato dal linguista George Armitage Miller presso l'Università di Princeton, che si propone di organizzare, definire e descrivere i concetti espressi dai vocaboli [13].

L'organizzazione del lessico si avvale di raggruppamenti di termini con significato affine, chiamati **"synset"** (dalla contrazione di synonym set), e del

collegamento dei loro significati attraverso diversi tipi di relazioni chiaramente definite. All'interno dei synset le differenze di significato sono numerate e definite.

Il lessico per la lingua italiana è stato sviluppato dall'Istituto di linguistica computazionale del CNR a Pisa [14]. Le relazioni semantiche sono le seguenti e sono suddivise in base alla componente grammaticale in questione, eccone elencati alcuni esempi.

I **sostantivi** godono delle seguenti relazioni:

- iperonimia: Y è un iperonimo di X se ogni X è una specie di Y (Canino è un iperonimo di cane);
- iponimia: Y è un iponimo di X se ogni Y è una specie di X (cane è un iponimo di canino);
- coordinazione: Y è un termine coordinato di X se X e Y hanno un iperonimo in comune;
- olonimia: Y è un olonimo di X se X è parte Y (Palazzo è olonimo di finestra);
- meronimia: Y è un meronimo di X se Y è parte X (finestra è meronimo di window);

I **sostantivi** godono delle seguenti relazioni:

- iperonimia: il verbo Y è un iperonimo del verbo X se l'attività X è una specie di Y (come viaggio rispetto a movimento);
- troponimia: il verbo Y è un troponimo del verbo X se nel fare l'attività Y si fa anche la X (come mormorare rispetto a parlare);
- implicazione: il verbo Y è un'implicazione del verbo X se nel fare X uno deve per forza fare Y (come russare rispetto a dormire);
- coordinazione: Y è un termine coordinato di X se X e Y hanno un iperonimo in comune.

WordNet è stato utilizzato per numerosi scopi nei sistemi di informazione, tra cui chiarimento del senso delle parole, recupero delle informazioni, classificazione automatica del testo, riepilogo automatico del testo, traduzione automatica e persino generazione automatica di cruciverba.
Nonostante il suo ampio utilizzo ne sono stati riconosciuti vari limiti.
Il limite più ampiamente discusso di WordNet è che alcune delle relazioni semantiche sono più adatte a concetti concreti piuttosto che a concetti astratti. Ad esempio, è facile creare relazioni come iponimie, ovvero, catturare che una "conifera" è un tipo di "albero", un "albero" è un tipo di "pianta" e una "pianta" è un tipo di "organismo" e così via, tuttavia risulta molto difficile classificare emozioni come "paura".

SentiWordNet

SentiWordNet è un altro database semantico-lessicale creato da Andrea Esuli and Fabrizio Sebastiani a partire da WordNet. Il loro scopo è quello di realizzare un database da utilizzare per la Sentiment Analysis basata sul lessico.
SentiWordNet applica ad ogni synset tre punteggi di polarità che possono essere *positivo*, *oggettivo*, *negativo*, e la cui somma sia sempre uguale a 1.
In questo dizionario tutti i termini appartenenti allo stesso synset hanno quindi la stessa polarità, e se un termine appartiene a più synset allora, la sua polarità sarà valutata in base al contesto.
Questi tre aspetti possono quindi essere rappresentati come un triangolo suddiviso su due assi [15].

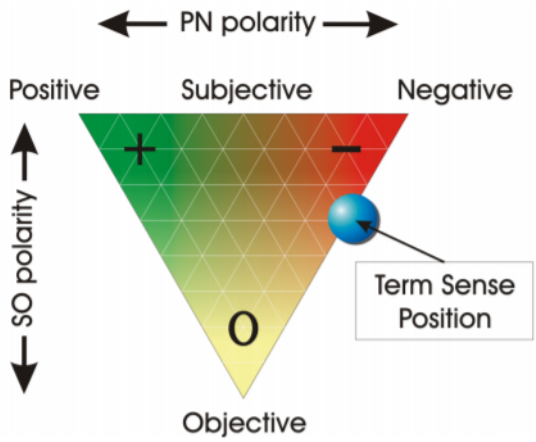

Figura 2.7: SENTIWORDNET for representing the opinion-related

Come mostrato nella figura 1.9, si distinguono due assi:
La **PN-polarity** che indica il grafo positivo/negativo.
La **SO-polarity** che indica il grado di "oggetività" del termine.

Come detto sopra un termine può avere diversi significati, e SentiWordNet deve essere in grado di riconoscerlo.
Eccone riportato un esempio.

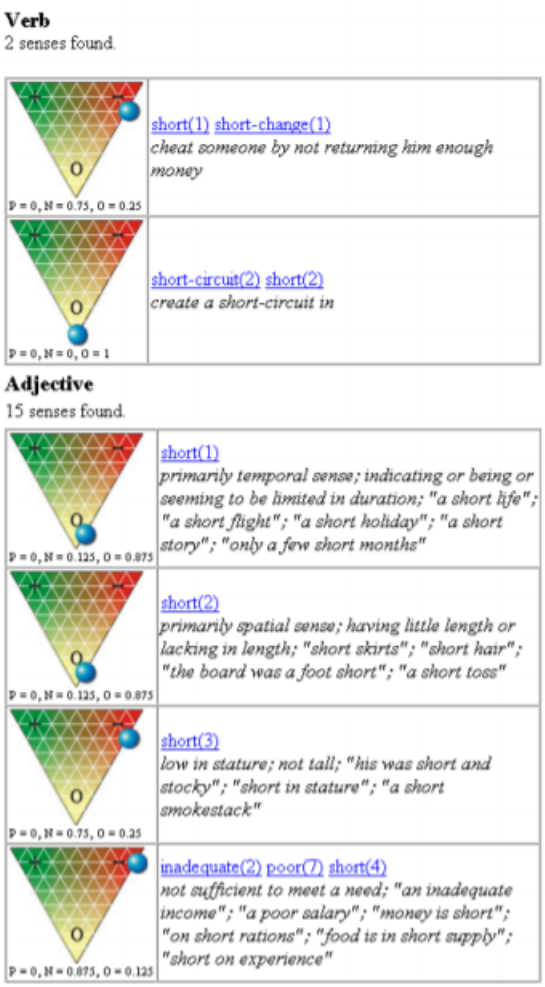

Figura 2.8: 3: SENTIWORDNET visualization of the opinion-related properties of the term short

Come mostrato in figura il termine "short" può avere una moltitudine di significati, e i valori sono differenti in base all'utilizzo che ne viene fatto.

2.4.2 Approccio basato su corpora

Utilizzato per cercare di ovviare i problemi della Sentiment Analysis basata sui dizionari, l'obbiettivo è quello di ottenere delle cosiddette *opinion words* applicate al contesto.
Si utilizzano pattern sintattici o, comunque, schemi che, in concomitanza insieme a delle seed words, consentono di trovare nuove parole all'interno di un corpus e di identificarne l'orientamento [16].
Uno di questi consiste nel partire da una lista si seed words, (solitamente aggettivi rilevanti, in grado di esprimere chiaramente opinioni), e tramite l'utilizzo di vincoli sintattici, solitamente connettivi, (e, come...) per appunto estendere la lista di opinion words. Se infatti si trovano vincoli connettivi come "e", allora i due aggettivi hanno la stessa polarità, viceversa se si trova un vincolo connettivo come "ma", i due aggettivi avranno polarità diversa se non direttamente opposta. Da questi collegamenti si formano grafi a cui poi saranno applicati algoritmi di clustering. Questi si basano sulla possibilità di definire una distanza (intesa come misura di *dissimilarità* tra oggetti che si vogliono classificare).
Un' altra tecnica si basa su una mappatura *feature taxonomy* ovvero una classificazione delle feature in base alle loro frequenze e combinazioni. Queste si riferiscono alla posizione del concetto in una categoria, o meglio in un dominio. Ancora un'altra tecnica si basa sulla collocazione di termini associando in seguito la polarità in base a questo.
In linea di massima gli approcci basati su corpora si sono rilevati meno efficaci degli approcci basati sui dizionari (per via della difficoltà di costruzione del corpus), tuttavia attraverso opportuni metodi statistici o semantici, si riesce a ovviare comunque i problemi del dictionary-based approach.

Approcci Statistici

Tale tecnica si basa sul calcolo delle occorrenze a partire dalle Opinion Word, presenti in una raccolta manuale. L'idea è che se la parola si presenta maggiormente in testi positivi allora avrà una polarità positiva, se si

verifica più frequentemente in testi negativi avrà una polarità negativa e se ha le stesse frequenze allora avrà una polarità neutrale. Attraverso questo ragionamento stesse parole all'interno dello stesso concetto dovranno avere la medesima polarità. Ciò consente quindi di determinare la polarità di un termine analizzandone la frequenza relativa, rispetto a una delle Opinion Word di partenza, con polarità nota. Per fare ciò un tecnica molto utilizzata, derivante dalla *teoria dell'informazione* è la PMI ovvero la **Point-wise Mutual Information**[17].

Cenni di PMI

Con Mutual Information si intende la quantità di informazione su una variabile aleatoria che può essere ricavata osservandone un'altra. Nel caso della Sentiment Analysis ,l'informazione tra le feature e le classi (positiva, negativa, neutra)[18].
Sotto vediamo un esempio di *mutua informazione*, tra un termine "X" ed una classe "Y".

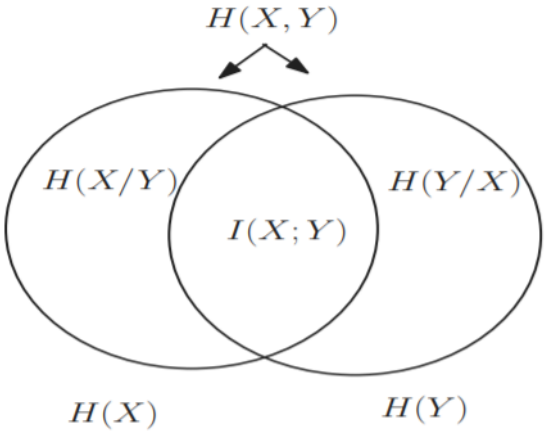

Figura 2.9: chart of mutual information

Ecco la formula della PMI applicata alla Sentiment Analysis

$$M_i(w) = log(\frac{F(w)P_i(w)}{F(w)P_i}) = log(\frac{P_i(w)}{Pi})$$

In questa formula:

- F(w) rappresenta la percentuale di documenti contenente il termine w
- P_i rappresenta la percentuale di documenti contenenti la classe i
- $P_i(w)$ la probabilità condizionata della classe "i-esima" rispetto al termine "w"
- $M_i(w)$ rappresenta la PMI tra il termine "w" e la "i-esima" classe
- $F(w)P_i(w)$ rappresenta la probabilità che vi sia coincidenza tra il termine "w" e la classe "i-esima"
- $F(w)P_i$ rappresenta la reale coincidenza tra il termine "w" e la classe "i-esima"

Approcci Semantici

Questo approccio è molto più diretto rispetto a quello visto precedentemente si basa infatti su principi semantici per analizzare le parole. In particolare sull'idea che due parole che si trovano vicine possano avere la stessa polarità attraverso principi semantici. Tale approccio è alla base dei maggiori database lessici presenti online come i precedentemente descritti WordNet e SentiWordNet.

2.5 Approcci Machine Learning

Entriamo ora nel caso della Sentiment Analysis effettuata tramite *Machine Leraning* ovvero sfruttando l'intelligenza artificiale.
In primo luogo occorre fare distinzione tra due approcci:

- Apprendimento Supervisionato
- Apprendimento Non Supervisionato

Apprendimento Supervisionato

Il sistema acquisisce conoscenza ed esperienza per la classificazione a partire da un *training set* di dati già classificati ed etichettati. Da questo acquisisce conoscenza ed esperienza per classificare i dati successivi

Apprendimento Non Supervisionato

Il sistema acquisisce conoscenza durante la fase di training, svolto con una serie di dati non precedentemente etichettati, e che egli quindi riclassificherà e riordinerà sulla base di caratteristiche comuni. I dati non sono quindi etichettati a priori ma appresi dal sistema in maniera autonoma.

Gli approcci più comuni tuttavia risultano essere quelli Apprendimento Supervisionato

2.5.1 Apprendimento Supervisionato

Come espresso sopra, gli algoritmi basati su Apprendimento Supervisionato necessitano di una pre-etichettatura del set di dati per istruire il cosiddetto *classificatore*.
Diventa quindi necessario che per i documenti da ispezionare siano definite delle proprietà che nel campo della Sentiment Analysis vengono chiamate *features*. Con questo termine si indicano quindi le principali proprietà del testo da ispezionare, evidenziandone quindi il sentiment sottostante.
Vengono riportate sotto le principali feature [19]:

- **I termini e la loro frequenza:** si parla quindi di parole, le quali possono essere singole o concatenazione, espresse in *n-grams*, oltre la loro frequenza, in alcuni casi può essere considerata la relazione d'ordine e la loro posizione

- **Parti del discorso:** consiste in una sorta di etichettatura dei termini, il più delle volte dal punto di vista grammaticale come nomi, aggettivi, avverbi e possono diventare anche molto specifici. E' stato dimostrato infatti da molte ricerche che gli aggettivi sono i più importanti indicatori di opinioni.

- **Opinion words e opinion phrases:** le opinion words sono termini usati comunemente per esprimere opinione diretta, sentimenti positivi o negativi, anche queste possono essere aggettivi, sostantivi, verbi, verbi ecc. Oltre alle opinion words esistono anche le opinion phrases, ovvero frasi, solitamente "modi di dire" che esprimono opinioni in maniera diretta.

- **Negazioni:** le negazione sono di fondamentale importanza poichè la loro presenza stravolge totalmente l'orientamento dell'opinione. Vanno tuttavia gestite con particolare attenzione in quanto la loro presenza non implica necessariamente un cambiamento di opinione.

- **Dipendenze sintattiche:** Si intendono quei rapporti di dipendenza tra termini in cui uno di un termine non potrebbe esistere senza l'altro. Anche questi sono notevolmente importanti in quanto un analisi su un termine di questo genere preso singolarmente sarebbe inutile se non fuorviante.

Come mostrato dallo schema precedente, per quanto riguarda gli algoritmi di machine learning supervisionato sono presenti diversi classificatori suddivisi in varie categorie.

2.5.2 Classificatori probabilistici: Naive Bayes

Si tratta di un algoritmo che utilizza il Teorema di Bayes per prevedere la categoria di un testo. Il Teorema di Bayes descrive la probabilità di una caratteristica, in base alla sua conoscenza precedente rispetto alle condizioni che potrebbero essere correlate a quella caratteristica.
Espressa in maniera più semplice, il Teorema di Bayes descrive il modo in cui le opinione nell'osservare un certo testo A, siano arricchite dopo avere osservato un altro testo B [20].
Fa utilizzo della bag of words ,di cui si è già parlato, la quale ignora la posizione delle parole nel documento (infatti si basa sulla distribuzione di parole nel documento), e del teorema di Bayes per predire la probabilità con la quale una data feature appartenga ad una particolare categoria. Per semplificare la categoria sarà indicata come tag.

$$P(\frac{tag}{features}) = \frac{P(tag) * P(\frac{features}{tag})}{P(features)}$$

- P(tag) rappresenta la probabilitàcon cui un insieme casuale di feattures ricada in quel tag.

- P(features) rappresenta la probabilità che hanno un certo insieme di feautures di apparire insieme.

- $P(\frac{features}{tag})$ rappresenta la probabilitàche una certa feature si classificata con quel tag.

Cenni di Baesyan Network

Le reti bayesiane sono un tipo di modello grafico probabilistico che utilizza l'inferenza bayesiana per i calcoli di probabilità. Le reti bayesiane mirano a modellare la dipendenza condizionale, e quindi la causalità, rappresentando la dipendenza condizionale come nodi in un grafico diretto [21]. Attraverso queste relazioni, si può condurre in modo efficiente inferenza sulle variabili

casuali nel grafico attraverso l'uso di fattori. Applicata alla Sentiment Analysis una rete bayesiana N è rappresentata come una distribuzione grafica della probabilità congiunta tra un insieme di valori casuali variabili.
Il grafico è formato da due componenti G = (Rn, Mr) che rappresentano la distribuzione strutturale di un insieme di variabili:
Rn = x1.....xn sono i nodi del grafico collegate da archi Mr.
E in insieme di distribuzioni di probabilità condizionata P=Pi.....Pn. L'arco diretto tra due variabili Xi Xj rappresenta quindi un un dipendenza condizionale tra le due variabili.
Questo approccio mira a includere informazioni sul sentiment come criteri di dipendenza tra le variabili rappresentate nel grafo.

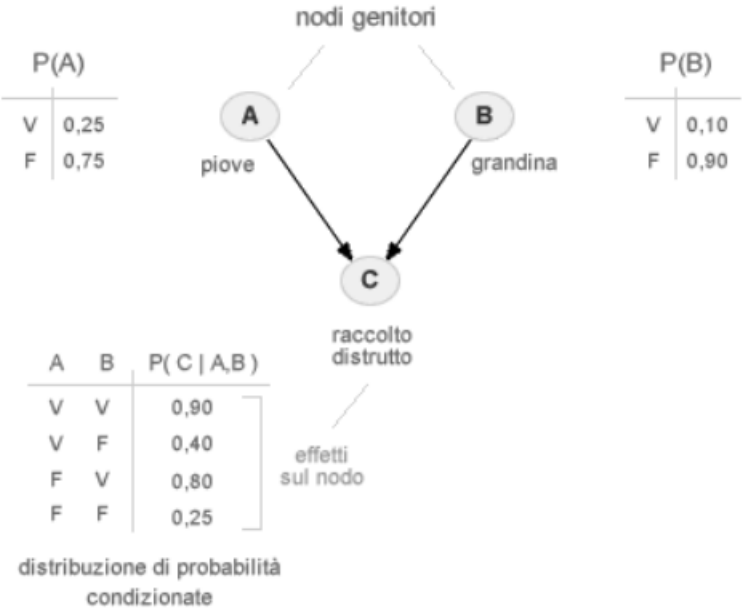

Figura 2.10: An example of Baesyan Network

2.5.3 Classificatori probabilistici: Maximum Entropy

Il classificatore Max Entropy è un classificatore probabilistico che appartiene alla classe dei modelli esponenziali. A differenza del classificatore Naive Bayes visto precedentemente, Max Entropy non assume che le caratteristiche siano condizionatamente indipendenti l'una dall'altra.
Questo classificatore si basa sul Principio della massima entropia e tra tutti i modelli che si adattano ai nostri dati di allenamento, seleziona quello che ha appunto entropia massima [22]. Il classificatore Max Entropy può essere utilizzato per risolvere una grande varietà di problemi di classificazione del testo come rilevamento della lingua, classificazione degli argomenti, analisi del sentiment e altro. Questo sistema garantisce che non vengano introdotti "pregiudizi".
Il classificatore Maximum Entropy viene utilizzato quando non possiamo assumere l'indipendenza condizionale delle funzionalità. Ciò è particolarmente vero nei problemi di classificazione del testo in cui le nostre funzionalità sono generalmente parole che ovviamente non sono indipendenti. Max Entropy richiede più tempo per l'addestramento rispetto a Naive Bayes, principalmente a causa del problema di ottimizzazione che deve essere risolto per stimare i parametri del modello. Tuttavia, dopo aver calcolato questi parametri, il metodo fornisce risultati affidabili ed è competitivo in termini di consumo di CPU e memoria Questo metodo converte gli insiemi di feature in vettori codificati, i quali possono essere combinati per determinare l'etichetta più adatta per un dato insieme di feature [23].
Sotto viene riportata la formula per il calcolo della probabilità di queste etichette:

$$P(\frac{f_s}{feature}) = \frac{dotprod(pesi, encode(f_s, label))}{sum(dotprod(pesi, encode(fs, I))}$$

Il numeratore indica il prodotto scalare tra i pesi e il vettore (ottenuto tramite codifica), mentre il denominatore rappresenta la sommatoria di tutti i prodotti scalari per ogni etichetta.

2.5.4 Classificatori Lineari: Support vector machine

Una macchina vettoriale di supporto (SVM) è un modello di apprendimento automatico supervisionato che utilizza algoritmi di classificazione per problemi di classificazione a due gruppi. Dopo aver fornito un set di modelli SVM di dati di allenamento etichettati per una delle due categorie, sono in grado di classificare nuovi esempi.
L'SVM è basato sull'idea di trovare un iperpiano che divida al meglio un set di dati in due classi.

Definizione Iperpiano:
In uno spazio a r dimensioni, l'insieme dei punti le cui coordinate (cartesiane o proiettive) soddisfano un'equazione lineare. Si tratta di uno spazio lineare, di dimensione r-1, subordinato allo spazio dato. [24]

Definizione Vettore di supporto:
I vettori di supporto sono i punti dati più vicini all'iperpiano. Tali punti dipendono dal set di dati che si sta analizzando e se vengono rimossi o modificati alterano la posizione dell'iperpiano divisorio. Per questo motivo, possono essere considerati gli elementi critici di un set di dati. [25]

Definizione Margine:
Il margine è definito come la distanza tra i vettori di supporto di due classi differenti più vicini all'iperpiano. Alla metà di questa distanza viene tracciato l'iperpiano, o retta nel caso si stia lavorando a due dimensioni. [26]
Immaginiamo quindi di avere due tag (per esempio rosso e blu) e che i nostri dati abbiano due caratteristiche x e y.
Vogliamo un classificatore che, data una coppia di coordinate (x, y), emetta se è rosso o blu. Tracciamo i nostri dati di addestramento già etichettati su un piano cartesiano e dopodichè si traccia l'iperpiano.

Iperpiano linearmente separabile

Non è detto che esista un limite di decisione che separa i valori di una classe dall'altro. Se esiste si parla appunto di iperpiano linearmente separabile.
Dal momento che (nella maggior parte dei casi) esistono infiniti iperpiani, bisogna cercare quello che ha margine più alto con i vettori di supporto, per migliorare l'accuratezza del modello. Infatti, più lontano dall'iperpiano si

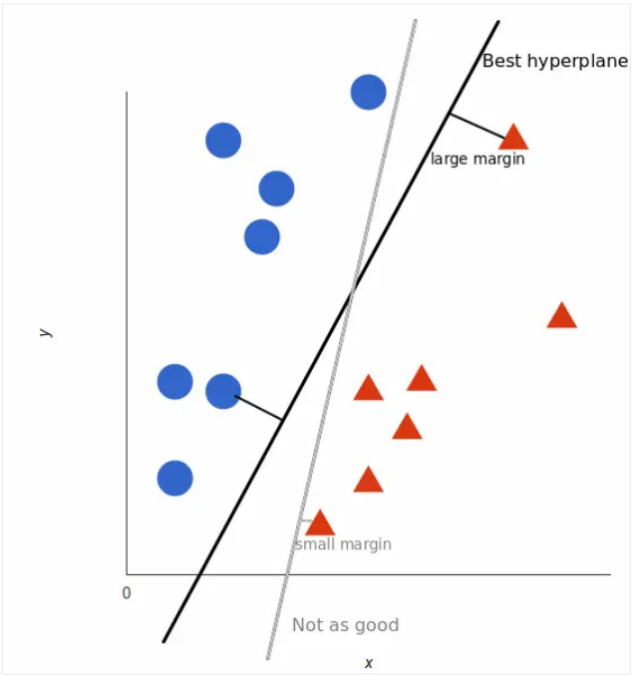

Figura 2.11: Optimal Hyperplane

trovano i nostri punti dati, più siamo fiduciosi che essi siano stati classificati correttamente. Pertanto, desideriamo che i nostri punti dati siano il più lontano possibile dall'iperpiano. [26]
A questo punto si riesce quindi agilmente ad eseguire una classificazione di qualsiasi testo identificando l'iperpiano di appartenenza, basandoci (come in Naive Bayes) sulla frequenza con cui i termini appaiono nel testo attraverso l'utilizzo di apposite *funzioni kernel*.

Tra i principali **vantaggi** della SVM vi sono la su efficacia in dimensioni spaziali elevate, efficienza dal punto di vista della memoria ed un elevata versatilità.
Tuttavia presenta anche alcuni **svantaggi** come la sua difficile interpretazione dei risultati (che può essere facilitata dalle tecniche di visualizzazione grafica) e il fatto che non probabilistica.

Iperpiano non linearmente separabile

Vi è comunque la possibilità di avere anche dataset non lineari, ovvero dove non sia possibile effettuare una retta (per semplificare si ragioni bidimensionalmente) per delineare le due differenti parti.

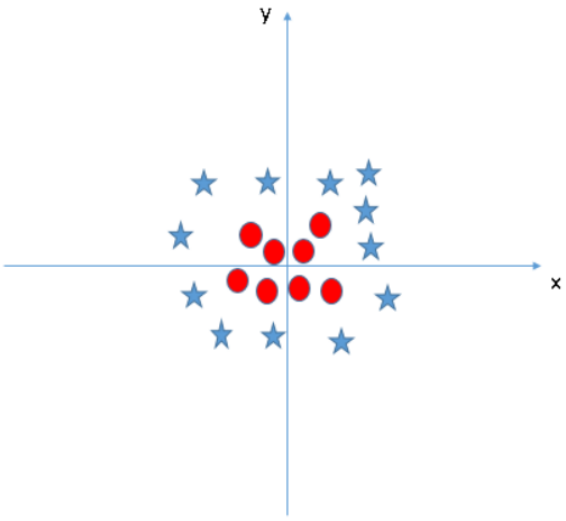

Figura 2.12: Not Optimal Hyperplane

Come nella figura sopra è abbastanza chiaro che non esiste un limite di decisione lineare (una singola linea retta che separa le classi). Nonostante ciò, i vettori sono chiaramente molto segregati e sembra che sia facile separarli. E' possibile infatti aggiungere una terza dimensione. Aggiungiamo quindi alle precedenti dimensioni x e y, una nuova dimensione z, e stabiliamo che venga calcolata in un certo modo per noi conveniente: $z = x^2 + y^2$.
La terza dimensione ci darà uno spazio tridimensionale. Una fetta di questo spazio, può essere rappresentata dalla seguente figura:

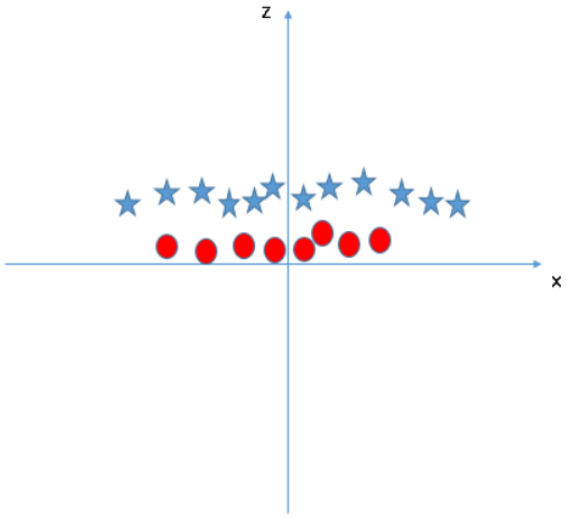

Figura 2.13: Optimal Hyperplane cartesian axes xz

Si noti che poiché ora siamo in tre dimensioni, l'iperpiano è un piano parallelo all'asse x a una certa quota z.
Nel nostro esempio abbiamo trovato un modo per classificare i dati non lineari mappando abilmente il nostro spazio a una dimensione tridimensionale, utilizzando quello che viene definito il metodo Kernel.
Gli algoritmi SVM utilizzano un insieme di funzioni matematiche definite come kernel. Il loro scopo è quello di prendere i dati come input e trasformarli nella forma richiesta qualora non sia possibile determinare un iperpiano linearmente separabile, come avviene nella maggior parte dei casi.

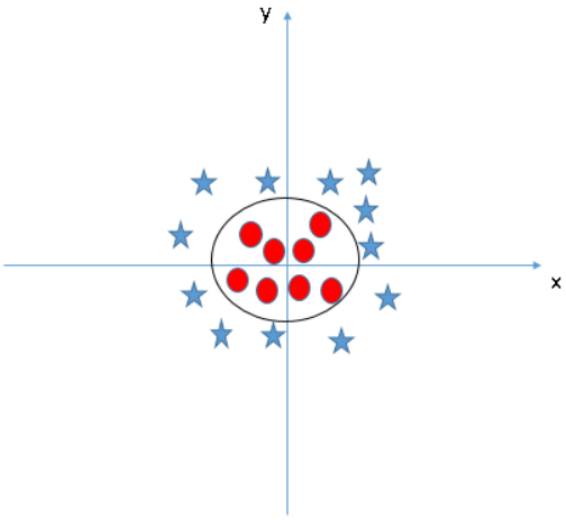

Figura 2.14: Optimal Hyperplane

2.5.5 Classificatori Lineari: Reti Neurali

Le Reti Neurali

Le reti neurali artificiali sono modelli di calcolo matematico-informatici che cercano di simulare le reti neurali biologiche, ovvero sistemi costituiti da migliaia di interconnessioni tra neuroni (sinapsi) che consentono di ragionare e di gestire ogni funzione del corpo.
Allo stesso modo, le reti neurali artificiali sono strutture non lineari di dati statistici organizzate come strumenti di modellazione: ricevono segnali esterni su uno strato di nodi d'ingresso (che rappresenta l'unità di elaborazione, il processore); ognuno di questi nodi d'ingresso è collegato a svariati nodi interni della rete che, tipicamente, sono organizzati a più livelli in modo che ogni singolo nodo possa elaborare i segnali ricevuti trasmettendo ai livelli successivi il risultato delle sue elaborazioni (quindi delle informazioni più evolute, dettagliate) [27]. In particolare, i nodi vengono dislocati su livelli che possono essere di tre tipi:

- Livello di Ingresso (Input Layer): livello progettato per ricevere le informazioni provenienti dall'esterno al fine di imparare a riconoscerle e processarle.
- Livello Nascosto (Hidden Layer): collegano il livello di ingresso con quello di uscita e aiutano la rete neurale ad imparare le relazioni complesse analizzate dai dati. Spesso i livelli nascosti sono più di uno
- Livello di Uscita (Output Layer): livello finale che mostra il risultato di quanto il programma è riuscito a imparare.

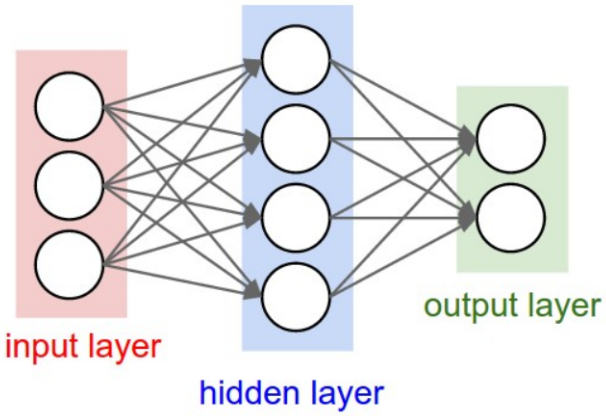

Figura 2.15: Artificial neural network structure

Ogni livello è formato da centinaia, se non migliaia di neuroni artificiali che si interconnettono tra i vari livelli.
La maggior parte delle reti neurali sono completamente connesse, cioè ogni neurone appartenente al livello nascosto risulta connesso con ogni neurone del livello di uscita.
Ad ogni connessione tra neuroni è associato un peso che determina l'importanza del valore di input. I pesi iniziali sono impostati casualmente.
L'esempio più semplice di rete neurale consiste nella Rete FeedForward. In questa rete, il flusso delle informazioni è monodirezionale: quando si impara (attraverso l'addestramento) o quando si opera in condizioni normali (dopo essere stati addestrati) le informazioni schematizzate sono alimentate nella rete dal livello di input. Successivamente sono "sparate" nei livelli nascosti e infine arrivano al livello di uscita.

Ogni livello nascosto riceve i neuroni dalla sua sinistra, e gli input sono moltiplicati per il peso delle connessioni che percorrono.

Ogni livello somma tutti gli input ricevuti in questo modo e se la somma è superiore a un certo valore di soglia, il livello "spara" e attiva il livello connesso alla sua destra.

Apprendimento reti neurali

In questa fase si fornisce alla rete un insieme di input ai quali corrispondono output noti (training set). Analizzandoli, la rete apprende il nesso che li unisce. In tal modo impara a generalizzare, ossia a calcolare nuove associazioni corrette input-output processando input esterni al training set. Man mano che la macchina elabora output, si procede a correggerla per migliorarne le risposte variando i pesi. Ovviamente, aumentano i pesi che determinano gli output corretti e diminuiscono quelli che generano valori non validi. Il meccanismo di apprendimento supervisionato impiega quindi l'Error Back-Propagation. Questo algoritmo prevede di confrontare il risultato ottenuto da una rete, con l'output che si vuole in realtà ottenere e, usando la differenza tra i due risultati, prevede di modificare i pesi delle connessioni tra i livelli della rete partendo dal livello output. In seguito, procedendo a ritroso, l'algoritmo modifica i pesi dei livelli nascosti e infine quelli dei livelli di input. Per far ciò sviluppa una funzione di costo appropriata al problema da risolvere. In definitiva, dal punto di vista matematico, una rete neurale può essere definita come una funzione composta, ovvero dipendente da altre funzioni a loro volta definibili in maniera differente a seconda di ulteriori funzioni dalle quali dipendono, rimane comunque di fondamentale importanza l'esperienza dell'operatore che istruisce la rete. Il motivo risiede nel non facile compito di trovare un rapporto adeguato fra le dimensioni del training set, quelle della rete e l'abilità a generalizzare che si tenta di ottenere. Un numero eccessivo di parametri in ingresso e una troppo potente capacità di elaborazione, paradossalmente, rendono difficile alla rete neurale imparare a generalizzare, perché gli input esterni al training set vengono valutati dalla rete come troppo dissimili ai sofisticati e dettagliati modelli che conosce. D'altro canto, un training set con variabili scarse porta per la via opposta alla stessa conclusione: la rete, in questo caso, non ha sufficienti parametri per apprendere a generalizzare. Il giusto compromesso, insomma, è un compito che necessita di molta preparazione ed esperienza. Le reti feedforward come il MLP utilizzano l'apprendimento supervisionato.

Deep Learning

Il Deep Learning può essere definito una rete neurale artificiale che è composta da almeno 2 livelli nascosti. In realtà le applicazioni di Deep Learning contengono molti più livelli (ad esempio 10 o 20 livelli nascosti). Lo svilup-

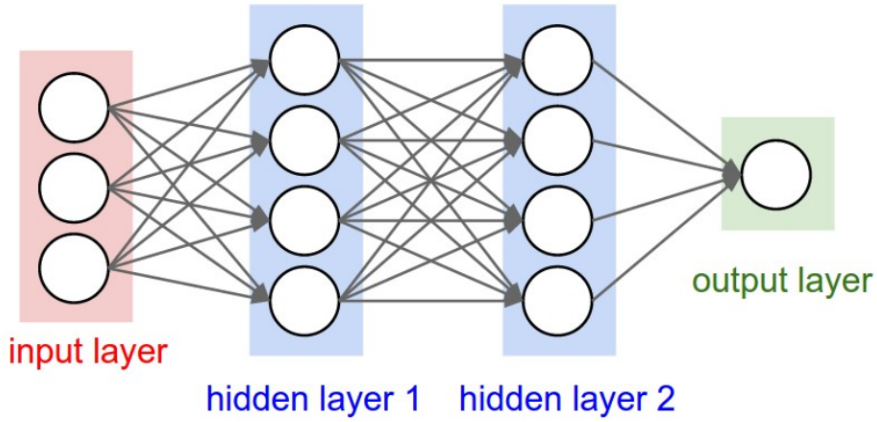

Figura 2.16: Deep learning structure

po di questa nuova tecnologia di apprendimento è dipeso da principalmente due fattori. In prrimo luogo dall'aumento esponenziale dei dati disponibili (i cosiddetti big data), infatti più dati abbiamo a disposizione, più è alto il livello di apprendimento del sistema [28].
In secondo luogo, il notevole incremento delle prestazioni dei computer moderni, che ha permesso migliori risultati con tempi di calcolo notevolmente ridotti.

In sintesi quindi il livello di input riceve i dati di input, Il livello di input passa quindi i dati al primo livello nascosto.
Gli strati nascosti eseguono calcoli matematici sui vari input fornitogli. Una delle sfide nella creazione di reti neurali è decidere il numero di strati nascosti, nonché il numero di neuroni per ogni strato. Una volta che l'informazione arriva all'ultimo livello , livello di output, questo restituirà quindi i risultati

cercando di fare delle specifiche previsioni.
Come anticipato ad ogni connessione tra i neuroni è associata a un peso. Questo peso determina l'importanza del valore di input. I pesi iniziali sono impostati in modo casuale. Quando si cerca di fare una previsioni, alcuni dati in input sono più importanti di altri. Quindi, le connessioni neuronali che partono da tale nodo avranno peso maggiore. Ogni neurone ha una funzione di attivazione. Scopo di queste è quello di "standardizzare" l'output dal neurone.
Tra le funzioni di attivazione più conosciute troviamo la funzione *Unit step* o *funzione a gradino* che normalizza tutti i valori tra 0 e 1, la funzione *sigmoide* molto simile a quella a gradino ma il passaggio da 0 a 1 è più graduale guadagnando stabilità anche per grosse variazioni di valori ed infine la funzione *Rectifier Linear Unit* o più semplicemente *ReLu* che ha la caratteristica di essere molto semplice da calcolare ed appiattire a zero la risposta a tutti i valori negativi, mentre lascia tutto invariato per valori uguali o superiori a zero [29].

Una volta che un insieme di dati ha attraversato tutti i layer della rete neurale, gli restituisce attraverso il layer di output.
Per l'apprendimento della rete è quindi necessario un set di dati di grandi dimensioni e di una grande quantità di potenza computazionale.
Per addestrare l'IA, dobbiamo fornirgli gli input dal nostro set di dati e confrontarne gli output con gli output del set di dati. La prima volta che ciò sarà eseguito, dal momento che l'IA non è ancora allenata, i suoi risultati saranno sbagliati.
Una volta che l'intero set di dati sarà stato eliminato, viene creata una funzione che ci mostra quanto gli output dell'IA siano errati rispetto ai risultati reali. Questa funzione è chiamata funzione di costo.
La funzione di costo $C : F \to \mathbb{R}$ tale che per la soluzione ottimale
$C(f^*) \leq C(f)\ \forall f \in F \forall f \in F$ dove $f^* \in F$ rappresenta la funzione soluzione che risolve il problema in modo ottimale [30].

Idealmente, vogliamo che la nostra funzione di costo sia zero. Ovvero che

gli output forniti dalla nostra rete siano gli stessi del set di dati. Per ridurre al minimo la funzione di costo viene utilizzata una tecnica chiamata Discendenza del gradiente. La discesa del gradiente è una tecnica che ci consente di trovare il minimo di una funzione. Nel nostro caso, stiamo cercando il minimo della funzione di costo [31].

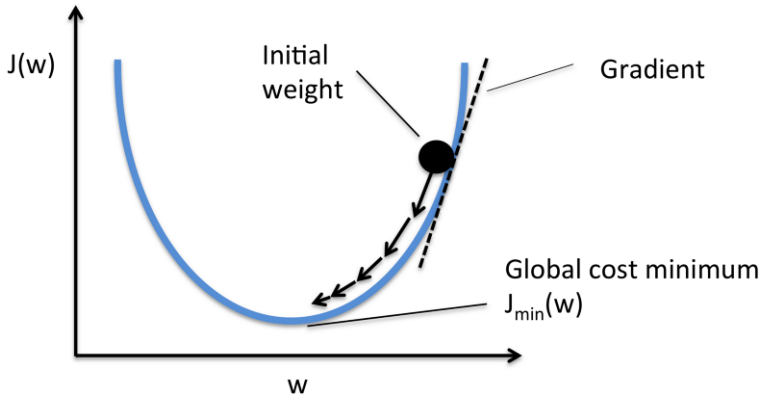

Figura 2.17: Gradient Descent chart

Essa si basa sulla modifica dei pesi in piccoli incrementi dopo ogni iterazione del set di dati. Calcolando la derivata (o gradiente) della funzione di costo con un determinato set di pesi, siamo in grado di vedere in quale direzione si trova il minimo. Per ridurre al minimo la funzione di costo, è necessario scorrere più volte il set di dati. Ecco perché è necessaria una grande quantità di potenza computazionale. L'aggiornamento dei pesi mediante discesa gradiente viene eseguito automaticamente.
Una volta addestrato il nostro strumento per esso sarà in grado di effettuare diverse previsioni

Vantaggi delle reti neurali

L'utilizzo delle varie tipologie di reti neurali nasce dagli importanti vantaggi che presentano:

- Elevato parallelismo, grazie al quale possono processare in tempi relativamente rapidi grandi moli di dati;
- Tolleranza ai guasti, anche questo grazie all'architettura parallela;
- Tolleranza al rumore, ossia la capacità di operare, in molti casi, in modo corretto nonostante input imprecisi o incompleti;
- Evoluzione adattiva: una rete neurale ben implementata è in grado di auto aggiornarsi in presenza di modifiche ambientali.

Capitolo 3

Un esempio di applicazione: Twitter e Bitcoin

I social media riflettono e influenzano sempre più il comportamento di altri sistemi complessi.
In questo capitolo viene analizzato uno studio svolto dall'università di Stoccolma che utilizza l'analisi del sentiment al fine di effettuare previsioni sul futuro. Durante questo studio vengono ispezionate le relazioni tra una nota piattaforma di microblogging, ovvero Twitter e l'andamento della più famosa criptovaluta presente attualmente in circolazione, ovvero il bitcoin.

3.1 Social Network e Criptovalute

3.1.1 Twitter

Twitter si è da sempre contraddistinto come il social dell'immediato, dove quasi "le parole le porta via il vento". In effetti, l'alto numero di tweet (circa 300 mila tweet al giorno) si susseguono in continuazione sulle bacheche degli utenti. Questo è il social che consente di esprimere e diffondere nel modo più veloce e conciso possibile tutto ciò che gli utenti pensano. E, soprattutto, le emozioni che provano.

La caratteristica strettamente collegata a questo modo di esprimersi rapidamente, e che è stata una delle fonte di successo di Twitter, è sicuramente il vincolo del numero di caratteri utilizzabili per scrivere il proprio messaggio. In questo modo gli utenti sono vincolati ad esprimere il loro parere senza giri di parole e in maniera concisa.

Caratteristiche di tweet:

- testi brevi
- hashtag
- neologismi

Considerato che i motivi che spingono le persone a twittare riguardano principalmente reazioni a fatti e notizie, può avere senso ricercare il mood espresso nel testo scritto, andando così a effettuare una vera e propria Sentiment Analysis. Con l'analisi del Sentiment dei tweet è possibile infatti analizzare il loro livello di positività o negatività (polarità) in tempo reale. I testi molto lunghi non sempre sono decifrabili in maniera ottimale, quindi il limitato numero di caratteri imposto da Twitter è sicuramente un elemento che può aiutare l'analisi [32].

3.1.2 Bitcoin

Il Bitcoin è una moneta virtuale, ovvero che non viene stampata come la normale cartamoneta, ma che viene creata, distribuita e scambiata in maniera completamente virtuale, attraverso i computer, e con una tecnologia peer to peer.
La tecnologia che fa funzionare bitcoin, ovvero la blockchain funziona in modo tale da avere una gestione digitalizzata della valuta. Questo significa che la moneta bitcoin non viene creata da una banca centrale, che produce e immette nel mercato nuova moneta (come accade con le monete oggi in tutti i Paesi); al contrario nel caso dei bitcoin le monete vengono conservate all'interno di giganteschi database condivisi (ovvero fisicamente installati su più

computer collegati tra loro alla rete internet), e attraverso sistemi avanzati di crittografia rendono possibile tracciare le transazioni, generare nuove monete, distribuirle e ai proprietari e effettuare transazioni.
Il bitcoin basandosi su tale una tecnologia non si svaluta a fronte dell'immissione sul mercato di nuova moneta, anche perché il sistema complesso rende sempre più difficile risolvere gli algoritmi per verificare ed accettare le transazioni, assicurando così un valore del bitcoin il meno possibile influenzato da svalutazioni date dall'inflazione. Inoltre il suo andamento non dipende da eventi particolari (come la moneta) se non dall'andamento del mercato[33].

3.2 Dichiarazione dei problemi

L'obbiettivo, come precedentemente annunciato è quindi quello di analizzare le informazione da prelevate da Twitter riguardanti i Bitcoin e compararli appunto con le variazioni di prezzo di questi ultimi.
Esiste quindi una correlazione tra il sentiment rilevate su Twitter e la fluttuazione del prezzo BTC? E soprattutto, è possibile produrre un modello di previsione basato sulla Sentiment Analysis di questi?

3.2.1 Precisazioni

Per quanto riguarda l'analisi nello specifico occorre fare alcune precisazioni, in quanto al fine di ottenere un'analisi più specifica è corretta sono state eseguite alcune restrizioni.
In primo luogo la *sentiment classification* è stata limitata al cosiddetto "valore binario", ovvero positivo oppure negativo, senza trattare forme di sentiment più complesse.
Sul lato BTC invece, il valore chiave sarà limitato a un aumento oppure riduzione del prezzo in determinati intervalli di tempo, ignorando il volume e altre metriche chiave.

3.2.2 Raccolta dei dati

Prezzi BTC

I prezzi storici dei Bitcoin sono stati raccolti giornalmente con le API di CoinDesk, disponibili pubblicamente [34].
Sono stati utilizzati diverse lunghezze di intervallo di tempo, a seconda delle quali CoinDesk restituisce diversi livelli di dettaglio.
Un esempio di *Api Call*:

```
\url{http :// api . coindesk . com/cha r t s/data ? output=csv\&data
    =c los e\&index=USD&s t a r tda t e =2017-05-09&enddate
    =2017-05-09&exchanges=bpi&dev=1}
```

Twitter in REAL TIME

Per raccogliere i dati da Twitter ed effettuarvi l'analisi del sentiment è stata usato **Tweepy**, ovvero una libreria open source scritta in Python che consente di accedere alle API Twitter. Questa libreria consente il filtraggio basato su particolari termini e ancora meglio sugli *"hashtag"*. Per questo motivo è stato considerato il modo più adeguato per raccogliere dati che fossero il più pertinenti e mirati possibile [35][36].

Inoltre è necessario effettuare il corretto ed adeguato **filtraggio**, ad esempio il termine "criptovaluta" non è adeguatamente filtrato in quanto se ne potrebbero intendere altre al di fuori di quella considerata nello studio, ovvero il bitcoin. Sono invece considerati adeguatamente filtrati termini come Bitcoin, BTC, XBT...in quanto sinonimi stretti di bitcoin.
Un esempio funzione che raccoglie flusso di tweet filtrati:

```
breaklines
    def btc_tweet_stream():
        api = TwitterAPIConnection()
        listener = StdOutListener()
```

```
stream = tweepy.StdOutListener()
stream.filter(track = ['btc','bitcoin',
'xbt', 'satoshi', languages =['en']])
```

3.3 Il processo

Il processo in questo studio è suddivisibile in tre macro fasi:

- Pulizia dei contenuti ricavati da Twitter
- Applicazione della Sentiment Analysis a livello individuale utilizzando VADER
- Aggregazione del sentiment in base agli intervalli temporali scelti

3.3.1 Riduzione del rumore

Oltre ai dati filtrati di cui si parla sopra, è necessario filtrare i tweet anche dai contenuti generati automaticamente da programmi e bot, in quanto influenti ai fini dell'analisi.
Per l'individuazione di questi è stata usata la seguente strategia. Viene preso un sottoinsieme del dataset contenente tutti i tweet che viene usato come base per trovare attributi comuni che probabilmente sono stati generati da robot o programmi.
Questi tweet vengono quindi scartati. I termini più frequenti di cui si parla sopra vengono passano poi al cosiddetto controllo manuale per identificare schemi sospetti dai quali si vanno a appunto gli N-grammi sospetti. Solitamente rientrano in questa categoria i messaggi generati automaticamente che vogliono convincere l'utente a fare qualcosa (come per esempio convincerlo a comprare bitcoin).

La tabella 3.1 mostra un esempio del processo sopra descritto, come si può notare, partendo da termini singoli si vanno a creare gli N-grammi che con maggiore probabilità sono generati automaticamente per poi scartarli.

Categoria	Esempio
Hashtags	#mgvip, #freebitcoin, #livescore, #makeyourownlane, #footballcoin
Parole	entertaining, subscribe
Bi-grammi	{free, bitcoin}, {current, price}, {bitcoin, price}, {earn, bitcoin}
Tri-grammi	{start, trading, bitcoin}

Tabella 3.1: Esempi di token sospetti

Esempi di tweets scartati
RT @mikebelshe: I'm incredibly risk averse. That's why I have all my money in Bitcoin.
RT @EthBits: EthBits ICO status: https://t.co/dLZk2Y5a88 bitcoins altcoins blockchain ethereum bitcoin cryptocurrency
Margin buying- profitable way of doing online trading tradingbitcoin on margin. $ellBuy https://t.co/aiYYyaCZhK#Bitcoin
RT @coindesk: The latest Bitcoin Price Index is 1241.17 USD https://t.co/lzUu2wyPQN https://t.co/CU1mmkP5mE

Tabella 3.2: Esempi di tweets scartati

Alcuni di questi n-grammi si intersecano con molti altri n-grammi su un token, o un hashtag o una parola. L'insieme dei token identificati costituiscono la base per la costruzione di un filtro. La tabella 3.1 mostra le "variabili" usate per la costruzione di questo filtro, ovvero l'insieme dei token sospetti. Questo filtro combinato con l'eliminazione dei duplicati è stato applicato all'intero set di dati tweets e la dimensione sostanzialmente ridotta passando da 2.271.815 tweets a 1.254.820 [37].
Questa fase consiste nella fase di riduzione del rumore ovvero la fase della pulizia.

3.3.2 La scelta di VADER

VADER (Valence Aware Dictionary e sEntiment Reasoner) è uno strumento *"lexicon based"* di analisi del sentiment basato su regole che è specificamente in sintonia con i sentimenti espressi nei social media [38]. In maniera simile a Senti Word Net (di cui si è già parlato). VADER utilizza una combinazione di un lessico sentimentale è un elenco di caratteristiche lessicali (ad esempio parole) che sono generalmente etichettate in base al loro orientamento semantico come positive o negative. Tuttavia essendo stato studiato specificatamente per i social media risulta più utili ai fini di questa specifica analisi.
VADER è un software completamente open source con licenza MIT.
Tra i principali vantaggi di VADER:

- Funziona eccellentemente sui testi dei social media ma è applicabile anche su altri domini
- Non richiede alcun dato di formazione
- Abbastanza veloce per essere usato online con dati streaming
- Buon compromesso tra velocità e prestazioni

Vediamo ora un esempio di utilizzo di VADER che ha il solo scopo di dimostrarne il funzionamento tramite l'utilizzo del metodo **polarity_scores()** che restituisce gli indici di polarità della frase data:

```
def sentiment_analyzer_scores(sentence):
    score = analyser.polarity_scores(sentence)
    print("{:-<40} {}".format(sentence, str(score)))
```

Passando quindi a questo metodo la frase *"The phone is super cool."* otteniamo il risultato seguente:

```
sentiment_analyzer_scores("The phone is super cool.")
```

```
The phone is super cool----------------- {'neg': 0.0,
'neu': 0.326, 'pos': 0.674, 'compound': 0.7351}
```

Sentiment	**Index**
Positivo	0.674
Neutro	0.326
Negativo	0.0
Composto	0.735

Tabella 3.3: Esempi di indici di VADER

I punteggi positivi, negativi e neutri rappresentano la percentuale di testo che rientra in queste categorie. Ciò significa che la nostra frase è stata valutata come positiva al 67%, neutra al 33% e negativa al 0% la somma di questi deve essere quindi uguale a 1.
Il punteggio composto è una metrica che calcola la somma di tutte le valutazioni del lessico **normalizzate tra -1 e 1**, ovvero tra l'estremo positivo e l'estremo negativo. Nel caso descritto sopra un punteggio composto di 0.735 significa quindi molto positivo.
In generale VADER utilizza la seguente classificazione:

- **positivo** Punteggio composto >= 0.05
- **neutro** Punteggio composto > -0.05 && Punteggio composto < 0.05
- **negativo** Punteggio composto <= -0.05

Come anticipato prima VADER è studiato appositamente per l'utilizzo sui micro-blog in quanto in grado di focalizzare e analizzare aspetti tipici dei social media come, il particolare utilizzo della **punteggiatura**, il quale può aumentare l'intensità (come il punto esclamativo) o aggiungere retorica alla domanda (come il simbolo !?), l'utilizzo di **lettere maiuscole** per dare enfasi a parole rilevanti, o ancora i cosiddetti **modificatori di grado** termini che possono indebolire o rafforzare le affermazioni in questione (termini come *estremamente* oppure *marginalmente*), l'utilizzo di **congiunzioni**, che può segnalare uno spostamento parziale o totale della polarità della frase. Infine di particolare importanza per i social network vi è il riconoscimento delle **emojis** e dello **slang**. Con le emoji si intendono simboli pittografici creati appunto con lo scopo di esprimere sentimento che sarebbe difficile esprimere con il solo testo scritto (esempio dare ironia alla frase) con sleng si intendono invece quei termini, spesso acronimi che esprimono un sentimento (ad esempio LOL per esprimere stupore negativo o TOP per esprimere stupore positivo).

3.4 L'utilizzo dei sentiment

VADER viene quindi utilizzato per ricavare la polarità di ciascun tweet **preso singolarmente**. Come spiegato sopra esso fornisce un punteggio che viene poi normalizzato in un punteggio *composto*, il cui valore è utilizzato per captare la sua negatività, positività o neutralità. Ai fini di questa analisi **i tweet con punteggio neutrale vengono scartati in quanto considerati irrilevanti**.
Ogni riga quindi del file contente set di tweet viene quindi aggiornata, scartando quelle irrilevanti e aggiungendo il punteggio individuale ai tweet rimanenti.
La tabella sottostante mostra come VADER valuta i tweets in analisi come negativi, positivi o neutrali in base al singolo tweet. Questi tweet seppur al solo scopo di fornire un esempio sono stati selezionati, in maniera casuale, dal reale set di dati su cui è stato svolto lo studio.
Si ricordi che i contenuti neutri, ovvero quelli della seconda riga della tabella sottostante vengono scartati in quanto non esemplificativi ai fini dell'analisi.

Classification	**Tweet text examples**
Positivo	:D :D :D ....[Bitcoin performance assessment (+6.18%)] #bitcoin
Neutro	I know somebody who is ALL ABOUT THE BITCOIN
Negativo	CYBER ATTACK FEARED AS MULTIPLE U.S. CITIES HIT WITH SIMULTANEOUS POWER GRID FAILURES OVER LAST 24 HOURS https://t.co/BzWfzlpZrc #Bitcoin

Tabella 3.4: Classificazione tweet con VADER

Dopo il lavoro svolto da VADER, i punteggi individuali di questi tweet vengono raccolti e raggruppati in serie temporali. Queste serie temporali sono una serie di 7 intervalli ovvero:

5min	15min	30min	45min	1h	2h	4h

Tabella 3.5: Intervalli di tempo scelti per l'analisi

Per ognuno di questi gruppi viene calcolata la media del sentiment sulla base dei punteggi dei tweet, in modo da indicare la media per ogni intervallo di tempo scelto.
Dopo questa fase si ottiene quindi un dataset di *sentiment score* ordinati in base ad intervalli di tempo.

3.5 Prerevisione

La previsione in questo campo non è affatto semplice in quanto, dipende da una serie di fattore non facilmente leggibili. Queste dipendono infatti da una combinazione di frequenza, lunghezza e fluttuazioni tra periodi. Nel tentativo di identificare una possibile correlazione tra la variazione del prezzo BTC e del sentiment su Twitter va prestata particolare attenzione al lunghezza e spostamento della frequenza, questa è la motivazione della scelta degli intervalli di tempo, che come si nota sono a breve termine, essi vanno infatti da 5 minuti a 4 ore. Il punteggio del sentiment in un dato periodo viene utilizzato per misurare il tasso di variazione dell'opinione nei periodi successivi. Questa operazione viene eseguita calcolando la differenza tra i punteggi del sentiment nei periodi limitrofi, in questo modo:

- Se il *sentiment change rate* è positivo, allora significa che vi è un aumento del sentimento positivo degli utenti riguardo a BTC.
- Se il *sentiment change rate* è negativo, allora significa che vi è un aumento del sentimento negativo degli utenti riguardo al BTC.

Ogni evento del primo tipo viene classificato come `"1"` e predirrà un incremento del valore dei bitcoin durante il periodo successivo. Viceversa i periodi con crescita negativa del sentiment vengono classificati come `"0"` predicendo un calo del valore dei bitcoin.
Date le previsioni di base il modello crea un vettore binario di previsione per una soglia definita e infine confronta le previsioni con i dati storici dei prezzi.

A questo punto nel set di dati si trova anche il vettore di previsione. Ogni vettore include previsioni riferite ad un determinato intervallo e queste possono essere filtrate in base alla variazione del sentiment in quell'intervallo di tempo in base al *sentiment change rate*, in modo da avere la possibilità di filtrare le informazione con variazioni di sentiment più significativo (ovvero quello con variazioni di sentiment più elevato, sia in crescita che in perdita) rispetto ad un valore soglia.
Le soglie utilizzate in questo studio vanno dallo `0%` al `10%` con un passo dello 0.05%.

Per quanto riguarda il prezzo BTC, il set di dati dei prezzi di cambio USD/BTC contiene aggiornamenti minuto per minuto. Quindi durante l'intero periodo di raccolta dei tweet, i dati dettagliati dei prezzi vengono aggregati alle frequenze indicate nella tabella 3.4. Infine ogni frequenza è classificata come `0` o `1`, a seconda della variazione di prezzo.

3.6 Confronto

Per scoprire l'effettivo rendimento dei vari vettori di previsione, ognuno di questi vettori viene confrontato con i dati storici corrispondenti al periodo in questione. Da queste comparazioni (ovvero il valore predetto con il valore storico) emergono quattro classificazioni:

- True Positive ovvero una previsione positiva corretta
- False Negative ovvero una previsione negativa errata

- False Positive ovvero una previsione positiva errata
- True Negative ovvero una previsione negativa corretta

Se quindi la previsione di un prezzo in crescita coincide con una crescita effettiva del prezzo allora si parla di **True Positive**, viceversa se a una discesa prevista coincide un'effettiva discesa dei pressi storici si parla invece di **True Negative**.
Caso diverso è il caso in cui dato previsto e dato storico non coincidono, in quel caso si parla di **False Negative** qualora ad un decremento previsto dei prezzi corrisponda un aumento effettivo rilevato dai dati storici, mentre si dice **False Positive** se a un incremento previsto corrisponde un effettivo decremento. Viene sotto mostrata la cosiddetta matrice di confusione.

	Predict Increase	**Predict Decrease**
Historical Increase	True Positive	False Negative
Historical Decrease	False Positive	True Negative

Tabella 3.6: Matrice di confusione tra valori reali e valori predetti

3.6.1 Misurazioni

Sulla base della matrice di confusione del paragrafo precedente vengono definite le seguenti funzioni che saranno poi utilizzate per confrontare i risultati ottenuti nei vari intervalli di tempo [39].

L'**Accuracy** misura la percentuale delle previsioni correttamente previste (di crescita o di crescita che siano) rispetto a tutte le previsioni fatte.
La formula è la seguente:

$$\textbf{Accuracy} = \frac{\sum(TruePositive) + \sum(TrueNegative)}{\sum(TotalPopulation)} \tag{3.1}$$

Il **Recall** misura la percentuale delle previsioni (solo) positive correttamente identificate rispetto a tutti gli eventi positivi (solo effettivamente positive, si

noti che al denominatore si trovano infatti anche le previsioni *false negative*).
La formula è la seguente:

$$\textbf{Recall} = \frac{\sum(TruePositive)}{\sum(TruePositve) + \sum(FalseNegative)} \tag{3.2}$$

La **Precision** misura la proporzione tra le previsioni (solo) positive, correttamente identificate, in relazione a tutte le previsioni positive (anche quelle erroneamente positive, come si nota dal denominatore in cui rientrano anche le *false negative*).
La formula è la seguente:

$$\textbf{Precision} = \frac{\sum(TruePositive)}{\sum(TruePositve) + \sum(FalsePositive)} \tag{3.3}$$

F1-score è un indicatore che misura la media armonica tra precision e recall, questa è una misura di accuratezza del test.
La formula è la seguente:

$$\textbf{F1-score} = \frac{2 * Recall * Precision}{Recall + Precision} \tag{3.4}$$

Queste quattro metriche vengono applicate su tutte le combinazioni di frequenza con le rispettive variazioni, in questo modo viene quindi effettuato il confronto tra i vettori di previsione in base ai differenti valori degli intervalli di tempo.

3.7 Risultati

Lo studio è stato eseguito per un totale di 31 giorni, nell'arco temporale dal 11/05/17 al 11/06/17, ovvero un mese. I dati quindi (Tweet e Tasso di cambio USD/BTC sono relativi a questo intervallo di tempo). Una volta al giorno sono stati richiesti dati tramite l'API CoinDesk relativa al valore dei Bitcoin. I dati sono stati restituiti in un fil.*csv* con intervallo di prezzo pari a 1 minuto.

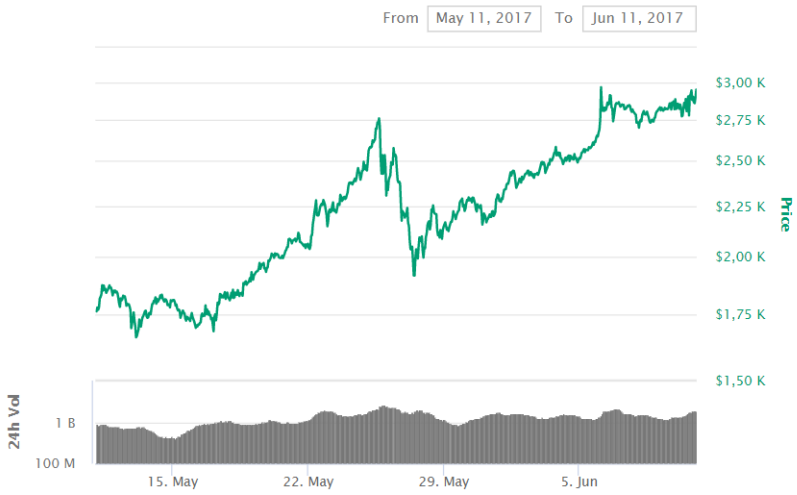

Figura 3.1: BTC/USD chart

Il grafico riporta quindi l'andamento del valore del Bitcoin nell'arco di tempo considerato [40].
Si mostrano ora i risultati che l'analisi del sentiment ha prodotto in questo mese, con particolare attenzione agli intervalli di tempo scelti, i quali si rilevano fondamentali per una previsione accurata. Prima di tutto si noti come le previsioni cambino con l'aumentare della soglia del *sentiment change rate*, infatti più viene incrementata la soglia maggiore è la diminuzione di previsioni (NB: il grafico è su scala logaritmica).

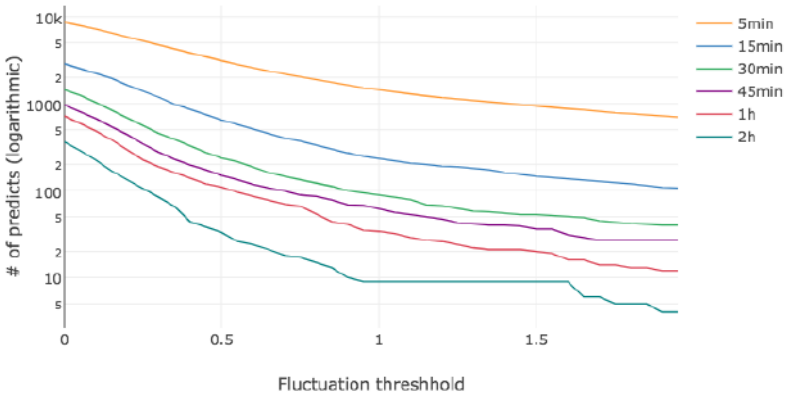

Figura 3.2: Numero di prevsioni per soglia scelta

Di seguito viene ora mostrata la tabella riguardante gli indici sopra descritti (Accurracy, Recall, Precision, F1-score) riguardante i dati predetti confrontati con i dati storici effettivi.
Ne vengono riportate alcune in base a frequenza temporale delle previsioni e alla soglia utilizzata (come spiegato nelle sezioni precedenti dello studio) al fine di individuare i valori ipoteticamente più veritieri qualora si volesse effettivamente effettuare trading BTC tramite l'analisi del sentiment.

Frequency	**Accuracy**	**F1 score**	**Precision**	**Recall**	**Threshold**
1h	0.833333	0.800000	1.000000	0.888889	1.90
30min	0.787879	0.866667	0.722222	0.787879	2.25
45min	0.705882	0.700000	0.777778	0.736842	3.15
4h	0.661017	0.658537	0.818182	0.729730	0.20
2h	0.647059	0.777778	0.636364	0.700000	0.75
5min	0.630137	0.658537	0.675000	0.666667	9.10
15min	0.586207	0.777778	0.636364	0.700000	7.40

Tabella 3.7: Valutazione degli indici per frequenza

Di fondamentale importanza oltre a frequenza e soglia è il numero di previsioni che i dati hanno permesso di effettuare appunto in una determinata frequenza di tempo. In quanto maggiore è il numero di valori predetti maggiore è la validità dell'analisi.
I dati che riportavano numero di predizioni normalizzato inferiore a 10 sono stati rimossi dai risultati.
La seguente tabella mostra il numero di previsioni in relazione agli intervalli di frequenza considerati.

Frequency	**Predicts**
1h	12.0
30min	33.0
45min	17.0
4h	59.0
2h	17.0
5min	146.0
15min	29.0

Tabella 3.8: Valore normalizzato delle previsioni in base all'intervallo di frequenza

Si nota quindi che sebbene le previsioni fatte con frequenza pari ad 1h e soglia di 1.90 mostrino il livello di accuratezza e precisione maggiore, il numero di predizioni normalizzato sia pari a 12, ovvero di sole due unità superiore alla soglia di scartamento.
Di maggiore rilevanza è invece il valore relativo alla frequenza pari 30 min e con una soglia pari al 2.25. In questo caso risulta un numero di predizioni normalizzato superiore (33) con un'accuratezza circa pari al 79%.

3.8 Considerazioni finali

Questa studio cerca quindi una correlazione tra le variazioni di prezzo dei bitcoin in relazione alle percezioni degli utenti di Twitter.
Sono stati analizzati 2,27 milioni di tweet relativi a Bitcoin al fine di identificare fluttuazione del sentiment riconducibili a una variazione del prezzo BTC nel prossimo futuro.
E' quindi stato indicato che l'analisi del sentiment sui dati di Twitter, relativa a Bitcoin, può servire come base predittiva per indicare se il prezzo di Bitcoin potrà subire variazione di crescita o perdita.
Dallo studio è emerso che il sentimento su Twitter cambia in periodi compresi tra 5 minuti e 4 ore, essendo così ristretto questo intervallo di tempo, il sentiment diventa quindi estremamente volatile.
Il risultato principale del modello rimane quello che con la frequenza di `1h` si abbia l'accuratezza maggiore, con soglia scelta pari al `1.9%`, tutta via il numero di previsioni normalizzato si dimostra appena superiore alla soglia di sbarramento il che deve mettere in guardia riguardo alla veridicità di questo risultato.
Nonostante ciò risulta molto interessante il risultato numero 2, ovvero le previsioni effettuate con frequenza pari a `30min` e soglia di variazione pari a `2.25%`. In questo caso il modello a mostrato un'accuratezza pari a circa il `79%`, mantenendo inoltre un numero di predizioni normalizzato più rilevante (33).
Visualizzando le previsioni con frequenza e soglia diverse da quelle sopracitate si nota come l'accuratezza vada a diminuire. I valori di accuratezza più bassi si hanno per quelle previsioni con frequenza più vicina agli estremi (gli intervalli di frequenza scelti vanno da 5 minuti a 4 ore infatti).
Aumentando eccessivamente la frequenza o diminuendola i risultati diventano quindi notevolmente meno accurati, aggirandosi intorno al `60%` di accuratezza, ciò significa che in intervalli troppo brevi o troppo lunghi. questo tipo di Sentiment Analysis, ovvero applicata ai prezzi BTC che sono estremamente volatili non fornisce i risultati interessanti.

3.8.1 Debolezze dell'analisi

Nonostante questa analisi abbia prodotto risultati interessanti occorre sottolineare alcuni sui punti di debolezza, che potrebbero averla influenzata. In primo luogo come si nota dalla figura 3.2 il numero di previsioni cala drasticamente quando si aumenta la soglia di variazione. Inoltre nello studio sono state utilizzate soglie statiche e non dinamiche che avrebbero potuto aiutare per lo meno ad avere un numero di previsioni più alto ad esempio cercando di applicare soglie più appropriate in base al momento della giornata.
In secondo luogo va effettuata una riflessione sull'utilizzo di VADER, sebbene questo abbia come grande punto di forza il fatto di essere elaborato appositamente per il linguaggio utilizzato sui Social Network, non è sicuramente studiato per il linguaggio delle criptovalute, tanto meno dei Bitcoin rischiando quindi di essere un'arma a doppio taglio (termini che comunemente sono considerati come negativi sui social network potrebbero non esserlo se riferiti al mondo delle criptovalute).
In ultimo va sottolineato che la durata dell'analisi è di solo un mese, in cui il bitcoin ha quasi raddoppiato il suo valore, sarebbe opportuno eseguire un'analisi in un periodo di tempo più lungo per controllare la validità di questo metodo.

Conclusioni

In questa tesi sono state esplicate prima di tutto le principali fonti di dati e la loro classificazione in relazione anche alla loro reperibilità. Con una particolare attenzione ai Social Media è stato dimostrato quanto siano utilizzabili per diversi scopi. Alla base dell'argomento principale della tesi (ovvero la Sentiment Analysis) vi è infatti il grande mercato dei Big Data che al giorno d'oggi è stato definito come *il nuovo petrolio*. Risulta infatti fondamentale qualora si applichi l'analisi del sentiment sapere con che tipo di dati si avrà a che fare.
Tuttavia l'enorme espansione dei Social Media, ha dato uno strumento di enorme potenza in mano principalmente alle aziende che lo utilizzano per scopi prevalentemente di marketing, ma non solo, esistono ormai moltissimi tool per effettuare Sentiment Analysis (sia free che a pagamento) alla portata di chiunque come ad esempio Meltwater, Google Alert, Google Analytics, People Browser e molti altri...
Questo per dimostrare che sebbene sia uno strumento poco conosciuto a chi non rientra prettamente nel settore è comunque ormai consolidato.
Partendo dai dati la tesi ha proseguito quindi nel mostrare come questi vengano prima pre-processati poi elaborati dandogli in pasto ai vari algoritmi.
La classificazione principale è quella tra l'approccio basato sul lessico e l'approccio basato sul machine learning con relativi punti di forza e di debolezza dei due approcci. Si ricorda infatti che sebbene l'approccio basato sul lessico non richieda nessun addestramento ne adattamento questo ha lo svantaggio di essere strettamente legato alla coerenza del lessico e quindi di non riuscire

ad adattarsi facilmente a specifici domini.
Per quanto riguarda invece gli approcci machine learning, questi sono si più applicabili a specifici domini ma richiede addestramento da parte del sistema molto costoso in termini di tempo e di complessità. Detto ciò risultano comunque alcuni problemi comuni che sembrano difficilmente ovviabili, questi sono dati dalla natura intrinseca del linguaggio umano che ha ancora troppo sofisticato e ricco di sbavature, rispetto al linguaggio macchina. Si pensa comunque che nel futuro questi problemi verranno risolti dallo sviluppo delle reti neurali, che stanno facendo passi da gigante, non solo per quanto riguarda l'analisi del sentiment, ma nel campo dell'intelligenza artificiale in generale. Esistono tuttavia anche alcuni approcci ibridi che non sono stati mostrati nello studio in quanto obbligavano a ricadere in argomenti di elevata complessità. Infine è stato presentato un caso studio, svolto nell'università di Stoccolma, rivisitato, ed a cui è stata applicata una interpretazione derivante dallo studio precedentemente eseguito e descritto. In questo capitolo è stato eseguito un tool per l'analisi del sentiment, studiato appositamente per i social media. Rientra nella categoria degli approcci basati sul lessico, infatti tra i punti di debolezza dell'analisi è stato citato appunto il fatto che non si riesca ad adattare al meglio allo specifico dominio delle criptovalute, in particolare del Bitcoin. Nonostante i suoi punti di debolezza, questo studio ha dato tuttavia risultati interessanti, dimostrando come se ben applicata la Sentiment Analysis può veramente dare risultati predittivi con elevata accuratezza. Per concludere si può affermare che la Sentiment Analysis, sebbene sia già utilizzata da gran parte della aziende più grandi (Amazon, Google, Apple...) sia comunque in espansione, e si può predire senza troppa immaginazione che ben presto quasi la totalità delle aziende (almeno tra quelle a contatto diretto con il cliente) utilizzeranno questa pratica, sia con scopo valutativo che predittivo.

Bibliografia

[1] A. De Mauro, M. Greco, M. Grimaldi: *A Formal definition of Big Data based on its essential features*

[2] http://www.treccani.it/enciclopedia/web-3-0_%28Lessico-del-XXI-Secolo%29/

[3] https://it.wikipedia.org/wiki/Analisi_del_sentiment.

[4] https://www.studiosamo.it/social-media-marketing/global-digital-2019-statistiche-social/

[5] A. Ceron, L. Curini, S.M. Iacus: *Social Media e Sentiment Analysis: L'evoluzione dei fenomeni sociali attraverso la Rete*

[6] http://www.rainews.it/dl/rainews/articoli/bluedot

[7] M. Prosser: *How AI Helped Predict the Coronavirus Outbreak Before It Happened*

[8] B. Liu, L. Zhang: *A survey of opinion mining and Sentiment Analysis*

[9] G. Angiani, L. Ferrari, T. Fontanini, P. Fornacciari, E. Iotti, F. Magliani, S. Manicardi: *A Comparison between Preprocessing Techniques for Sentiment Analysis in Twitter*

[10] Gagandeep Singh: *Updated Text Preprocessing techniques for Sentiment Analysis*

[11] A. Minini: *Algoritmo di stemming*

[12] W. Medhat, A. Hassan and H. Korashy: *Sentiment analysis algorithms and applications: A survey*

[13] https://wordnet.princeton.edu/

[14] http://www.ilc.cnr.it/

[15] Andrea Esuli and Fabrizio Sebastiani: *SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining*

[16] K. Stuart, A, Botella, and I. Ferri: *A Corpus-Driven Approach to Sentiment Analysis of Patient Narratives*

[17] F. H. Khan, U. Qamar, S. Bashir: *SentiMI: Introducing point-wise mutual information with SentiWordNet to improve sentiment polarity detection*

[18] https://en.wikipedia.org/wiki/Pointwise_mutual_information

[19] https://monkeylearn.com/sentiment-analysis/

[20] https://monkeylearn.com/blog/practical-explanation-naive-bayes-classifier/

[21] https://www.okpedia.it/rete_bayesiana

[22] N. Mehra, S. Khandelwal, P. Patel: *Sentiment Identification Using Maximum Entropy Analysis of Movie Reviews*

[23] V. Vryniotis: *Machine Learning Tutorial: The Max Entropy Text Classifier*

[24] http://www.treccani.it/enciclopedia/iperpiano_%28Dizionario-delle-Scienze-Fisiche%29/

[25] http://www.treccani.it/enciclopedia/vettoredisupporto_%28Dizionario-delle-Scienze-Fisiche%29/

[26] L. Govoni: *Algoritmo Support Vector Machine*

[27] http://www.intelligenzaartificiale.it/reti-neurali/

[28] J Brownlee: *What is Deep Learning?*, Machine Learning Mastery

[29] M. Bicego: *Riconoscimento e recupero dell'informazione per bioinformatica Reti Neurali*, Università di Veronsa, Bioinformatica

[30] https://it.wikipedia.org/wiki/Rete_neurale_artificiale

[31] L.Govoni: *Algoritmo Discesa del Gradiente*

[32] https://monkeylearn.com/blog/sentiment-analysis-of-Twitter/

[33] https://it.wikipedia.org/wiki/Bitcoin

[34] https://www.coindesk.com/coindesk-api

[35] Rishabh Bansal: *Tweet using Python*

[36] https://www.tweepy.org/

[37] E. Stenqvist, J. Lönnö: *Predicting Bitcoin price fluctuation with Twitter Sentiment Analysis*

[38] https://github.com/cjhutto/vaderSentiment

[39] R. Joshi: *Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures*

[40] https://coinmarketcap.com/it/currencies/bitcoin/