Alma Mater Studiorum · Università di Bologna

Corso di Laurea in Scienze Statistiche Finanziarie e Attuariali
Sede di Rimini

L'effetto dei Tweet sulla volatilità del mercato finanziario

Relatrice:
Prof. Cristina Bernini

Presentata da:
Francesco Gabellini

Anno Accademico 2019/2020

Indice

Introduzione

Questa tesi nasce dal mio interesse, sviluppato dall'esperienza sul campo dell'analisi dei dati, verso i dati non strutturati e le nuove tecniche di data mining create ad hoc per risolvere tali problematiche. Negli ultimi decenni, grazie alla diffusione dei social media, la quantità e la tipologia di dati a disposizione della ricerca scientifica sono aumentate esponenzialmente. In tali dati sono contenuti i pensieri, le riflessioni e i commenti di milioni di persone, quindi è innegabile che celino un contenuto altamente informativo. In questa tesi in particolare si è scelto di utilizzare i dati provenienti da Twitter.
Quello che verrà analizzato è la correlazione tra la volatilità dell'indice FTSE-MIB e ciò che viene discusso nei Tweet che fanno riferimento ad aziende incluse nel paniere dell'indice.
Tale idea nasce dall'intuizione per cui la volatilità del mercato finanziario italiano è influenzata sia dalle aspettative dei diversi operatori che ne fanno parte, sia dal flusso di notizie e dalle reazioni ad esse relative. Perciò viene naturale pensare che ci possa essere una qualche utilità informativa nelle opinioni liberamente scambiate sulla piattaforma di Twitter dagli utenti interessati al mercato.
Nella prima sezione del capitolo due verrà chiarito come viene approssimata la volatilità utilizzata in questa ricerca e quali sono altri possibili approcci alla modellistica della volatilità del mercato azionario. Una volta chiarita la definizione di volatilità, verrà trattato come trasformare le opinioni espresse nei Tweet in un indice di sentiment, così da permetterci di utilizzare tali informazioni come una tradizionale serie storica. A tale scopo, ci avvarremo di diverse metodologie di natural language processing, che verranno eseguite utilizzano il software R.
In questa tesi saranno utilizzati due degli approcci più comuni nel mondo della sentiment analysis, il primo è quello di utilizzare dizionari ricavati da altre ricerche e il secondo è quello di usufruire dei pesi di una rete neurale allenata su un set di dati specifico. Il problema è affrontato utilizzando entrambi gli approcci perché non esiste una metodologia che a priori ci può fornire l'aspettativa di un risultato più accurato.
Successivamente verranno presentate le metodologie di analisi di serie storiche e saranno esposte alcune ricerche simili che hanno utilizzato come fonte dati Twitter per l'estrazione di informazioni utili allo studio del comportamento dei mercati finanziari. Una volta ottenuto l'indice che rappresenta il sentiment nel tempo e approssimata la volatilità nel mercato, nel capitolo successivo sarà indagata l'esistenza di una correlazione tra il sentiment e la volatilità del FTSE-MIB di Milano. Infine si procederà ad analizzare la quota informativa offerta da questo indice di sentiment per migliorare le previsioni sulla volatilità, utilizzando come benchmark i modelli tradizionali.

Capitolo 1

Letteratura di riferimento

Gli economisti sono spesso interessati a capire le relazioni tra diverse variabili, come l'indice di inflazione e l'occupazione. L'econometria è la materia che cerca di quantificare queste relazioni sulla base dei dati disponibili e utilizzare diverse tecniche statistiche per interpretare i risultati ottenuti. Di conseguenza, l'econometria è l'interazione di teoria economica, dati osservati e metodologia statistica. Tradizionalmente, l'econometria si concentra sullo studio delle relazioni macroeconomiche, che permettono di fornire strumenti per le valutazioni e le previsioni di politiche economiche. Di recente il focus si è spostato sui modelli microeconomici che permettono di modellare il comportamento individuale o aziendale. Un ulteriore campo di utilizzo sono i mercati finanziari, in cui l'applicazione di modelli è diventata centrale sia in termini di pricing, sia nella gestione del rischio. Dal punto di vista operativo, quando si deve verificare una determinata teoria riguardante una specifica relazione tra diverse variabili, si formalizza in termini matematici la suddetta relazione e poi se ne verifica la validità empirica tramite un modello statistico. Tale modello deve contenere una qualche misura di incertezza per tener conto dei comportamenti che non sono prevedibili o casuali. In questo elaborato i modelli econometrici verranno quindi utilizzati allo scopo di verificare l'ipotesi per cui il sentiment relativo ai Tweet potrebbe prevedere in qualche misura la volatilità dell'indice di borsa italiano. Comprendere e modellare la volatilità è di fondamentale importanza nei modelli statistici, perché ci offre una misura dell'incertezza del fenomeno sotto esame. A sua volta, misurare l'incertezza di un fenomeno ci permette di capire quanto tale fenomeno sia prevedibile e quanto le realizzazioni si discostino dalla media. In più, avere una misura dell'incertezza ci fornisce anche la possibilità di capire se determinati effetti o parametri siano significativi o meno. Perciò è fondamentale avere una misura affidabile che rappresenti l'incertezza, sia per prevedere un determinato fenomeno, sia per modellarne i comportamenti rispetto ad altri.

1.1 La volatilità nei mercati finanziari

L'evidenza empirica derivante dallo studio dei mercati finanziari ha portato alla luce alcuni fatti fondamentali sulle serie storiche finanziarie [1] :

- Le serie dei prezzi sono processi integrati.
- Le serie dei rendimenti sono processi stazionari.
- I rendimenti non sono autocorrelati.
- Il quadrato dei rendimenti mostra autocorrelazioni rilevanti.

Le serie storiche economiche, in particolare quelle di indici finanziari, riscontrano spesso periodi con alta o bassa concentrazione di volatilità. Per queste tipologie di serie storiche è essenziale proporre un modello che riesca a cogliere una volatilità che cambia nel tempo. Questo fenomeno è noto da molto tempo, già nel 1963 Mandelbrot affermava che "Large changes tend to be followed by large changes - of either sign - and small changes tend to be followed by small changes".
Nei mercati finanziari la volatilità è una misura della variazione del prezzo di un indice in uno specifico intervallo di tempo. Esistono fondamentalmente due definizioni di volatilità, quella storica, che deriva dalla effettiva serie dei prezzi misurati nel passato e quella implicita, che deriva dal prezzo di mercato delle opzioni con scadenze future ma scambiate oggi.
È interessante notare che, mentre il valore della volatilità storica è facilmente ottenibile, non esiste una formula diretta per calcolare la volatilità implicita. Quest'ultima non viene infatti calcolata, ma ricavata ripercorrendo all'indietro la procedura di pricing delle opzioni. Una stima della volatilità è richiesta come uno degli input nel modello di pricing delle opzioni di Black-Scholes. Di seguito è mostrato l'esempio del pricing di un opzione call:

$$C = S_o e^{-qt} N(d_1) - X e^{-rt} N(d_2) \tag{1.1}$$

Dove:

- $d_1 = \frac{ln(\frac{S_t}{K}) + (r + \frac{\sigma_v^2}{2})t}{\sigma_s \sqrt{t}}$
- $d_2 = d_1 - \sigma_s \sqrt{t}$
- C = Prezzo dell'opzione call scambiata sul mercato
- S = Prezzo del sottostante
- K = Strike

- r = Tasso Risk-Free
- t = Orizzonte temporale
- N = Distribuzione normale

Quindi conoscendo il prezzo di vendita di una determinata opzione call scambiata sul mercato possiamo ricavare quale σ_s è stato supposto per prezzare tale prodotto. Tale valore è appunto la volatilità implicita.
Molti dei paper che verranno citati in seguito utilizzano un indice di volatilità implicita molto popolare nell'analisi dei mercati finanziari, il VIX. Il VIX, acronimo di Volatility Index, è l'indice che rappresenta la volatilità implicita a 30 giorni delle opzioni scambiate sull'indice S&P 500.
La superficie di volatilità implicita spesso raggiunge un valore medio leggermente più alto della volatilità storica. Questo fenomeno è in parte spiegato dal fatto che la volatilità implicita comprende un premio che i venditori di opzioni si assegnano per assorbire il rischio del vendere tali strumenti.
Pertanto in questa tesi si utilizzerà un indice di volatilità storica in quanto quella implicita non rappresenta a pieno la volatilità del mercato ma una sua approssimazione che è ricca di assunzioni e distorsioni dovute al fatto che viene ricavata da un modello di pricing e non da un indice creato ad hoc per osservarla. Tale approccio sarebbe corretto a fini di pricing di ulteriori prodotti derivati ma non per il fine di ricavare il migliore stimatore possibile della volatilità dell'indice.
Una volta chiariti i motivi per cui siamo alla ricerca della volatilità storica, ci si trova di fronte a due ulteriori approcci dal punto di vista della modellistica statistica. La prima scuola di pensiero è quella dei modelli ARCH (Auto-Regressive Conditional Heteroskedasticity), la seconda è quella dei modelli di volatilità realizzata.
Partiamo ad analizzare brevemente il pensiero della prima corrente proposta. Il modello ARCH è stato presentato per la prima volta nel 1982 da Robert Franklin Engle [2]. Tale modello si basa sull'intuizione per cui essendo le serie storiche finanziarie caratterizzate da cluster di volatilità, questi dovranno necessariamente essere correlati serialmente, cioè la volatilità di oggi deve essere condizionata a quella di ieri. Dal punto di vista econometrico questo discorso si traduce nel fatto che la volatilità mostra dinamiche autoregressive nel tempo. Perciò tali modelli si sviluppano sull'analisi della volatilità condizionata ai tempi precedenti.
I modelli di tipo ARCH rappresentano quindi le espressioni analitiche per la varianza condizionale, come riportato di seguito:

$$P_t = \sigma_t \epsilon_t \tag{1.2}$$

$$\sigma_t = \sqrt{\alpha_0 + \alpha_1 P_{t-1}^2} \tag{1.3}$$

Dove:

- σ = Volatilità condizionata.
- P_t = Valore dell'indice al tempo t.
- P_{t-1} = Valore dell'indice al tempo t-1.

Per quali ragioni tali modelli non sono ottimali nella stima della variabilità? Innanzitutto perché il modello presuppone che gli shock positivi e negativi abbiano lo stesso effetto sulla volatilità, mentre in pratica è noto che i prezzi rispondono in maniera molto diversa a shock positivi e negativi. Inoltre la struttura teorica del modello presenta un problema fondamentale perché la varianza condizionale è una variabile latente, e quindi non è direttamente osservabile. Ciò crea grossi problemi nella stima della validità e dell'affidabilità di tali modelli. Quindi in questo elaborato si è preferito optare per il secondo approccio, cioè quello della volatilità realizzata che ci fornisce uno strumento osservabile nel tempo, direttamente misurabile e modellabile. Il concetto di volatilità realizzata, introdotto nei primi anni 2000 (Andersen et al., 2001, 2003)[3], ha come idea fondamentale quella di calcolare le stime della volatilità utilizzando le informazioni sulle frequenze giornaliere, in modo tale da rappresentare la volatilità tramite dati empirici.
La Formula più semplice che si può utilizzare per stimare la volatilità realizzata utilizza come base la formula tradizionale per la deviazione standard e la modifica in alcuni punti chiave. In primo luogo, viene assunta la media uguale a zero perché si suppone che i rendimenti abbiano un andamento stazionario. In secondo luogo,si imposta il fattore di annualizzazione con una costante, 252 rappresenta il numero di giorni di trading in un anno tipico negli Stati Uniti. A causa delle festività e dei diversi giorni di calendario che si possono avere in un particolare anno, il numero effettivo di giorni di apertura dei mercati può essere leggermente superiore o inferiore a quello fissato. Tuttavia, è preferibile avere una costante approssimativa piuttosto che diversi valori per ogni anno e per ogni nazione. Infine, essendo il risultato tipicamente inferiore a uno, il valore viene moltiplicato per cento così da essere espresso in percentuale, come è comune fare tra gli operatori finanziari.

$$\sigma = 100\sqrt{\frac{252}{n}\sum_{t=1}^{n}(\ln(\frac{P_t}{P_{t-1}}))^2} \tag{1.4}$$

Dove:

- σ = Volatilità realizzata.
- 252 = Costante che rappresenta i giorni di trading annuali.
- t = Contatore temporale per ogni giorno

- n = Numero di giorni nell'orizzonte temporale analizzato.
- P_t = Valore dell'indice al tempo t.
- P_{t-1} = Valore dell'indice al tempo t-1.

La metodologia di stima precedentemente presentata è una buona approssimazione della volatilità realizzata, ma esistono approssimazioni migliori che prendono in considerazione anche gli avvenimenti intraday, come lo stimatore di Garman-Klass [4].
Garman-Klass è uno stimatore di volatilità che incorpora prezzo di apertura e chiusura, e prezzo massimo e minimo della giornata. La volatilità di Garman-Klass estende la formula precedente tenendo conto del prezzo di apertura e degli estremi giornalieri (massimo e minimo). Poiché i mercati sono più attivi durante l'apertura e la chiusura, la stima che se ne ricava rappresenta in maniera più fedele la reale volatilità giornaliera. Inoltre è intuitivo che il massimo e il minimo siano informazioni forti in termini di stima della variabilità dell'indice.
Di seguito viene presentata la formula di Garman-Klass implementata per il calcolo della volatilità realizzata.

$$\sigma = \sqrt{\frac{252}{n}\sum_{t=1}^{n}\left[\frac{1}{2}\left(\log\frac{H_t}{L_t}\right)^2 - (2\log 2 - 1)\left(\log\frac{C_t}{O_t}\right)^2\right]}. \tag{1.5}$$

Dove:

- σ = Volatilità realizzata.
- 252 = Costante che rappresenta i giorni di trading annuali.
- t = Contatore temporale per ogni giorno.
- n = Numero di giorni nell'orizzonte temporale analizzato.
- H_t = Valore di massimo al tempo t.
- L_t = Valore di minimo al tempo t.
- C_t = Valore di chiusura al tempo t.
- O_t = Valore di apertura al tempo t.

Riassumiamo ora brevemente questa rassegna degli strumenti per modellare la volatilità. Per quanto riguarda la variabile obiettivo, si è scelto di optare per la volatilità realizzata, rispetto a quella implicita o condizionata, perché tale volatilità approssima in maniera più fedele il mercato ed è possibile osservarla direttamente nel tempo.

Ciò ci permette di creare un modello diretto e di effettuare diversi test trattandola come una qualsiasi altra serie finanziaria senza particolari assunzioni a priori sul processo. Nella sezione riguardante le serie storiche saranno approfonditi i modelli utilizzati per descrivere il comportamento della serie qui definita. Una volta definita la variabile dipendente dell'elaborato, cioè la volatilità, nelle prossime due sezioni saranno definite le variabili indipendenti utilizzate come regressori.

1.2 Ricerche precedenti

In questa sezione passeremo in rassegna la letteratura che si è occupata di utilizzare i Tweet per ricavare informazioni utili alla modellazione dei comportamenti sui mercati finanziari.
Per primo va citato il paper che ha maggiormente influenzato questo elaborato, cioè "Twitter mood predicts the stock market" di Johan Bollen, Huina Mao, Xiaojun Zeng [5].
In questo paper Bollen e colleghi analizzano un campione di Tweet pubblicati tra il 28 febbraio e il 19 dicembre 2008. Questi Tweet vengono puliti dalle stopwords e trasformati in singole parole per ogni giorno. Successivamente, viene applicata la sentiment analysis tramite un software chiamato Opinion Finder e infine vengono sommati i valori giornalieri di sentiment e creato uno score standardizzato in base alla media e alla varianza settimanali.
Una volta ottenuto questo score per ciascuno dei sentiment forniti dal software, è stata misurata la capacità predittiva di queste serie tramite il Granger Test, fino a 7 giorni di ritardo. Dal paper risulta che la sentiment binaria ottenuta da Opinion Finder è significativa nel fornire capacità previsiva in relazione ai rendimenti dell'indice del mercato statunitense, cioè il Dow Jones.
Questo paper è stato all'avanguardia per l'intuizione di utilizzare degli indici di sentiment provenienti dell'analisi dei social per comprendere l'andamento del mercato. Tuttavia, l'approccio è ancora un po' acerbo perché non vengono filtrati in alcun modo i Tweet provenienti dal social, quindi si potrebbe incorrere in falsi segnali provenienti da discussioni o argomenti che nulla hanno a che fare con l'andamento dei maggiori indici dei mercati finanziari. Per quanto riguarda invece la metodologia, l'approccio di studio delle relazioni tra le diverse serie tramite il Granger test risulta un'idea molto valida perché è difficile dimostrare relazioni causali tra le serie, anche se l'unico fattore utile che ne deriva per l'analisi del mercato finanziario è la capacità di fornire del segnale utile per prevedere l'andamento futuro dell'indice. Un altro studio che ha dimostrato delle correlazioni significative tra i Tweet e il mercato azionario è "Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear"" di Xue Zhang, Hauke Fuehres e Peter A. Gloo [6].
In questo articolo vengono analizzati per sei mesi diversi indici del mercato azionario

americano: Dow Jones, Nasdaq e S&P 500. Tali indici sono confrontati con un campione di tweet che rappresenta un millesimo del totale dei tweet scambiati giornalmente. La raccolta dei dati è avvenuta dal 30 marzo al 7 settembre 2009, con un range di tweet che oscillava tra gli 8100 e 43040 al giorno. Una volta ottenuti i tweet, i ricercatori non hanno utilizzato un dizionario, come nel caso della ricerca di Bollen, ma hanno semplicemente utilizzato una lista di parole che posseggono un significato chiaro dal punto di vista del sentiment. Vediamo ad esempio hope, cioè speranza, oppure happy, cioè felice. Una volta individuate queste parole chiave hanno creato un indice dividendo il numero di tweet contenenti tale parola con il numero totale di tweet campionati in quella giornata. Dalla ricerca si evince che alcuni degli indici di sentiment individuati risultano significativamente correlati con alcuni degli indici di mercato presentati. Inoltre va notato che questo paper, nella sua sezione finale, presenta un elemento di novità rispetto a quello di Bollen, perché dimostra che l'indice di sentiment riferito alla paura è significativamente correlato al VIX, che è un misura della volatilità dell'indice S&P 500, e non una semplice misura di rendimento.
Rispetto al paper di Bollen, c'è un passo avanti in termini di utilizzo di indici, non solo di rendimento, ma anche di volatilità. In quanto, secondo la teoria classica, l'andamento dei rendimenti non è prevedibile o perlomeno è molto difficile trovare dei trend chiari, mentre la volatilità presenta degli elementi di trend nel tempo. Per quanto riguarda invece le tecniche di natural language processing, l'indice di sentiment ricavato nello studio risulta troppo soggettivo a causa della scelta arbitraria della lista di parole, l'introduzione di un sistema di raccolta basato su una metodologia linguistica avrebbe sicuramente fornito un risultato più oggettivo. L'ultimo paper presentato in questa sezione è:"Twitter financial community sentiment and its predictive relationship to stock market movement" di Yang, Steve e Mo [7].
Tra i diversi articoli presentati questo è quello più recente, è stato infatti pubblicato nel 2015, rispetto, ad esempio, al paper di Bollen che è del 2011. In questo paper è presente un livello di sofisticazione in più rispetto agli altri, infatti, i Tweet giornalieri non vengono campionati casualmente ma sono selezionati solo i Tweet di 50 account di esperti di finanza e di 7 delle maggiori testate finanziarie, come ad esempio Bloomberg, Reuters e il Financial Times. Questi criteri di selezione fanno in modo che il sentiment ricavato dai tweet non provenga dalla massa di tutti gli utilizzatori, ma da utenti che si occupano professionalmente di finanza e che molto probabilmente hanno un impatto maggiore sul mercato rispetto alla massa indistinta degli utilizzatori di Twitter.
Oltre a una network analysis, su cui non ci soffermeremo, il paper arriva alla conclusione che, dopo aver eseguito una tradizionale sentiment anaylsis su i Tweet raccolti, l'indice di sentiment è significativamente correlato a diversi indici di mercato, tra cui il VIX, il Dow Jones, S&P 500 e il NASDAQ.
I risultati fortemente significativi di quest'ultimo paper ci forniscono lo spunto di non utilizzare un campione casuale di Tweet ma di trovare un qualche criterio per individuare Tweet che hanno una maggiore relazione con il mercato finanziario che ci poniamo di

prevedere con la sentiment analysis.
L'idea di scegliere un subset di tweet in base alla popolarità degli utenti nel mondo finanziario è un grande punto di forza, perché permette di focalizzare il campione solo sul sentiment di coloro che si presume siano i veri operatori del mercato. Si tratta tuttavia di un approccio molto sofisticato che richiede un certo tipo di informazioni e un grande volume di Tweet. Per semplicità, in questo elaborato si è scelto di filtrare i tweet in base al contenuto, che deve citare almeno una delle aziende del paniere del FTSE-MIB e non in base alla "popolarità" dei diversi account Twitter. Di seguito sono riportati brevemente ulteriori ricerche che hanno trattato l'argomento negli ultimi anni:

- "Sentiment analysis of Twitter data for predicting stock market movements" [8]. Che come elemento di novità presenta, al posto dei dizionari, dei modelli di machine learning allenati su Tweet classificati manualmente.

- "The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices" [9]. In cui è presentata l'integrazione della sentiment con altri indici del mood di mercato tramite un Kalman Filter.

- "Predicting Fluctuations in Cryptocurrency Transactions Based on User Comments and Replies" [10]. Dove la variabile di interesse non è un indice finanziario ma il volume di scambio delle cryptocurrency.

Ovviamente questi sono solo alcuni esempi degli articoli esistenti sull'argomento, ma credo siano quelli più calzanti per fornire una breve panoramica di quello che è stato ad oggi ricercato e sviluppato in letteratura. Per ulteriori approfondimenti sulla ricerca nell'ambito dell'analisi di dati alternativi per la finanza si consiglia di far riferimento a: "Special Issue of Quantitative Finance on 'Financial Data Analytics'" di Jessica James, Dietmar Maringer, Vasile Palade, Antoaneta Serguieva [11]. In linea generale ciò che possiamo apprendere da questi articoli è che nei messaggi pubblicati su Twitter c'è sicuramente dell'informazione utile che, adeguatamente trasformata in un determinato indice, può essere correlata sia con l'andamento dei rendimenti di borsa sia con l'indice di volatilità della borsa stessa. Forte delle conclusioni raggiunte dalla letteratura scientifica l'elaborato si prefigge di ricercare un indice di sentiment e un modello che possano descrivere al meglio la volatilità nel mercato azionario italiano.
Per concludere, in questo capitolo sono stati presentati tutti gli strumenti teorici e i punti di riferimento in letteratura che poi verranno utilizzati per analizzare e interpretare i dati raccolti.

Capitolo 2

Sentiment analysis

L'econometria ha sempre cercato di fornire uno strumento per comprendere le relazioni tra determinati fenomeni. Per gli economisti comprendere se il mercato sia razionale o meno nelle sue scelte è sempre stato un campo di ricerca di grande interesse. Da molto tempo infatti si cerca di comprendere tali relazioni tramite, ad esempio , l'utilizzo di questionari in grado di quantificare le opinioni del soggetto intervistato. L'avvento di enormi quantità di dati testuali, audio e visivi ha stimolato lo sviluppo di nuove metodologie per trasformare i dati qualitativi in variabili quantitative, così da poter poi formalizzare e testare statisticamente le relazioni tra le suddette variabili. La sentiment analysis è la metodologia che ci permette di passare da dati testuali (qualitativi) a un indice (quantitativo) così da darci la possibilità di analizzare statisticamente le relazioni di tali informazioni con altre variabili economiche. Perciò è di fondamentale importanza comprendere quali sono le possibili metodologie per sintetizzare concetti espressi in un determinato testo in un valore numerico interpretabile.

2.1 La sentiment analysis basata su dizionari

Grazie alla crescita esponenziale delle informazioni disponibili online, i ricercatori si sono trovati a confrontarsi con l'opportunità e la sfida di trovare nuovi metodi di estrazione di informazioni che riuscissero a trarre vantaggio da questa grande disponibilità di dati. Uno dei recenti strumenti nato per questo scopo è la sentiment analysis, che permette di ricavare la polarità di un testo cercando di stimare l'opinione di chi lo ha scritto.
La sentiment analysis è una metodologia di estrazione dell'informazione dal testo fortemente interdisciplinare. Infatti rientra nel mondo della linguistica computazionale ma le sue implicazioni teoriche ricadono anche sulla psicometria e sulla teoria delle variabili latenti. Inoltre i dati testuali non sono in un formato "tradizionale" e quindi sono necessarie anche metodologie informatiche specifiche per la loro elaborazione. In questo elaborato non cercheremo di giustificare il framework teorico per cui è possibile, tramite

il testo, approssimare l'opinione latente di chi lo ha scritto, ma ci limiteremo ad utilizzare questo potente strumento per verificare empiricamente se può avere un utilizzo rilevante nel campo della previsione economica.
Come per la volatilità la letteratura si divide in svariati approcci alla risoluzione di tale problema: in questa sezione verrà trattata la metodologia basata sull'utilizzo di dizionari, invece nella prossima sezione verrà analizzata la metodologia basata sul supervised learning di reti neurali.
Il dizionario si basa su un concetto intuitivo per cui emozioni diverse si esprimono attraverso parole diverse. Ad esempio, felice e piacevole indicano l'emozione della gioia, mentre oscurità e pianto sono indicativi di tristezza, e così via. Perciò un dizionario manualmente annotato ci fornisce lo strumento per identificare quali parole descrivano determinate emozioni. Un dizionario di sentiment ci indica l'emozione corrispondente a ogni parola e il grado di intensità con cui quella parola esprime l'emozione. Definire le emozioni espresse da un testo basandosi solo sull'elemento atomico parola può presentare delle problematiche, ad esempio, la stessa parola può evocare emozioni diverse in contesti o in frasi diverse. Inoltre non vanno dimenticati i rafforzativi o i prefissi negativi che possono modificare completamente il significato generale di una frase.
Infatti una frase non è semplicemente la "somma" delle sue parti ma rappresenta qualcosa di più profondo per l'interpretazione umana. Tuttavia approcciare il problema tramite le singole parole ci permette di semplificare la metodologia abbastanza da fornirci uno strumento rapido di descrizione del fenomeno, strumento che non è privo di difetti, ma che sicuramente ci può dare delle informazioni importanti sul lessico utilizzato
. Da ultimo va considerato che i Tweet generalmente sono di breve lunghezza o al massimo possono raggiungere i 140 caratteri, perciò tendenzialmente sono scritti con un lessico semplice e chiaro, quindi più facile da scomporre. Per questo elaborato sono stati utilizzati due diversi dizionari per calcolare gli indici di sentiment: NRC e OpeNER. OpeNER [12] è un dizionario sviluppato in modo semi-automatico da ItalWordNet v.2 partendo da una lista di 1.000 parole chiave a cui è stato associato manualmente uno score di sentiment. Contiene 24.293 entrate lessicali annotate con polarità positiva/negativa/neutra. OpeNER è un progetto finanziato dalla Commissione Europea nell'ambito del 7° Programma Quadro. Il nome è l'acronimo di Open Polarity Enhanced Name Entity Recognition. L'obiettivo principale di OpeNER è quello di fornire una serie di strumenti pronti all'uso per eseguire alcuni task di natural language processing gratuitamente e con facilità.
È adatto per il mondo accademico, la ricerca e le piccole e medie imprese per integrare tali metodologie nel flusso di elaborazione dati. Più precisamente, OpeNER mira eseguire la sentiment analysis sul testo, in particolare su testi provenienti da fonti web come recensioni o commenti sui blog. Il difetto di tale dizionario è che la maggior parte è stato annotato in modo semi-automatico quindi fa molto affidamento sulla capacità del corpus sottostante di assegnare la giusta relazione tra le diverse parole.
Per calcolare l'indice di sentiment in prima istanza è stato utilizzato il dizionario Ope-

NER, ma poiché questo fornisce solo la polarità negativa e positiva, è stato utilizzato anche un ulteriore dizionario in grado di distinguere un numero maggiore di emozioni che compongono il linguaggio.
Si tratta del dizionario NRC Sentiment and Emotion Lexicons [13], il cui nome è acronimo di National Research Council Canada. È una raccolta di svariati lessici, tra cui l'ampiamente utilizzato Word-Emotion Association Lexicon. Il dizionario è stato sviluppato tenendo conto di un'ampia gamma di applicazioni e contesti; fornisce una lista di 25.000 entrate lessicali e le loro associazioni con otto emozioni di seguito elencate:

- Rabbia
- Paura
- Attesa
- Fiducia
- Sorpresa
- Tristezza
- Gioia
- Disgusto

Il dizionario è stato sviluppato in inglese ma può essere utilizzato per analizzare testi anche in altre lingue perché fornisce traduzioni automatiche delle voci del dizionario in altre 40 lingue, tra cui quella di nostro interesse: l'italiano. Il dizionario è molto interessante dal punto di vista della robustezza degli score ottenuti perché è stato creato tramite crowdsourcing e quindi utilizzando un campione molto eterogeneo in termini di età, etnia e ambiente sociale. Questa caratteristica lo dovrebbe rendere capace di descrivere i diversi sentiment che derivano da connotazioni provenienti da slang e termini specifici di diversi ambiti. Per svilupparlo è stato utilizzato il servizio Mechanical Turk di Amazon come piattaforma per ottenere annotazioni su una larga scala di individui. Tramite tale servizio i ricercatori che hanno sviluppato il dizionario hanno creato dei task per ogni parola del lessico e il sistema ha assegnato casualmente a diverse persone delle domande che permettessero di identificare quale emozione evocasse quella determinata parola.
I dizionari presentati sono stati scelti perché entrambi rappresentano lo stato dell'arte delle risorse open-source disponibili per questo specifico problema. Esistono altri servizi offerti dai grandi player dell'elaborazione dati, come ad esempio i servizi cognitivi di Microsoft Azure. Ma i servizi offerti commercialmente non offrono della documentazione chiara riguardo a come siano stati generati i diversi lessici e quindi è difficile valutarne

la qualità e le possibilità di utilizzo.
In questa sezione sono stati presentati i due dizionari e le motivazioni per cui sono stati scelti, nella sezione metodologica sarà trattato nel dettaglio il passaggio dallo score fornito dal dizionario per ogni singola parola a un indice di sentiment vero e proprio.

2.2 Le reti neurali per l'analisi del linguaggio naturale

In questa sezione verrà presentato il modello di rete neurale che è stato allenato per riconoscere il sentiment delle parole in automatico, a differenza dei dizionari presentati in precedenza che derivavano da annnotazioni manuali.
Le reti neurali sono algoritmi di machine learning, ispirati dal funzionamento del cervello umano, per risolvere problemi con una forte complessità e non linearità. Le reti neurali possono essere utilizzate per risolvere questo tipo di problemi grazie ad un processo di apprendimento, che può essere supervisionato o meno. La rete neurale è un insieme di unità che hanno semplicemente un input e un output collegati tra loro, in cui ogni connessione ha un bias e un peso associato. Nella fase di apprendimento, la rete cerca di approssimare la funzione che può generare un determinato output dati i diversi input. Ciò avviene regolando i propri pesi in relazione alla funzione di perdita assegnata a tale previsione.
Esistono due diversi approcci per la sentiment analysis automatica che utilizza le reti neurali: supervised ed unsupervised. Quello supervised si basa sulla presenza di una label o di una qualche metrica che descrive il fenomeno e quindi che la rete si deve allenare a prevedere. Invece la metodologia unsupervised si basa sul ricercare dei gruppi di parole che sono più comunemente associate tra loro e che quindi rappresentano concetti simili. Il problema dell'approccio supervised è che spesso avere delle label non è economico oppure non è possibile, invece il difetto del modello unsupervised è che non è facile capire cosa rappresentino effettivamente i gruppi ottenuti e se possono essere di nostro interesse.
In questo elaborato si è utilizzato un approccio supervised anticonvenzionale, infatti come label, al posto della classica positive o negative annotata manualmente, si è utilizzato l'andamento della volatilità dell'indice di borsa. Con l'ipotesi che, dato come input il testo dei Tweet del giorno precedente e come label il valore di borsa del giorno corrente, la rete sia capace di capire quali parole hanno un' "influenza" positiva sull'indice futuro e quali una negativa.
Nel campo dell'analisi del linguaggio naturale, negli anni sono state sviluppate le più disparate architetture di reti neurali per cercare di cogliere la struttura del linguaggio. Tra le tante possibili opzioni a nostra disposizione è stato scelto di utilizzare una struttura di media complessità, cioè i deep neural network. Tale architettura è stata scelta perché

ai fini dell'elaborato non ci interessa ricercare lo strumento migliore per la classificazione del testo ma uno strumento che sia ormai consolidato e affidabile nella risoluzione di questo tipo di task.
I deep neural network sono semplicemente reti neurali con più strati disposti gerarchicamente, da questo deriva l'aggettivo "deep". La ricerca ha dimostrato che un deep neural network sufficientemente complesso può descrivere una qualsiasi funzione matematica e può quindi essere un classificatore universale, come espresso nell'universal approximation theorem [14].
Una volta individuata la necessità di avere un Hidden Layer per soddisfare il teorema di approssimazione generale e quindi assicuratisi che la rete sia in condizione di poter descrivere il fenomeno al meglio, va determinata quale dimensionalità associare ad ogni layer. Tale scelta è stata presa su base empirica, allenando diverse reti e rispettando la best practice di scegliere il modello che a parità di performance presenta il minor livello di complessità e di numero di nodi dell'input layer rispetto al numero di nodi nell'hidden layer [15].I modelli sono stati tutti valutati in base all'errore quadratico medio out of sample e l'architettura di seguito presentata è emersa come la migliore.

Per facilità di rappresentazione, l'input dei dati, che in questo elaborato corrisponde a 1500 valori, non è rappresentato in figura, ma sono rappresentati solo i nodi attivi della rete. Perciò è stata utilizzata una configurazione con due hidden layer, rispettivamente di 10 e 5 neuroni e un output layer di un neurone.
Una volta presentata l'architettura ad alto livello va capito esattamente cosa c'è al suo interno e che funzioni vengono calcolate. Come si può vedere dalla figura, ogni strato è costituito da uno o più nodi, rappresentati in questo diagramma dai piccoli cerchi. Le linee tra i nodi indicano il flusso di informazioni da un nodo all'altro. In questo particolare tipo di rete neurale, le informazioni fluiscono solo dall'ingresso all'uscita (cioè da sinistra a destra). Altri tipi di reti neurali hanno connessioni più complesse, come i percorsi di feedback. I valori che entrano in un nodo vengono moltiplicati per dei pesi, che sono in prima istanza inizializzati casualmente e poi vengono modificati durante l'allenamento. Quindi ogni nodo come prima operazione riceve una combinazione lineare dei diversi input che poi viene sommata per ottenere un determinato output, che è mostrato nell'equazione seguente.

$$\text{Output} = A_1 x_1 + \ldots + A_n x_n \tag{2.1}$$

Dove x_1 rappresenta il primo input che la rete riceve, A_1 il relativo peso e n il numero di input. Prima di lasciare il nodo, questa somma viene fatta passare attraverso una funzione matematica non lineare chiamata funzione di attivazione, in questo caso è utilizzata la funzione sigmoide. Tale funzione è simile a una curva a forma di "s" che limita l'uscita del nodo. L'ingresso al sigmoide è quindi un valore compreso tra $-\infty$ e

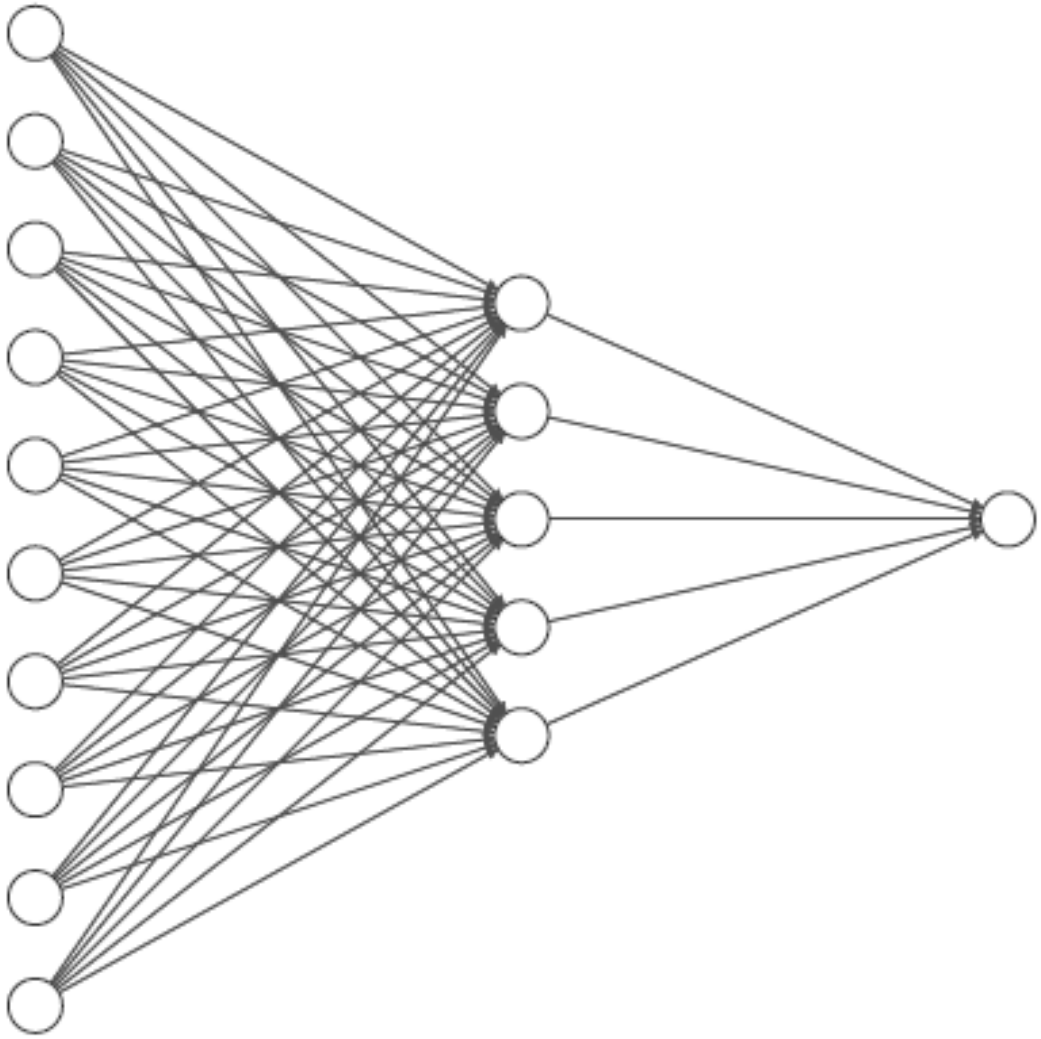

Figura 2.1: Hydden Layer utilizzati nella rete neurale

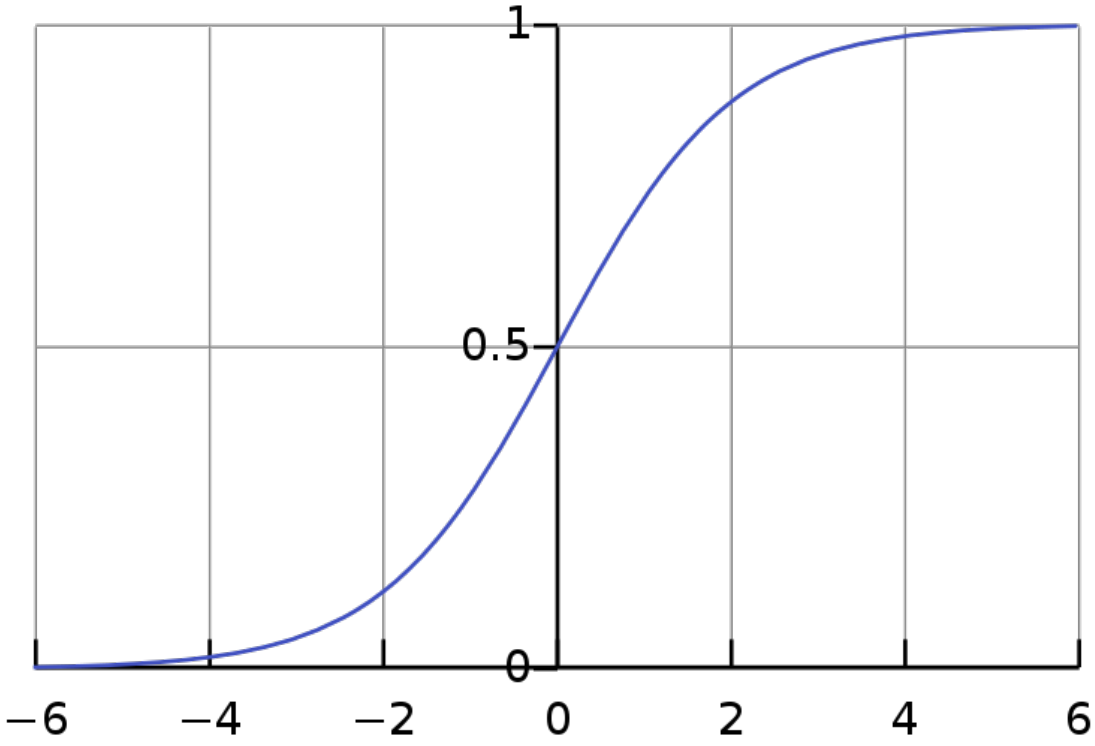

Figura 2.2: Funzione di attivazione Sigmoide

$+\infty$, mentre il suo output può essere solo tra 0 e 1. Di seguito è mostrata e definita la funzione in questione:

$$S(z) = \frac{1}{1 + e^{-z}} \tag{2.2}$$

Questo calcolo viene ripetuto per ogni nodo, che accetta come input tutti gli output dei nodi precedenti e fornisce un output a tutti i nodi successivi. Infine l'ultimo nodo differisce dagli altri solo per la funzione di attivazione, che non è più una sigmoide ma una semplice funzione lineare degli input, così da poter descrivere appunto una regressione, visto che stiamo cercando di addestrare la rete a stimare il valore dell'indice di borsa nel giorno seguente.
Le reti neurali possono avere un numero qualsiasi di strati e un numero qualsiasi di nodi per strato. La maggior parte delle applicazioni usa la struttura a tre strati con un massimo di alcune centinaia di nodi di input.
Una volta definita l'architettura della rete nel dettaglio ed esposte le funzioni che ne fanno parte, va chiarito come tale rete possa stimare i pesi di tutte queste combinazioni lineari

in modo da apprendere informazioni. Di seguito sono esposti i passaggi che costituiscono l'allenamento vero e proprio:

- Inizializzazione casuale dei pesi.
- Calcolo di tutti gli input e di tutti gli output.
- Confronto della stima con il valore reale.
- Backpropagation e aggiornamento dei diversi pesi.
- Calcolo di tutti gli input e di tutti gli output.
- Ripetizione dei due passaggi precedenti.

A questo punto va chiarito come avviene il processo di aggiornamento dei pesi tramite la backpropagation. Prima di tutto è definita una loss function che nel nostro caso è l'errore quadratico medio:

$$\text{rmse} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2} \tag{2.3}$$

Dove:

- y_i = Previsione della rete.
- x_i = Valore reale.
- n = Numero di elementi osservati nel campione.

Una volta ottenuto un determinato valore della loss function viene calcolato il gradiente della loss function rispetto a tutti i parametri della rete. Un gradiente non è nient' altro che la generalizzazione di una derivata su più variabili, quindi è un vettore che contiene le derivate parziali di ciascuna variabile rispetto alla loss function. Una volta determinato il gradiente, i pesi vengono aggiornati in modo da avvicinarsi al minimo della loss function. Per non incorrere in problemi di minimo locale, viene utilizzato un hyperparameter chiamato learning rate che rappresenta una percentuale del gradiente, in modo che ad ogni passaggio di backpropagation l'aggiornamento dei parametri sia graduale e non troppo repentino.
Questo processo viene iterato per un numero di volte scelto dal ricercatore oppure in base a un valore di loss function previsto.
Va evidenziato che un altro motivo fondamentale per cui sono stati utilizzati i feed forward network e non architetture più complesse è il fatto che da questo tipo di architettura possiamo ricavare i pesi che le singole parole hanno all'interno della rete, tramite l'algoritmo di Olden.

L'agoritmo di Olden[16] è simile all'algoritmo di Garson [17], in quanto i pesi di connessione tra gli strati di una rete neurale costituiscono la base per determinare l'importanza delle variabili. Tuttavia, Olden descrive un algoritmo di computo dei pesi che ha costantemente superato l'algoritmo di Garson nel rappresentare la reale importanza delle variabili in un set di dati. In questo algoritmo Olden calcola l'importanza di una singola variabile come il prodotto dei pesi di connessione di ogni combinazione lineare di input/output e poi somma tale prodotto per tutti i neuroni.

$$Imp_i = \sum_{i=1}^{n} (W_i Z_i) \tag{2.4}$$

Dove:

- Imp_i = Importanza della variabile i-esima.
- W_i = Peso assegnato alla connessione tra l'input e l'hidden layer.
- Z_i = Peso assegnato alla connessione l'hidden layer e l'output.

Un vantaggio di questo approccio è che i contributi relativi di ogni peso sono rappresentativi sia in termini di grandezza che di segno, rispetto all'algoritmo di Garson che considera solo la grandezza assoluta. Un ulteriore vantaggio è che l'algoritmo di Olden è in grado di valutare reti neurali con più hidden layer, come nel nostro caso, mentre quello di Garson è stato sviluppato per reti con un singolo hidden layer. I valori effettivi devono essere interpretati solo in base al segno e alla grandezza rispetto alle altre variabili della stessa rete. Non si possono fare confronti tra i risultati di diversi modelli ma solo rispetto la rete stessa.

Oltre all'utilizzo delle reti neurali, esistono ulteriori metodi per l'analisi del sentiment che non fanno affidamento sull'annotazione manuale del testo o sull'utilizzo di questionari. Tali metodi sono chiamati annotazioni semi–automatizzate. Partendo da un campione di parole annotate utilizzano determinate relazioni per stimare il valore delle restanti parole. Tali relazioni vengono ricavate automaticamente tramite l'estrazione della struttura di un corpus linguistico. I corpora sono raccolte di testi in formato digitale specificatamente create per l'analisi linguistica. Grazie ad esse i linguisti riescono ad utilizzare le metodologie statistiche per comprendere i pattern e la struttura che emerge dal linguaggio. Una volta definita una funzione obiettivo come quella della sentiment, sono annotate manualmente le parti di testo più importanti del corpus e le restanti ricavano il loro valore dalla relazione rispetto a quelle annotate estratta dai dati. In questo elaborato si è preferito optare per l'utilizzo di reti neurali rispetto al corpus perchè forniscono uno strumento meno dipendente dalla raccolta dati, infatti i corpora necessitano di una struttura di raccolta dei dati e delle relazioni linguistiche ben precisa. In conclusione, in questa sezione è stata presentata la Deep neural network, con le

relative funzioni di attivazioni, che poi verrà allenata a prevedere il valore dell'indice di volatilità dato come input il testo raccolto dai Tweet per ogni giornata di borsa. Inoltre è stato presentato l'algoritmo di Olden, che ci sarà utile per ricavare i pesi delle singole parole date come input alla rete così da estrarre un dizionario non più basato su raccolta manuale ma su un processo automatico di addestramento.

Capitolo 3

Dati

In questo capitolo tratteremo il procedimento di raccolta, trasformazione, e utilizzo dei dati necessari per l'analisi proposta nell'elaborato. Verranno presentati innanzitutto gli strumenti che hanno permesso l'estrazione di dati relativi ai Tweet, ai dizionari e all'indice di borsa.

3.1 Tweet

I Tweet sono stati scaricati dal software R tramite la funzione SearchTwitter del pacchetto Twitter R sviluppato da Jeff Gentry [18]. SearchTwitter è una funzione che permette al programmatore di scegliere diversi criteri di ricerca, tra cui: geolocalizzazione dell'utente che ha inviato il Tweet, lingua in cui il Tweet è stato scritto, orizzonte temporale, e numero di elementi da campionare. Una volta esplicitati tali parametri il pacchetto richiama l'API ufficiale di Twitter per eseguire l'operazione di download. Un API, acronimo di Application Programming Interface, è un set di protocolli che ci consente di comunicare con un determinato prodotto o servizio tramite l'impostazione di parametri di chiamata, senza conoscere come tale prodotto risponde internamente alla richiesta. Tale API, nella sua versione gratuita, ha delle forti limitazioni. Le limitazioni più significative riguardano l'orizzonte temporale e il limite al campionamento dei tweet richiesti. Nello specifico, è possibile scaricare solamente i dati relativi alla settimana precedente al giorno di download, ciò ci obbliga ad effettuare un diverso download ogni settimana per tutta la durata della ricerca. L'altra grande limitazione riguarda il campione di Tweet ottenuti dall'API, infatti, è possibile chiedere di richiamare al massimo 10.000 Tweet in un singolo download, ma Twitter non fornisce informazioni chiare su come questi Tweet siano stati campionati e quanto siano rappresentativi della popolazione da cui sono stati estratti. In questo elaborato i Tweet sono stati selezionati in base a tre filtri:

- Tweet pubblicati dal lunedì al venerdì.

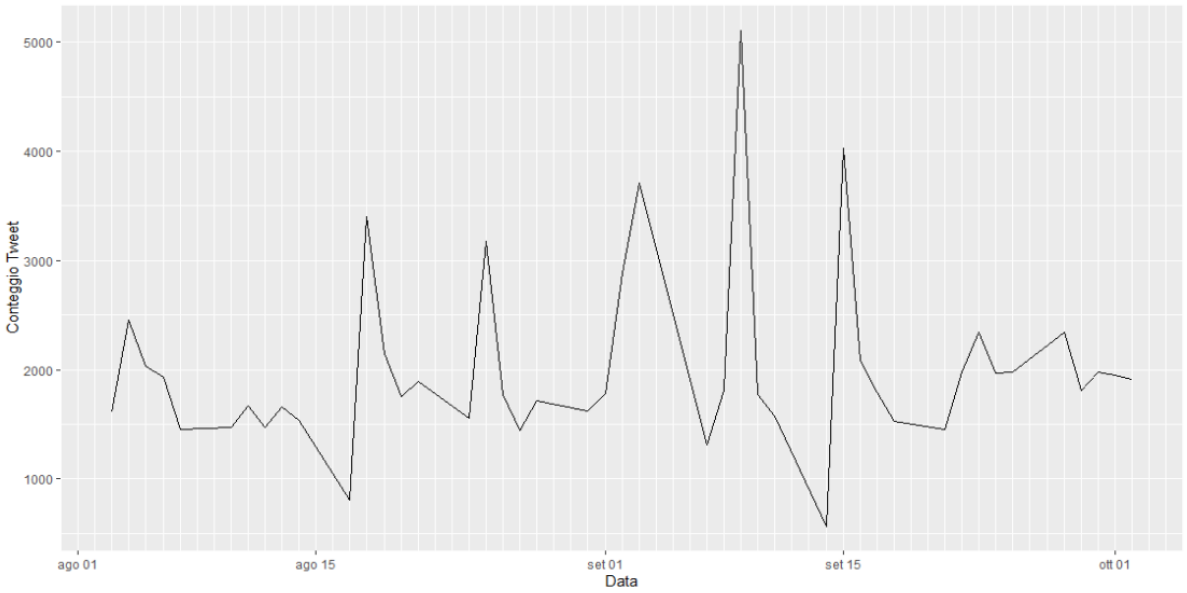

Figura 3.1: Conteggio dei Tweet nel periodo analizzato

- Tweet in lingua italiana.
- Tweet contenenti come parola chiave almeno un'azienda presente nel FTSE-MIB.

Questi filtri sono stati imposti per rispecchiare le caratteristiche dell'indice. Infatti l'indice è principalmente scambiato in Italia da operatori italiani e il mercato è aperto dal lunedì al venerdì.
Il risultato finale dei due mesi di elaborazione dati è il seguente: sono stati raccolti 90.000 Tweet, cioè 10.000 per ogni settimana. I Tweet sono stati estratti in formato Json, una volta scaricati sono stati salvati settimanalmente in formato CSV su una cartella nel Cloud per poi essere letti e concatenati su R per l'analisi finale. Il giorno in cui è stato raccolto il numero minore di Tweet è stato il 14 settembre, in cui ne sono stati estratti 567. Il giorno con il numero maggiore di Tweet è stato invece il 3 settembre, con 3714 Tweet. Di seguito è riportata la serie temporale relativa al conteggio di Tweet raccolti per questo elaborato.

Una volta definito il flusso dati relativo ai Tweet, vanno definiti gli strumenti che hanno permesso la loro manipolazione. Per passare dal file Json al file CSV e successivamente ai dataframe è stato utilizzato il pacchetto per l'analisi dei dati più popolare su R, ossia Dplyr di Hadley Wickham [**?**]Dplyr). Una volta trasformati i Tweet in un dataframe è stato possibile applicare il processo che viene definito tokenization, ogni Tweet è stato cioè scorporato nelle diverse parole che lo compongono.
Successivamente, sono state rimosse tutte le parole che risultano inutili al fine della sentiment analysis, come le stopwords (articoli, congiunzioni e segni di interpunzione), i

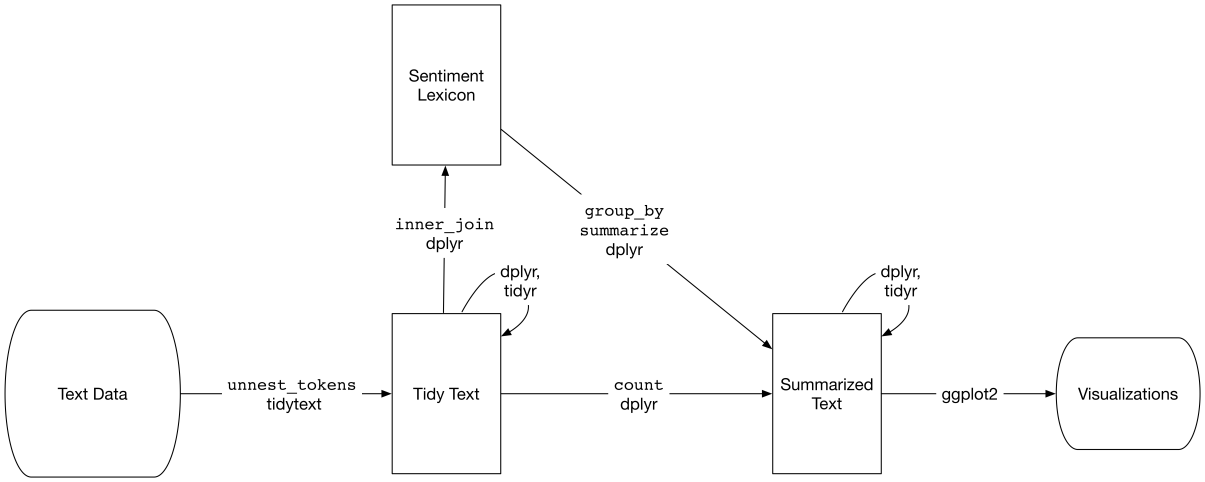

Figura 3.2: Processo di elaborazione dei dati testuali

simboli non alfabetici (chiocciole o hashtag), e le parole chiave utilizzate come filtri per l'estrazione dei Tweet. Una volta "puliti" i Tweet, è stato possibile caricare i dizionari nell'ambiente di R e associare a ogni parola il sentiment e lo score presenti nel dizionario tramite una semplice operazione di left join.
Questa operazione è stata eseguita seguendo la metodologia "Tidy", ossia un approccio all'analisi dei dati su R che fa grande utilizzo dei dataframe e tende ad evitare pratiche come i loop e il functional programming. Ciò è possibile grazie all'utilizzo del sistema di pacchetti e librerie che fa parte del "Tidyverse". Nello specifico, per l'elaborazione dei Tweet è stato utilizzata la metodologia di pulizia e tokenization presentata nel libro " Text Mining with R: A Tidy Approach" di J. Silge e D. Robinson [19].
Di seguito è fornito il diagramma, proveniente dal libro appena citato, che mostra chiaramente quale è il processo di ingestion dei dati testuali in R per la sentiment analysis:

Il risultato finale dell'elaborazione dati dei Tweet è un dataframe ripulito dalle diverse problematiche che si possono trovare nel testo e arricchito dalle informazioni sulla sentiment dei dizionari Opener e NRC già presentati in precedenza.

3.2 Dizionari

Nella sezione precedente abbiamo mostrato come avviene l'ingestion dei Tweet che successivamente vengono uniti tramite left join sui dizionari, ora è necessario soffermarci un momento su come i dati relativi ai dizionari siano stati raccolti.
Per quanto riguarda il dizionario Opener [12], è stato possibile scaricare il file direttamente dal repository open source inizializzato su Github da parte dell'Istituto di Linguistica

Computazionale “A. Zampolli”. Successivamente, il file è stato convertito dal formato LMF al dataframe convenzionale di R. A quel punto è stato possibile utilizzarlo per la left join sui Tweet.
Il dizionario NRC [13], invece, è stato scaricato tramite il pacchetto Syuzhet di Matthew Jockers [20] che fa da "wrapper" su R e ne permette il download in formato CSV selezionando già la lingua di interesse. Una volta scaricato il dizionario con l'ausilio del pacchetto è stato convertito in dataframe e poi utilizzato per la sentiment.

3.3 FTSE-MIB

Fino ad ora ci siamo occupati dei dati relativi a Tweet e dizionari, di seguito invece è presentata la procedure che ha permesso l'estrazione dei dati relativi al FTSE-MIB.
Il FTSE-MIB, acronimo di Financial Times Stock Exchange Milano Indice di Borsa, è il più importante indice azionario della Borsa italiana. Rappresenta un paniere delle quaranta più importanti aziende italiane quotate, valutate in termini di capitalizzazione, liquidità e flottante. L'indice di riferimento italiano è stato inizializzato il 31 dicembre 1992, con assegnato il valore di 100 punti base.
I dati relativi all'andamento di tale indice sono stati scaricati su R tramite la funzione tqget del pacchetto Tidyquant, di Matt Dancho e Davis Vaughan [21]. Tale funzione permette di scaricare svariate serie storiche finanziare semplicemente elencando l'orizzonte temporale di interesse e il ticker del titolo. Il ticker è il codice identificativo univoco associato a ogni prodotto scambiato nei mercati finanziari. Oltre al semplice prezzo di chiusura sono stati scaricati anche il minimo, il massimo e il prezzo di apertura giornalieri, in quanto saranno necessari al calcolo della volatilità realizzata. Questo pacchetto si è rivelato di grande utilità in quanto la sua API interna fornisce i dati già nel formato dei dataframe di R, quindi non sono state necessarie ulteriori operazioni di data wrangling. Come per i Tweet, l'indice è stato scaricato con un orizzonte temporale che va dal 3 agosto al 3 ottobre, di seguito è riportato il suo andamento:

Dal grafico possiamo notare il forte andamento negativo dell'indice; ciò è probabilmente dovuto alle forti problematiche socioeconomiche causate dal Coronavirus. In periodi non di crisi sarebbe raro trovare un indice di mercato con perdite così marcate poiché in un paniere le possibili forti perdite in cui può incorrere una azienda sono "attutite" dai restanti elementi dell'indice. In questo caso, invece, la forte discesa ha comportato perdite di valore generalizzate a tutte le aziende presenti nel paniere. Riassumendo, in questo capitolo sono stati mostrati i dati raccolti e gli strumenti che hanno permesso tale estrazione. Nel prossimo capitolo tratteremo l'utilizzo dei dati per l'analisi vera e propria.

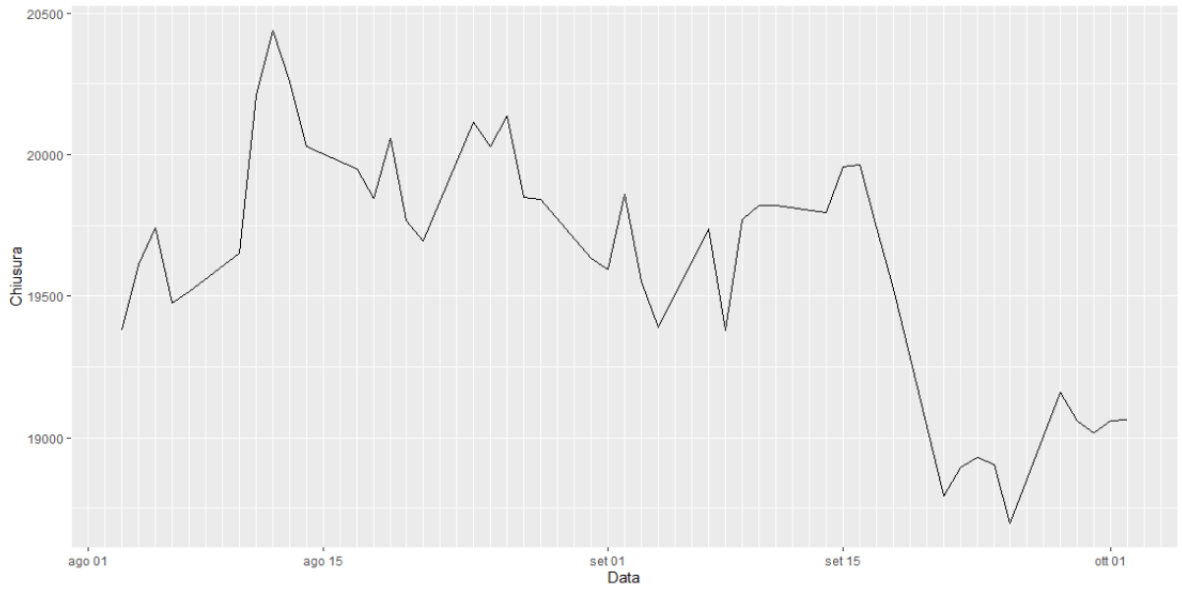

Figura 3.3: Andamento dell'indice FTSE-MIB

Capitolo 4

Metodologia statistica

In questo capitolo sono presentati gli strumenti statistici che ci permettono di studiare i fenomeni relativi a eventi che si realizzano nel tempo. In primis verrà definito l'approccio di analisi e la definizione di serie storiche, successivamente verranno presentati i modelli basati sullo studio delle componenti a medie mobili e autoregressive. Infine verranno giustificate le scelte metodologiche e il workflow seguito nell'analisi.

4.1 Modelli per le serie storiche

Dal punto di vista intuitivo, una serie storica è definita come un insieme di informazioni relative a un dato fenomeno ordinate nel tempo. Indicando con Y il fenomeno e con t l'indice temporale, che va da 1 a T, dove T è il numero complessivo di osservazioni, una serie storica è cosi definita : $Y_t = \{Y_1, Y_2 \ldots, Y_T\}$.
Esempi di serie storiche sono gli indici finanziari, la temperatura misurata nel tempo o le entrate di un determinato conto corrente. Da un punto di vista formale invece le serie storiche sono definite come la realizzazione finita di un processo stocastico, una legge probabilistica che descrive come un certo fenomeno si è evoluto e si potrà evolvere nel tempo. Un processo stocastico è definito come un insieme ordinato di variabili aleatorie, cioè una famiglia di variabili casuali indicizzate da un parametro t, che nelle serie storiche rappresenta il tempo, indicato come: $\{y_t\}_{t \in T}$.

Una variabile aleatoria X è una funzione che associa ad ogni elemento dello spazio campionario Ω un numero reale :$X : \Omega \to R$. Nel nostro Ω è discreto e quindi lo sarà anche la variabile aleatoria X.

Perciò con questo approccio la serie storica è una sequenza di valori campionari relativi a variabili casuali differenti generate da un determinato processo stocastico. Nella pratica della modellazione delle serie finanziarie è necessario presupporre alcune ipotesi per poter stimare i parametri del processo stocastico che genera la serie storica:

- Il momento primo e secondo devono essere sufficienti a descrivere la distribuzione della variabile aleatoria.
- Il processo deve avere una memoria finita.
- Il processo deve essere stazionario nel tempo.

Tali ipotesi saranno testate sui dati realmente raccolti nella parte relativa ai risultati in cui verrà studiata la forma distributiva, la struttura di autocorrelazione e la sua stazionarietà. In questa sezione non tratteremo di tutti gli strumenti per modellare tali fenomeni ma ci si soffermerà solo su quelli che sono stati utilizzati nell'elaborato per modellare la volatilità e l'indice di sentiment.
Il primo strumento di analisi delle serie storiche utilizzato nell'elaborato è una tecnica di previsione chiamata Exponential smoothing. L'Exponential smoothing è una tecnica di smoothing sviluppata da Holt [22] per prevedere i valori futuri delle serie temporali. Nel dettaglio, mentre nella media mobile semplice le osservazioni passate vengono ponderate tutte allo stesso modo, nelle Exponential smoothing vengono utilizzate delle funzioni esponenziali per assegnare dei pesi decrescenti nel tempo. Di seguito è riportata la formula base di tale metodologia:

$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1-\alpha)y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + \cdots , \tag{7.1}$$

Dove $0 \leq \alpha \leq 1$ è il parametro di smoothing, all'avvicinarsi di α a uno si dà maggior peso alle ultime osservazioni, viceversa all'avvicinarsi di α a 0 le osservazioni passate acquisiscono un peso maggiore. Tale tecnica può essere utilizzata non solo a fini predittivi, ma, come nel nostro caso, anche per ridurre il rumore della serie pur mantenendo il contenuto informativo, ossia il segnale.
Infatti le medie mobili sono strumenti semplici e molto utili per ridurre le oscillazioni di serie con molto rumore ed estrarre il trend fondamentale della serie. Per questo motivo vengono spesso utilizzate come filtri nell'analisi dei segnali. La funzione esponenziale contribuisce dando un peso maggiore alle ultime osservazioni, così da rendere la serie "smussata" più reattiva agli ultimi cambiamenti ma allo stesso tempo non troppo soggetta alla casualità riscontrata nella serie.
Un altro strumento che verrà utilizzato nell'elaborato è il test di Granger. Il test di causalità di Granger è un test di ipotesi per determinare se i lag di una determinata serie temporale X sono utili nella previsione di un'altra serie Y [23].
Va evidenziato che la causalità dimostrata dal test di Granger non può essere in alcun modo intesa come la vera causalità, perché tale concetto è filosoficamente e scientificamente molto più forte. Quello che può essere dimostrato tramite il test di Granger è chiamata "Granger Causality" ed è una mera misura della capacità di prevedere i valori futuri di una serie dati i valori passati di un determinato regressore. In pratica, il test di Granger ci garantisce che il regressore prevede la variabile dipendente ma non che la

causa direttamente.
Una serie temporale X è detta Granger causality di Y se si può dimostrare ,tramite test statistici, che quei valori X forniscono informazioni statisticamente significative sulle realizzazioni future di Y. Nonostante le sue debolezze teoriche, tale test è un metodo popolare per l'analisi della causalità nelle serie temporali, grazie alla sua semplicità di calcolo.
Tecnicamente il test si basa su una regressione OLS:

$$\hat{y}_i = \alpha_0 + \sum_{j=1}^{n} \alpha_j y_{i-j} + \sum_{j=1}^{n} \beta_j x_{i-j} + \varepsilon_i \tag{4.1}$$

Dove α_j sono i coefficienti relativi ai ritardi della variabile Y e β_j sono i coefficienti relativi ai ritardi del regressore X. Il test si basa sull'ipotesi nulla H_0:

$$H_0 = \beta_1 = \beta_2 = \cdots = \beta_n = 0 \tag{4.2}$$

Quando tale ipotesi è rifiutata possiamo affermare che X è Granger casuality di Y. Per verificare l'ipotesi viene utilizzato il test F:

$$F = \frac{(SSE_r - SSE)/n}{SSE/(n-p)} \tag{4.3}$$

Dove:

- $SSE = \sum(y_i - \hat{y}_i)^2$,rappresenta i residui del modello completo.
- $SSE_r = \sum(y_i - \hat{y}_i)^2$,rappresenta i residui del modello ridotto.
- $q =$ Numero dei parametri del modello completo.
- $p =$ Numero dei parametri del modello ridotto.

Una volta dimostrato tramite il test di Granger che una certa variabile X è in grado di prevedere una variabile Y è necessario formalizzare in quali modelli inserire tale regressore per migliorare la capacità previsiva su Y. I modelli più utilizzati per le analisi delle serie storiche sono i modelli ARMA, perché riescono a descrivere molte delle caratteristiche fondamentali di tipologia di serie. In questo elaborato utilizzeremo i modelli ARMA come benchmark per poi estenderli inserendo regressori aggiuntivi relativi alla sentiment così da poter verificare se le performance predittive siano migliori.

ARMA è acronimo di autoregressive–moving-average ed è un modello che è stato reso popolare dalla pubblicazione del 1970 di Box e Jenkins [24].Come suggerisce l'acronimo è un modello composto da due elementi, il primo, AR rappresenta la regressione della

variabile sui propri valori ritardati (cioè passati),il secondo, MA rappresenta la modellazione del termine di errore come una combinazione lineare dei termini di errore che si sono verificati nei vari istanti passati. Il modello è solitamente indicato come ARMA(p,q) dove p è l'ordine della parte autoregressiva e q è l'ordine della parte a media mobile.
L'idea del processo AR è che il valore della serie al tempo t sia una funzione lineare dei propri ritardi più un white noise. Procediamo a formalizzare la formula dell'elemento autoregressivo:

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t \tag{4.4}$$

Dove ϕ_i sono i diversi parametri di regressioni dei vari ritardi, c è una costante e ε_t è il termine di errore che si distribuisce come un white noise, per la precisione : $\varepsilon_t \sim WN(o, \sigma^2)$. Il modulo di ϕ_i deve essere minore di uno, altrimenti la serie risulterebbe non stazionaria, e nei modelli ARMA non è possibile trattare quel tipo di serie. Nel caso in cui la serie non risultasse stazionaria una soluzione comune consiste nel modellare non più la serie stessa ma la serie delle differenze prime, formalizzando un modello integrato. Una casistica di particolare interesse si verifica quando il parametro ϕ_1 non rispetta le condizioni di stazionarietà ed è uguale ad uno. In questo caso il processo AR(1) risulta:

$$y_t = c + y_{t-1} + \varepsilon_t \tag{4.5}$$

Questo è il famoso processo chiamato Random Walk, che nella maggior parte dei casi corrisponde al processo seguito dalle serie dei rendimenti finanziari. La stima dei parametri del processo AR(p) avviene attraverso la metodologia OLS, che ha due importanti proprietà: la consistenza e la normalità asintotica.

Un processo è un processo a media mobile di ordine q se può essere espresso come una combinazione lineare finita di q variabili aleatorie che costituiscono un processo white noise. Di seguito ne è riportata la struttura :

$$y_t = \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t \tag{4.6}$$

Dove θ_i sono i diversi parametri di regressione dei vari ritardi del termine di errore e ε_t è il termine di errore che si distribuisce come un white noise. Il processo a media mobile MA(q) ha una memoria finita, nel senso che le osservazioni oltre l'ordine q non sono correlate. Invece il modello AR(p) decade in modo esponenziale, ma la sua memoria non si interrompe mai.
Unendo in uno stesso modello la parte autoregressiva e la parte a media mobile si ottiene il modello ARMA:

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t \tag{4.7}$$

Il processo esiste ed è debolmente stazionario se e solo se tutte le radici del polinomio $P(z)1 - \phi_1 z + \ldots + \phi_p z^p$ sono definite al di fuori del cerchio di raggio unitario. Inoltre il processo è invertibile se tutte le radici del polinomio $Q(z)1 - \theta_1 z + \ldots + \theta_p z^q$ si trovano al di fuori del cerchio di raggio unitario. Dopo aver determinato le condizioni per cui è possibile definire i parametri, la stima vera e propria avviene tramite lo stimatore di massima verosimiglianza.
Una volta che abbiamo definito il modello ARMA(p,q), va studiato come tale modello può produrre una previsione ai tempi successivi a t. Prendiamo come esempio un caso semplice in cui il miglior modello per il fenomeno analizzato risulta essere un ARMA(1,1).

$$y_t = c + \phi_1 y_{t-i} + \theta_1 \varepsilon_{t-i} + \varepsilon_t \tag{4.8}$$

A questo punto dopo la stima dei parametri saremmo interessati a capire quale sia il valore al tempo y_{t+1}. Per effettuare una qualsiasi previsione va studiato il set informativo disponibile fino al tempo t I_t. Per ottenere una stima puntuale abbiamo bisogno di definire una funzione di costo per giudicare le diverse alternative, in questo caso, come di consuetudine, utilizziamo l'errore quadratico medio come tale funzione. Data come funzione di costo l'errore quadratico medio la stima ottimale risulta essere il valore atteso condizionato $\hat{y}_{t+1} = E[y_{t+1}|I_t]$. Il calcolo di tale valore atteso dipende dalla distribuzione di ε_t, ma visto che ε_t si distribuisce come un White Noise a media nulla e varianza finita $E[\varepsilon_{t+1}|I_t] = 0$, ci semplifica di molto la procedura di stima. Perciò nel caso dell'ARMA (1,1) la stima al tempo $\hat{y}_{t+1}$ risulterà:

$$\hat{y}_{t+1} = c + \phi_1 y_t + \theta_1 \varepsilon_t \tag{4.9}$$

Dopo aver fornito uno stima puntuale è necessario definire il grado di variabilità di tale stima tramite un intervallo di confidenza. Tale intervallo è banale da costruire, perché dipende dalla varianza σ della variabili di errore ε_t. Quindi l'intervallo risulterà:

$$\hat{y}_{t+1} \pm z_{\alpha/2}\sigma \tag{4.10}$$

Nel caso in cui la previsione richiesta fosse a tempi successivi al primo, si applicherebbe ricorsivamente la formula 4.9, con l'unica differenza che l'intervallo di confidenza aumenterebbe di ampiezza ad ogni tempo successivo al primo in base alla seguente regola:

$$\hat{y}_{t+i} \pm z_{\alpha/2}\sigma \sqrt{\sum_{i=o}^{t-1} \phi_i} \tag{4.11}$$

Dopo aver formalizzato la definizione di un modello ARMA(p,q), averne elencato le restrizioni e i metodi di stima, rimane da chiedersi come selezionare tra tutti i possibili p,q quelli adeguati al fenomeno in analisi.

Un metodo rapido per determinare quali p,q utilizzare è quello di produrre una visualizzazione dell'ACF e della PACF della serie storica. Il grafico di ACF (Autocorrelation Function) è la rappresentazione della funzione di autocorrelazione, quindi ci fornisce i valori di autocorrelazione di una serie relativi ai suoi elementi ritardati. La sua formulazione risulta quindi banale:

$$ACF_k = \frac{cov(y_t, y_{t-k})}{v(y_t)} \tag{4.12}$$

Dove k rappresenta il k-esimo ritardo della serie, *cov* la covarianza e v la varianza. Il grafico di PACF (Partial Autocorrelation Function) è la rappresentazione della funzione di autocorrelazione parziale, fondamentalmente, invece di mostrare le correlazioni con i diversi ritardi come l'ACF, mostra la correlazione depurata dai ritardi intermedi. Tale correlazione è quella che può essere definita "pura", nel senso che calcolando la semplice autocorrelazione tra y_t e y_{t-2} si avrà anche l'effetto di y_{t-1}. Invece nel caso dell'autocorrelazione parziale, si elimina la parte di correlazione spiegata da y_{t-1} così da ricavare la correlazione che intercorre solo tra y_t e y_{t-2}. Ad esempio la PACF in questione viene calcolata come segue:

$$y_t = \phi_{21} y_{t-1} + \phi_{22} y_{t-2} + e_t \tag{4.13}$$

La stima di ϕ_{22} ci darà il valore di autocorrelazione parziale di ordine due. Estendendo questo tipo di regressione a k ritardi avremo la stima dell'autocorrelazione parziale di ordine k. Perciò una volta definiti questi due strumenti al ricercatore non rimane che scegliere p in base a quante correlazioni risultano significative nell grafico dell'ACF e q in base a quante correlazioni parziali risultano significative nel PACF.
In linea di massima un buon approccio è partire con il modello che ci suggerisce l'ACF-PACF e poi raffinare la scelta in base alle performance previsive out of sample del modello oppure in base a dei criteri informativi. In questo elaborato come criterio informativo di riferimento viene presentato il BIC, acronimo di Bayesian Information Criterion.

$$BIC = -2\log(L) + k\log(n) \tag{4.14}$$

Dove k è il numero di parametri stimato all'interno del modello, L è il valore della log-verosimiglianza e n il numero di osservazioni.
All'aumentare della complessità del modello il valore del BIC aumenta e all'aumentare della verosimiglianza il valore del BIC diminuisce. Quindi per selezionare il modello migliore è necessario scegliere quello con il valore di BIC minore.
L'ultimo strumento di analisi delle serie storiche che ci servirà presentare per l'elaborato sono i modelli ARMAX, che non sono nient'altro che i precedentemente presentati ARMA con l'aggiunta di un regressore X.
Un modello ARMA può essere considerato un tipo particolare di modello di regressione in cui la variabile dipendente è in funzione delle variabili indipendenti, che sono tutte ritardi della variabile dipendente e/o ritardi degli errori. Quindi in linea di principio è

semplice estendere un modello ARMA in modo tale da incorporare le informazioni fornite da altre variabili esogene: è sufficiente aggiungere uno o più regressori all'equazione di previsione. Perciò la formulazione del modello non differirà da quella del modello ARMA se non per una parte aggiuntiva riguardante X:

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \sum_{i=1}^{p} \gamma_i X_{t-i} + \varepsilon_t \tag{4.15}$$

Nel modello ARMAX valgono tutte le stesse restrizioni ai parametri dei modelli precedentemente presentati, inoltre le stesse restrizioni di stazionarietà applicate alla parte AR(p) sono imposte ai parametri relativi al regressore esogeno. La stima dei parametri avviene tramite la stima di massima verosimiglianza, raggiunta in maniera iterativa, visto che la complessità del modello non permette soluzioni chiuse. Come per l'ARMA valgono le stesse considerazioni sulla scelta del modello più accurato e sulla scelta dei ritardi p e q esplicitate in precedenza.
Riassumendo, in questa sezione sono stati presentati gli strumenti per l'analisi di serie storiche che verranno utilizzati nella parte di analisi vera e propria. Ripercorrendoli brevemente, il primo a essere stato presentato è l'Exponential smoothing che è un valido strumento per rimuovere del rumore ma mantenere il segnale nelle serie storiche. Il secondo è stato il test di Granger che ci permetterà di dimostrare la capacità previsiva di una serie storica X rispetto ad un altra Y. Infine sono stati presentati i modelli ARMA e ARMAX che sono i modelli che verranno utilizzati per effettuare le previsioni sulla volatilità con e senza l'aiuto degli indici di sentiment.

4.2 Scelte metodologiche

In questa sezione affronteremo il processo decisionale che sta alla base dello studio esposto, facendo riferimento agli strumenti teorici e alla letteratura scientifica presentati nei capitoli precedenti.
Per prima cosa va definito il motivo per cui è stato scelto questo orizzonte temporale. Tale scelta è stata obbligata a causa delle limitazioni dell'API di Twitter,che permette il download solo dei dati della settimana corrente, pertanto il periodo scelto corrisponde alle settimane in cui è stato scritto questo elaborato. Per quanto riguarda la scelta della lingua, sono stati scelti i Tweet in lingua Italiana perché, come riportato da UniImpresa su elaborazione dei dati ufficiali di Banca D'Italia [25], "meno della metà delle quote di società presenti sui listini finanziari è detenuta da soggetti stranieri". Perciò la maggior parte degli operatori sul mercato italiano risulta non straniera ed è quindi meglio rappresentata dalla lingua italiana.
Inoltre, è stato applicato un ulteriore filtro per mantenere i soli Tweet riguardanti aziende

che compongono il FTSE-MIB (nell'appendice A è disponibile la lista completa estratta da Borsa Italiana [26]). Questi filtri nascono dall'intuizione presente già nel paper di Yang [7] e nelle conclusioni del paper di Bollen: " As Twitter.com's user base becomes increasingly international and the use of smartphones equipped with geo-location increases, future analysis will have to factor in location and language to avoid geographical and cultural sampling errors." [5]. Per cui limitare la massa dei Tweet in base a criteri che identificano opinioni più vicine agli operatori che effettivamente partecipano al mercato può portare a risultati più calzanti rispetto a una banale estrazione casuale dalla massa indistinta degli utenti.
Una volta giustificate le scelte riguardanti i filtri dell'estrazione dei Tweet va giustificata la scelta di non indagare la correlazione tra l'indice di sentiment e il rendimento del FTSE-MIB. Innanzitutto va riportato che, secondo la teoria dei mercati efficienti, le previsioni sui rendimenti delle serie finanziari non sono possibili, in quanto ogni operatore possiede tutte le informazioni disponibili al pubblico. Tale teoria è stata esposta da Fama nel 1970 [27].
Inoltre anche l'esperienza empirica dell'econometria riconosce con certezza che la serie dei rendimenti sia nella maggior parte dei casi descrivibile tramite un processo stocastico simile ad una Random Walk con cluster di volatilità. Perciò, vista la dubbia prevedibilità dei rendimenti finanziari e la loro forte natura erratica, che potrebbe portare a incorrere in correlazioni spurie, è più utile e opportuno concentrarsi sulla volatilità di tali serie.
Quest'ultima presenta infatti delle strutture di autocorrelazione e una minore erraticità, caratteristiche che ci fanno premettere la possibilità di creare modelli più affidabili rispetto a quelli basati sui rendimenti. Ora che abbiamo chiarito la scelta della nostra variabile dipendente (volatilità) è necessario spiegare la metodologia che ci permette di ottenere i diversi indici di sentiment, ossia il passaggio da una lista di parole a un indice che fornisce uno score giornaliero.
Per quanto riguarda lo score binario generato con il dizionario Opener, si è proceduto nel seguente modo: dopo aver pulito i dati, sono state filtrate solo le parole presenti nel dizionario. Successivamente, ad ogni parola è stato associato un valore di score in base alla loro connotazione (positiva o negativa). In seguito, abbiamo ottenuto lo score relativo a ogni giorno calcolando la somma degli score delle singole parole divisa per il numero di parole presenti nei Tweet estratti in quel giorno, come riportato nella formula seguente:

$$I_t = \sum_{i=1}^{p} S_i/n \tag{4.16}$$

Dove I_t indica l'indice di sentiment al giorno t,S_i il valore di score della singola parola (valore negativo nel caso delle parole negativamente connotate), n il numero parole complessive estratte in quel giorno e P il numero di parole presenti sia nel dizionario che nei Tweet. L'indice è stato definito dividendo per il numero di parole di quella giornata

perché l'API di Twitter non ci fornisce un campione costante all'interno della settimana quindi è necessario normalizzare per la diversa numerosità campionaria dei singoli giorni. Per quanto riguarda gli indici di sentiment provenienti dal dizionario NRC, l'unica differenza rispetto al calcolo qui presentato è che non vengono assegnati valori negativi, ma si sommano tutti i valori relativi a quel determinato sentiment.
Dopo aver calcolato tutti i diversi indici di sentiment per ogni giorno di analisi ci si è accorti che erano fortemente perturbati, perciò si è deciso di procedere a "smussarli" tramite l'exponentially weighted moving average, così da mantenere il segnale ma rimuovere parte del rumore. La media utilizzata per lo smoothing è a 5 giorni (che corrisponde alla settimana di apertura dei mercati) e con una ratio di 0.4.
Un ulteriore indice di sentiment è stato ricavato tramite l'ausilio della rete neurale. È stata allenata una rete neurale per prevedere l'andamento della volatilità date le informazioni nei Tweet del giorno precedente. Dopo aver addestrato la rete tramite l'agoritmo di Olden, è stato possibile ricavare i pesi assegnati a ciascuna parola fornita in input alla rete. Visto che l'algoritmo di Olden fornisce una scala di pesi che ha significato solo all'interno di quella singola rete, i valori sono stati normalizzati così da poter essere trasformati nella stessa scala dei pesi ottenuta dai dizionari tradizionali. Tali pesi sono stati trattati esattamente come i pesi del dizionario Opener, è stato quindi sviluppato un nuovo indice che fa riferimento ai pesi generati automaticamente dalla rete.
Dopo aver ottenuto i diversi indici di sentiment si è proceduto a confrontarli con l'indice di volatilità e a provarne con il Granger test la significatività a fini predittivi fino al quinto ritardo, che rappresenta il range temporale della settimana lavorativa sui mercati. Successivamente, individuati gli indici e i ritardi significativi, sono stati sviluppati dei modelli ARMAX che integrassero le informazioni provenienti dagli indici di sentiment. Infine sono state confrontate le performance out of sample, i criteri informativi e i residui dei modelli ARMAX contro un semplice modello ARMA, cercando di dimostrare che le informazioni contenute negli indici di sentiment possono diventare uno strumento utile alla previsione tramite modelli econometrici. Di seguito è riportato uno schema riassuntivo del processo logico di questo elaborato:

Riassumendo brevemente il processo: prima di tutto vengono estratti I tweet e i dati relativi al FTSE-MIB, poi viene allenata la rete neurale a prevedere la volatilità e vengono estratti e normalizzati i pesi relativi per generare un un nuovo dizionario di sentiment. Successivamente, tale dizionario,Opener e NRC vengono utilizzati per generare gli indici di sentiment. A questo punto gli indici di sentiment vengono valutati tramite il Granger test. Infine gli indici significativi in termini di capacità previsiva sulla volatilità vengono inseriti nel modello ARMAX, che poi viene confrontato con il bechmark, rappresentato dal modello ARMA.

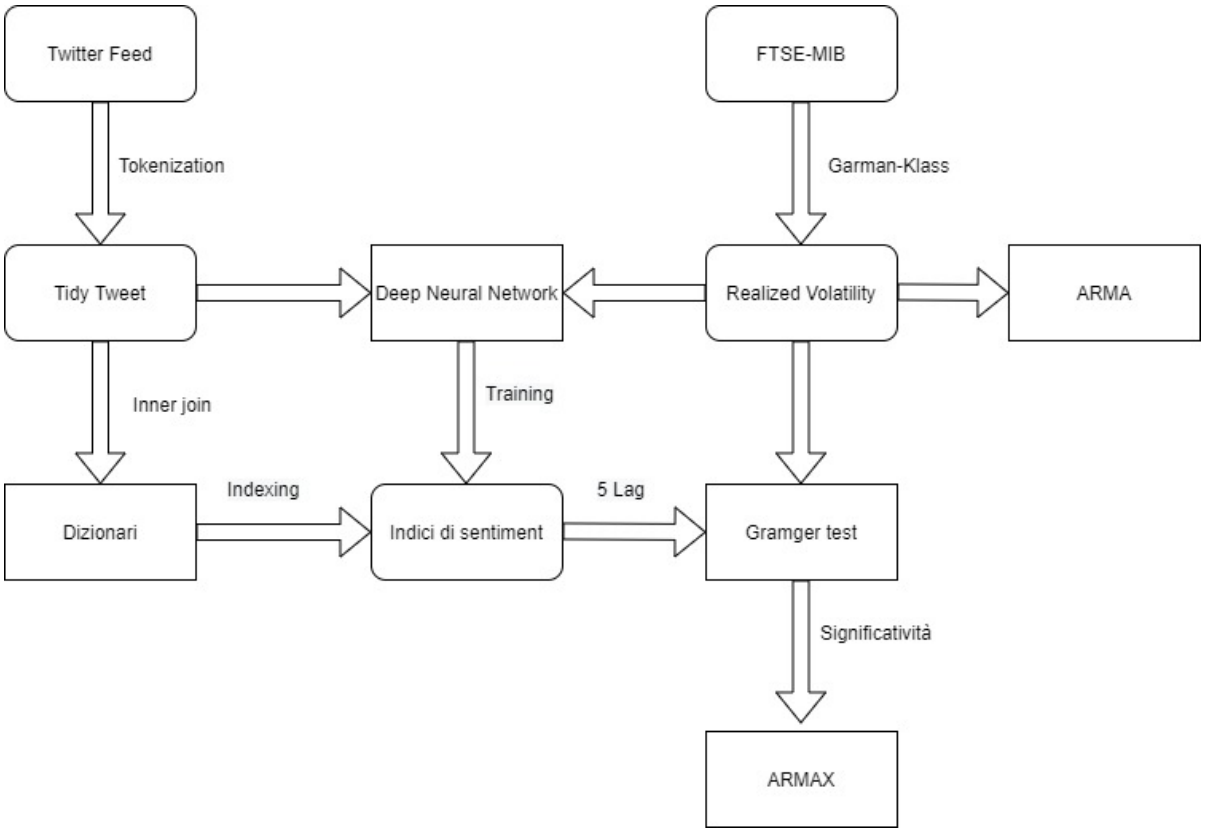

Figura 4.1: Flowchart

Capitolo 5

Risultati

Nel capitolo precedente è stato definito il processo che verrà seguito nell'elaborato e le motivazioni che hanno portato a fare determinate scelte metodologiche. In questo capitolo sono presentati e commentati i risultati ottenuti con i dati raccolti e gli strumenti presentati.

5.1 Volatilità

Prima di tutto viene presentato l'andamento della volatilità realizzata settimanale del FTSE-MIB, calcolata tramite la formula di Garman-Klass, con orizzonte temporale che va dal 3 agosto al 3 ottobre 2020.

Possiamo subito notare che sulla sinistra sono presenti dei dati mancanti, ciò è dovuto al fatto che la volatilità è calcolata su base settimanale, quindi è necessario avere almeno cinque osservazioni precedenti a quell'istante temporale per poterla calcolare. Dalla visualizzazione notiamo che è presente una certa ciclicità, il che rafforza l'ipotesi della presenza di cluster di volatilità. Ciò significa che sicuramente almeno un ordine autoregressivo sarà significativo nello studio del processo che genera questa serie. A prima vista non sembra esserci un forte trend direzionale, anzi, la volatilità oscilla attorno a una media. Un'altra caratteristica fondamentale che si evince da questo grafico è che la volatilità è sicuramente un processo stazionario quindi non dovremo integrarla o fare trasformazioni ulteriori per modellarla tramite gli ARMA.

5.2 Indici di Sentiment

Per quanto riguarda gli indici, come è stato spiegato nel capitolo precedente, una volta che i dati relativi ai Tweet sono stati puliti e arricchiti dai dizionari è stato calcolato l'indice

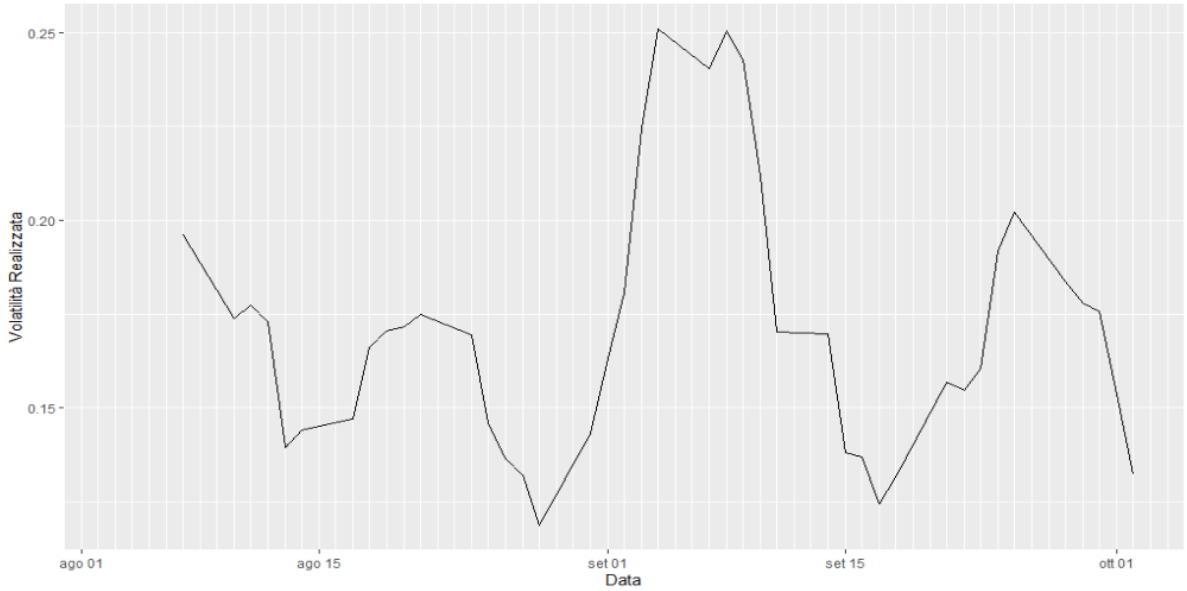

Figura 5.1: Andamento Volatilità

presentato nella formula 4.16 per ognuno dei sette sentiment derivanti dal dizionario NRC, per l'indice binario Opener e per l'indice derivante dalla rete neurale. Prima di procedere a mostrare le serie è necessario soffermarsi a comprendere come è stata allenata la rete e come è stato possibile ricavarne un indice.
La rete neurale di tipo Feed Forward, di dimensioni (10,5,1), è stata allenata utilizzando una matrice di 1501 colonne e 35 righe, dove 1500 colonne sono il conteggio delle 1500 parole più utilizzate all'interno dei tweet più una, che è la variabile dipendente. In questo caso la variabile dipendente è il valore della volatilità il giorno successivo rispetto ai Tweet. Le 35 righe sono i 35 giorni relativi alle sette settimane di allenamento. Le colonne sono state limitate a 1500 perché altrimenti si incorrerebbe in una matrice troppo sparsa che potrebbe portare al problema del vanishing gradient, cioè alla non convergenza verso il minimo della loss function da parte della rete.
Riassumendo, la rete ha ricevuto come input il testo relativo ai Tweet di una giornata e come label il valore di volatilità della giornata successiva, così da apprendere quali informazioni di oggi possono aiutarci a prevedere domani.
L'allenamento della rete vero e proprio è stato svolto utilizzando il pacchetto neuralnet [28] che utilizza l'algoritmo di backpropagation di Riedmiller [29]. Come funzione di attivazione è stata utilizzata la sigmoide per tutti i neuroni tranne l'ultimo, che ne è privo. Infatti, essendo questo un problema di regressione, l'output finale deve essere lineare e non una probabilità tra 0 e 1 come nel caso delle classificazioni.
Una volta che i parametri della rete neurale sono stati allenati, è stata testata la sua capacità previsiva out of sample sulla nona settimana a nostra disposizione. Questo è un

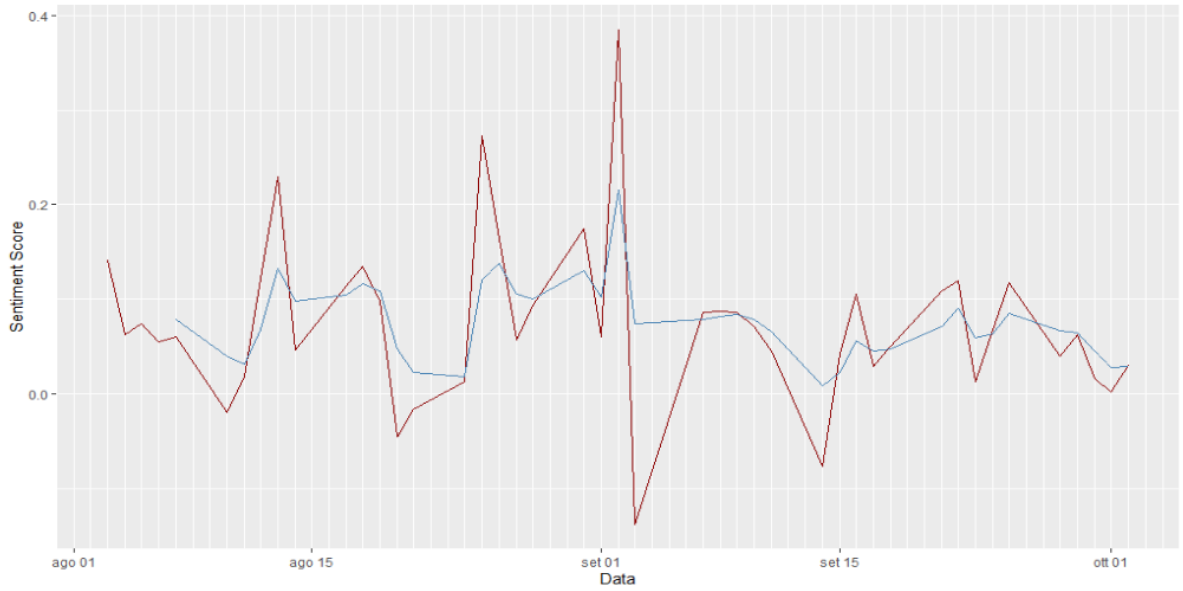

Figura 5.2: Andamento dell'indice di sentiment

passaggio fondamentale visto che le reti tendono ad arrivare molto facilmente all'overfitting, cioè a non imparare più dai dati, limitandosi a memorizzarli. Il risultato in termini di errore quadratico medio è soddisfacente e risulta uguale a $0,020$.
Dopo aver allenato la rete ed esserci assicurati che l'allenamento abbia prodotto capacità previsive utili anche out of sample, si è utilizzato l'algoritmo di Olden (presentato nella formula 2.4) per estrarre il peso assegnato a ciascuna parola fornita in input alla rete. Tali pesi sono stati successivamente standardizzati come presentato nella prossima formula:

$$Z = \frac{S - \mu}{\sigma} \tag{5.1}$$

Dove S è lo score della singola parola, μ la media degli score e σ la deviazione standard. Infine i pesi normalizzati derivanti dalla rete sono stati trattati esattamente allo stesso modo dello score binario proveniente da Opener, per generare un nuovo indice di sentiment.
Per concludere questa sezione sugli indici di sentiment, è riportato a titolo di esempio l'andamento dell'indice relativo al dizionario Opener.

In rosso è mostrato l'andamento dell'indice "puro" e in blu l'andamento dell'indice dopo lo Smoothing. Come per l'indice di volatilità i primi cinque valori non sono calcolati visto che la media mobile necessità di cinque osservazioni per essere generata. Dal grafico possiamo apprezzare la capacità dello smoothing di seguire il trend della serie originale senza cadere nella "trappola" delle variazioni troppo estreme. Probabilmente tale rumore è causato dalla forte differenza tra i diversi campioni di Tweet, sia in termini

Indice	Lag	F	Pvalue
Opener	1	8,0437	0,007449
Opener	2	3,3924	0,045726
Opener	3	3,6179	0,024282
Opener	4	2,9061	0,040288
Opener	5	1,9036	0,131103
gioia	1	16,9564	0,000213
gioia	2	4,5719	0,017675
gioia	3	2,8815	0,052202
gioia	4	1,9915	0,124273
gioia	5	1,7655	0,158229
NN	1	10,0814	0,00307
NN	2	1,4457	0,25012
NN	3	0,629	0,60197
NN	4	0,2642	0,89835
NN	5	0,4693	0,7953

Tabella 5.1: Test di Granger Significativi

di numerosità sia in termini di qualità; con campioni più ampi e con più controllo sul campionamento il rumore sarebbe stato sicuramente inferiore. In appendice sono presenti gli andamenti di tutti i restanti sentiment e le loro relative serie "smussate".

5.3 Test Di Granger

Una volta studiata la nostra variabile dipendente e calcolati i diversi indici di sentiment possiamo testare se tali indici forniscano delle informazioni utili per prevedere la volatilità realizzata. Nella tabella di seguito sono presentati i risultati dei test di Granger sulle serie che sono risultate significative.

Per una visione completa di tutti i test effettuati si rimanda alla tabella nell'appendice B.

La tabella riportata mostra: l'indice testato (Indice), il ritardo relativo a quell'indice (Lag), il valore della statistica test (F) e il relativo Pvalue (Pvalue). Data la formalizzazione del test di Granger se il pvalue è inferiore all' α prefissato, rifiutiamo l'ipotesi nulla che i parametri relativi ai regressori siano uguali a zero e quindi tali regressori esprimono informazioni rilevanti ai fini predittivi della volatilità. Da questi test possiamo capire

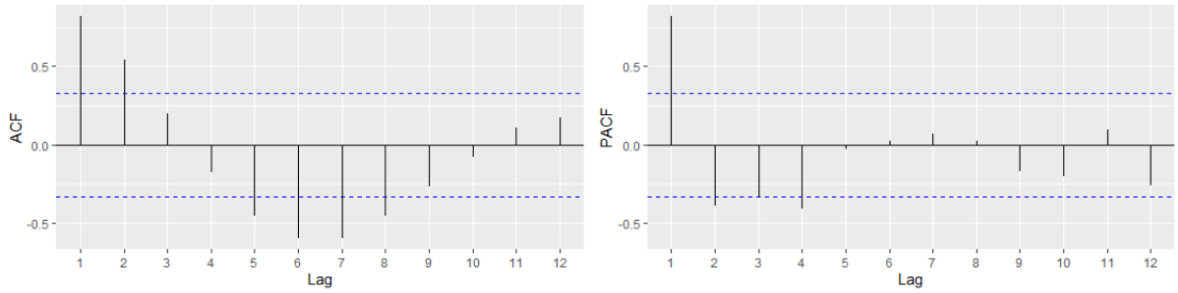

Figura 5.3: Autocorrelazione storica e parziale

che almeno il primo lag dell'indice di sentiment binario (Opener), dell'indice relativo alla gioia e dell'indice di sentiment della rete neurale (NN) sono significativi in termini di capacità previsiva sulla volatilità.
Dalla tabella si può notare anche che sono presenti altri ritardi che sono debolmente significativi in termini di Pvalue. Ma visto che ricerchiamo degli indici che siano capaci di fornirci un reale vantaggio in termini previsivi, in via precauzionale si preferisce considerare solo gli indici che hanno una significatività oltre l'uno percento.

5.4 Modelli Econometrici

Dopo aver individuato quali sono gli indici di sentiment che possono essere utili nella previsione della volatilità, è stato stimato un modello ARMA allenato solo sulla serie storica della volatilità così da avere un benchmark di confronto rispetto a modelli più complessi. Il Granger test ci ha dimostrato che tre degli indici di sentiment forniscono informazioni utili alla previsione, ma all'atto pratico va verificato se, rispetto a un modello più semplice che non utilizza regressori esogeni, l'inserimento di tali indici produca modelli più informativi e con capacità previsive migliori.
Procediamo ora ad identificare quale possa essere il miglior modello ARMA per descrivere la volatilità. Come da best practice nell'analisi delle serie storiche, per decidere quale dei possibili modelli ARMA(p,q) utilizzare sono stati studiate le visualizzazioni relative all'ACF e al PACF, riportate di seguito:

Da queste visualizzazioni si comprende che il modello ARMA da stimare deve avere almeno due ordini autoregressivi e un ordine a media mobile. Di seguito è riportata la stima di tale modello:

Dalla stima possiamo osservare che tutti i parametri stimati risultano molto più alti del loro relativo standard error, quindi sono significativi nel prevedere il processo. Sono stati testati anche modelli più complessi rispetto a quello presentato, ma non c'è

```
ARIMA(2,0,1) with non-zero mean

Coefficients:
          ar1      ar2      ma1    mean
      1.7235  -0.8874  -0.7629  0.1718
s.e.  0.0867   0.0697   0.1997  0.0038

sigma^2 estimated as 0.0002121:  log likelihood=99.21
AIC=-188.41   AICc=-186.34   BIC=-180.63
```

Figura 5.4: Parametri del modello ARMA stimati

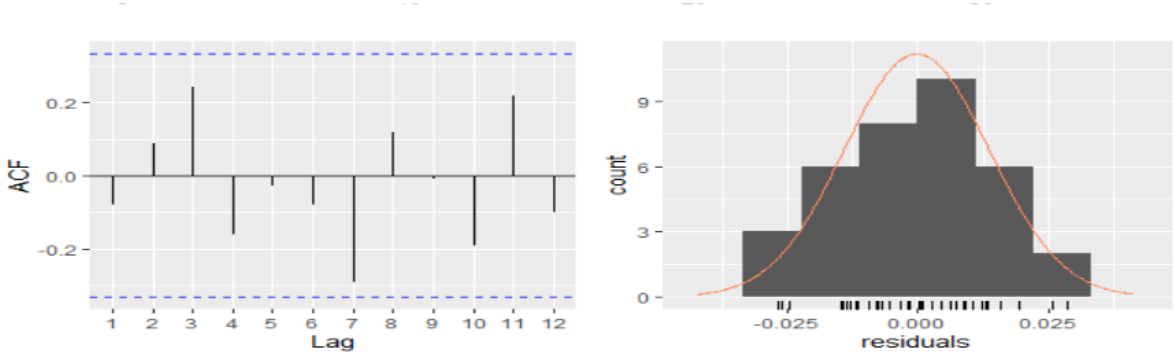

Figura 5.5: Analisi dei residui

ragione di utilizzarli in quanto no presentano un valore BIC inferiore al modello appena presentato. Inoltre l'analisi dei residui del modello ARMA(2,1) ci garantisce che i residui sono approssimabili da una normale e non sono autocorrelati, le assunzioni sul processo alla base del modello sono quindi rispettate.

I parametri del modello ARMA sono stati stimati solo su otto delle nove settimane di dati disponibili perché l'ultima settimana è stata utilizzata per valutarne le performance Out of Sample. Per stimare tale modello e produrre in pochi e semplici comandi le visualizzazioni sull'analisi dei residui e sull'ACF e PACF è stato utilizzato il pacchetto Forecast di Rob Hyndman [30]. Inoltre, per tutto ciò che riguarda il forecasting e la modellazione di serie storiche su R son stati utilizzati come riferimento i suggerimenti e gli esempi del libro : "Forecasting: Principles and Practice" di John Hyndman e George Athanasopoulos [31].

Modello	BIC	RMSE
ARMA(2,1)	-180,63	0,02970051
ARMAX gioia	-173,29	0,02326456
ARMAX NN	-174,21	0,01738032

Tabella 5.2: Performance dei diversi modelli

Dopo ad avere definito il modello ARMA(2,1) sulla serie della volatilità, sono stati definiti altri modelli con all'interno come regressori ausiliari i tre sentiment risultati significativi al Granger Test. Di seguito è riportata una tabella con i risultati dei modelli più significativi ottenuti, ossia l'ARMA (2,1), un ARMAX(2,1) con regressore ausiliare equivalente al primo lag del sentiment gioia e un ARMAX(2,1) con regressore ausiliare equivalente al primo lag del sentiment proveniente dalla rete neurale.

Nella tabella vengono mostrate le performance dei modelli in termini di criteri informativi (BIC) e dell'errore quadratico medio ottenuto sulla settimana out of sample (RMSE). Va chiarito che tra l'indice binario di Opener e l'indice relativo all gioia è stato mostrato solo il modello relativo a gioia perché è quello che raggiungeva una capacità previsiva maggiore.
Ciò che si evince da tali risultati è che, anche se dal Granger test gli indici di sentiment provenienti dai dizionari risultano significativi in termini di capacità previsiva, una volta messi alla prova su un modello già molto informativo come può essere l'ARMA(2,1), non forniscono un grande vantaggio in termini di set informativo. Ciò è sintomo del fatto che l'introduzione del parametro di stima aggiuntivo non fa aumentare a sufficienza la stima di verosimiglianza per giustificarne l'inserimento. L'indice informativo rappresenta una guida in termini di confronto tra modelli, ma non è l'unico criterio di scelta. In termini di errore quadratico medio il modello contenente l'indice relativo alla gioia presenta delle performance out of sample migliori del benchamark. Perciò il sentiment gioia fornisce delle informazioni in qualche modo utili al modello, ma in termini pratici il miglioramento delle performance è misero e non giustifica l'utilizzo di modelli più complessi rispetto al benchmark.
Anche il secondo modello ARMAX fallisce nel migliorare il criterio informativo BIC rispetto al benchmark, però il regressore ausiliario relativo all'indice di sentiment generato dalla rete neurale fornisce una forte riduzione dell'errore out of sample, infatti si passa dai 0,029 dell'ARMA(2,1) a 0,017 di RMSE. Tale risultato giustifica a pieno l'inserimento nel modello di un regressore ausiliario.
In somma, anche se gli indici di sentiment erano fortemente significativi nei test di Granger, solo quello ottenuto dalla rete neurale riesce a fornire informazioni utili a produrre un modello ARMAX con delle performance chiaramente superiori al modello ARMA di

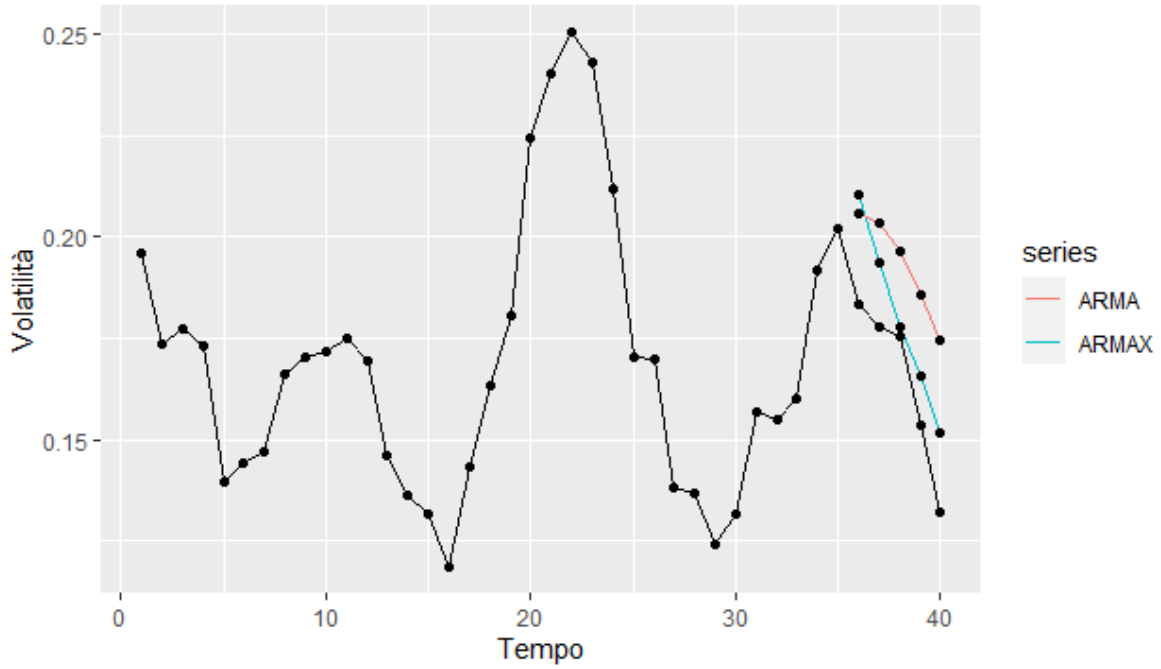

Figura 5.6: Previsione del modello ARMAX

benchmark. Infine è riportato il grafico relativo alla previsione out of sample del modello ARMAX (2,1) che utilizza lo score di sentiment della rete neurale e la previsione ottenuta dal modello ARMA (2,1).

Capitolo 6

Conclusioni

In questo elaborato si è verificato se un campione di messaggi pubblicati su Twitter, dal 3 agosto 2020 al 3 ottobre 2020, potesse fornire o meno informazioni utili sul valore futuro della volatilità realizzata dell'indice FTSE-MIB.
I risultati presentati dimostrano che almeno alcuni indici ricavati dall'elaborazione di questi Tweet sono utili a fini previsivi. Ripercorrendo i risultati ottenuti, è stato dimostrato, tramite il Granger test, che l'indice di sentiment relativo alla gioia, l'indice binario proveniente da Opner e l'indice proveniente dalla rete sono significativamente utili a prevedere l'andamento della volatilità. Ciò di per se è un risultato molto importante, perché è stato statisticamente verificato che le informazioni e le opinioni scambiate tra i vari utenti su Twitter contengono informazioni utili a prevedere la volatilità dell'indice.
Altro fatto che si evince dall'elaborato è che le deep neural network riescono ad ottenere delle ottime performance previsive anche con dati non strutturati. Tale risultato non è banale vista la grande dimensionalità che la rete riceve come input. Inoltre, è sorprendente riscontrare che la sentiment generata estraendo i pesi della rete neurale è superiore, in termini di capacità previsiva, a quelle generate manualmente o semi-manualmente. Ciò può essere giustificato dal fatto che la rete è stata allenata su un corpus di parole proveniente solo da Twitter e con determinati filtri sulle aziende quotate nel FTSE-MIB. Tali filtri probabilmente hanno reso la rete un agente esperto nel dare importanza ai termini finanziari, rispetto ai dizionari utilizzati che invece sono stati generati da utenti senza particolari focus sull'ambito di applicazione. Pertanto, la teoria per cui il valore di una parola dipende fortemente dal contesto in cui è calata si è rivelata indirettamente grazie a questa differenza di performance.
Infine si è dimostrato che includere gli indici di sentiment come regressori ausiliari nei modelli ARMAX permette di produrre modelli con capacità previsiva out of sample superiori rispetto al benchmark. Tuttavia, l'aumento significativo in termini di performance è stato fornito solo dall'indice derivante dalla rete e non dagli indici derivanti dai dizionari. Ciò ci suggerisce che sicuramente una sentiment analysis basata su dizionari specializzati per il settore della finanza avrebbe potuto fornire risultati ancora migliori,

ma visto che ad oggi questo tipo di dizionari non sono disponibili in lingua italiana, l'approccio di allenamento automatico ha fatto la differenza per fornire informazioni al modello ARMA.
Per concludere, possiamo affermare che l'elaborato è riuscito dimostrare l'utilità dei dati provenienti dal feed di Twitter nella previsione della volatilità dell'indice del mercato azionario italiano.
Ricerche successive potrebbero ampliare le fonti di dati non strutturati aggiungendo news da Bloomberg oppure articoli da Reuters. Inoltre, un altro approccio possibile potrebbe seguire l'analisi delle relazioni tra le diverse variabili anche tramite modelli capaci di gestire relazioni non lineari.
Le nuove fonti di dati non strutturati e l'evoluzione delle metodologie di data mining offrono ai ricercatori un nuovo mondo di strumenti che possono essere utili a comprendere sempre più a fondo i comportamenti umani ed economici.

Appendice A

Elementi del FTSE-MIB

Di seguito è riportata la lista delle 40 aziende che fanno parte del paniere del FTSE-MIB ad oggi.

Società	Ticker
A2A S.p.A.	A2A.MI
Amplifon S.p.A.	AMP.MI
Atlantia S.p.A.	AMP.MI
Azimut Holding S.p.A.	AZM.MI
Banca Generali S.p.A.	BGN.MI
Banca Mediolanum S.p.A.	BMED.MI
Banco BPM S.p.A	BAMI.MI
BPER Banca S.p.A.	BPE.MI
Buzzi Unicem S.p.A.	BZU.MI
Campari - Milano S.p.A.	CPR.MI
CNH Industrial N.V.	CNHI.MI
DiaSorin S.p.A.	DIA.MI
Enel S.p.A.	ENEL.MI
Eni S.p.A.	ENI.MI
Exor N.V.	EXO.MI
Ferrari N.V.	RACE.MI

Tabella A.1: Composizione FTSE-MIB

Società	Ticker
Fiat Chrysler Automobiles N.V.	FCA.MI
FinecoBank S.p.A.	FBK.MI
Generali S.p.A.	G.MI
Hera S.p.A.	HER.MI
Interpump Group S.p.A.	IP.MI
Intesa Sanpaolo S.p.A.	ISP.MI
Inwit S.p.A.	INW.MI
Italgas S.p.A.	IG.MI
Leonardo S.p.A.	LDO.MI
Mediobanca S.p.A.	MB.MI
Moncler S.p.A.	MONC.MI
Nexi S.p.A	NEXI.MI
Pirelli C. S.p.A.	PIRC.MI
Poste italiane S.p.A.	PST.MI
Prysmian S.p.A.	PRY.MI
Recordati S.p.A.	REC.MI
Saipem S.p.A.	SPM.MI
Snam S.p.A.	SRG.MI
STMicroelectronics N.V.	STM.MI
Telecom Italia S.p.A.	TIT.MI
Tenaris S.A.	TEN.MI
Terna - Rete Elettrica Nazionale S.p.A.	TRN.MI
UniCredit S.p.A.	UCG.MI
Unipol S.p.A.	UNI.MI

Tabella A.2: Composizione FTSE-MIB

Appendice B

Indici di sentiment e test

Di seguito sono riportate le serie storiche relative a tutti gli indici di sentiment calcolati tramite il dizionario NRC e le tabelle relative a tutti i test di granger effettuati.

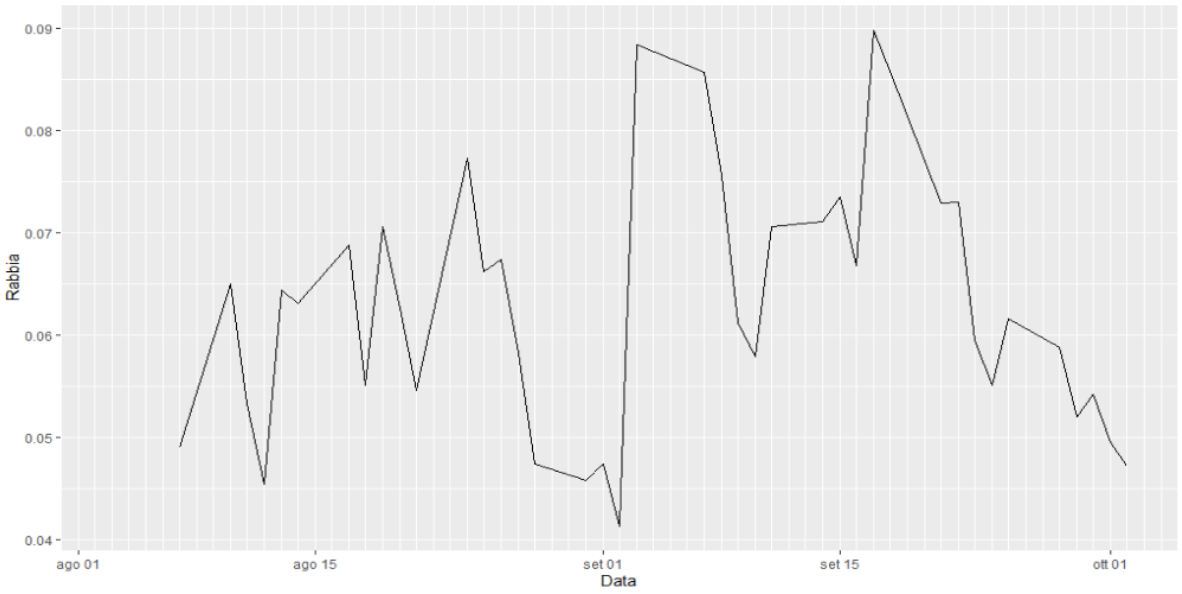

Figura B.1: Andamento sentiment relativo alla rabbia

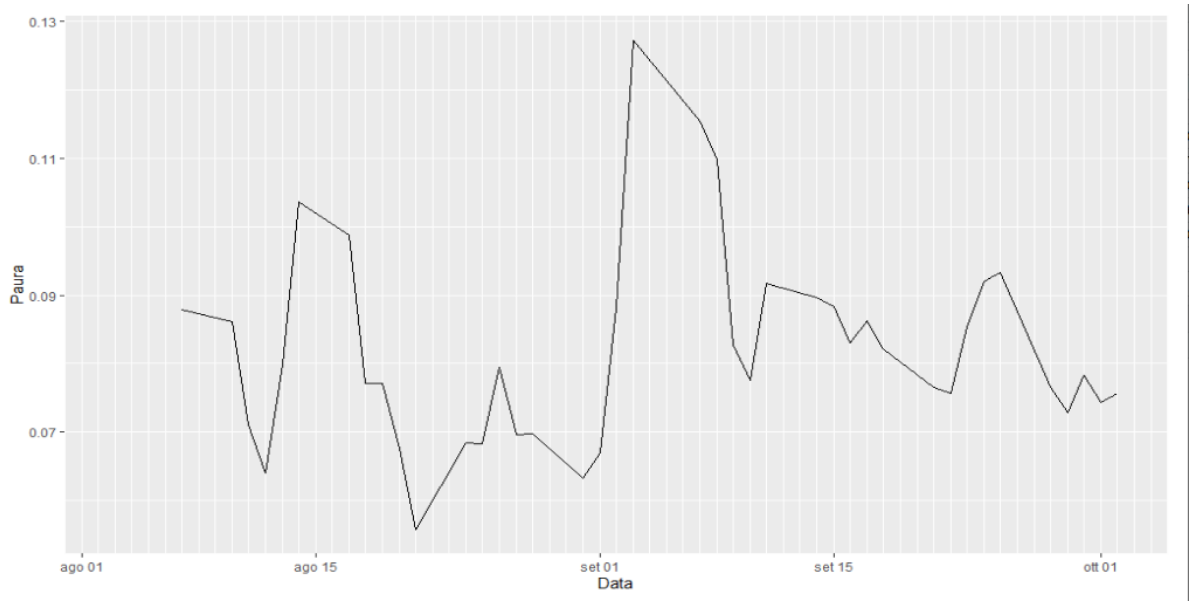

Figura B.2: Andamento sentiment relativo alla paura

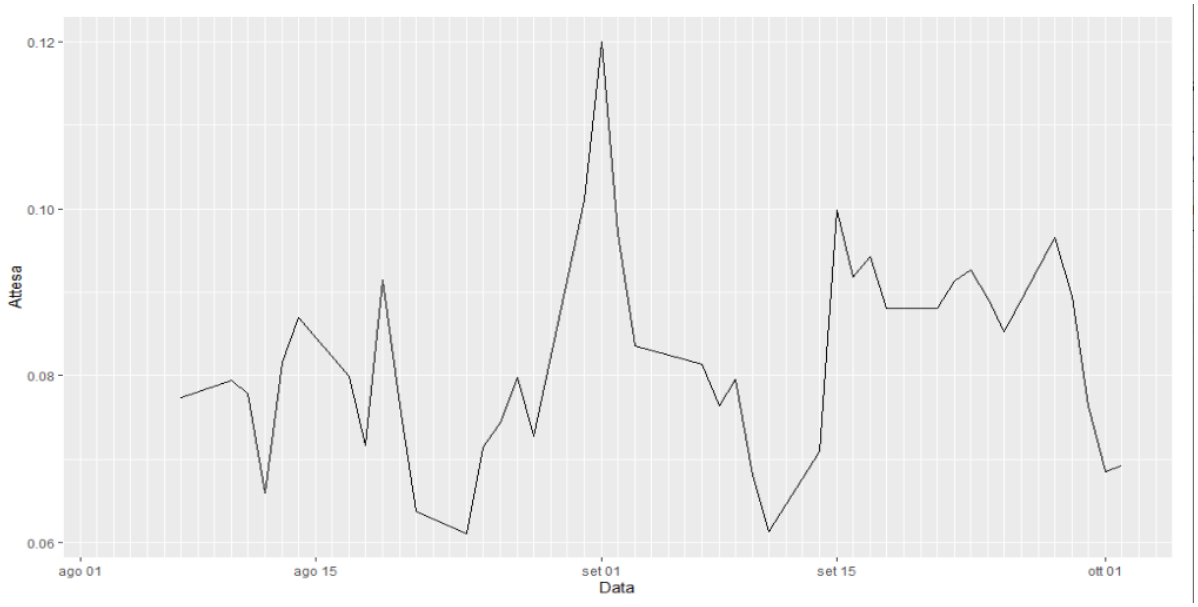

Figura B.3: Andamento sentiment relativo alla attesa

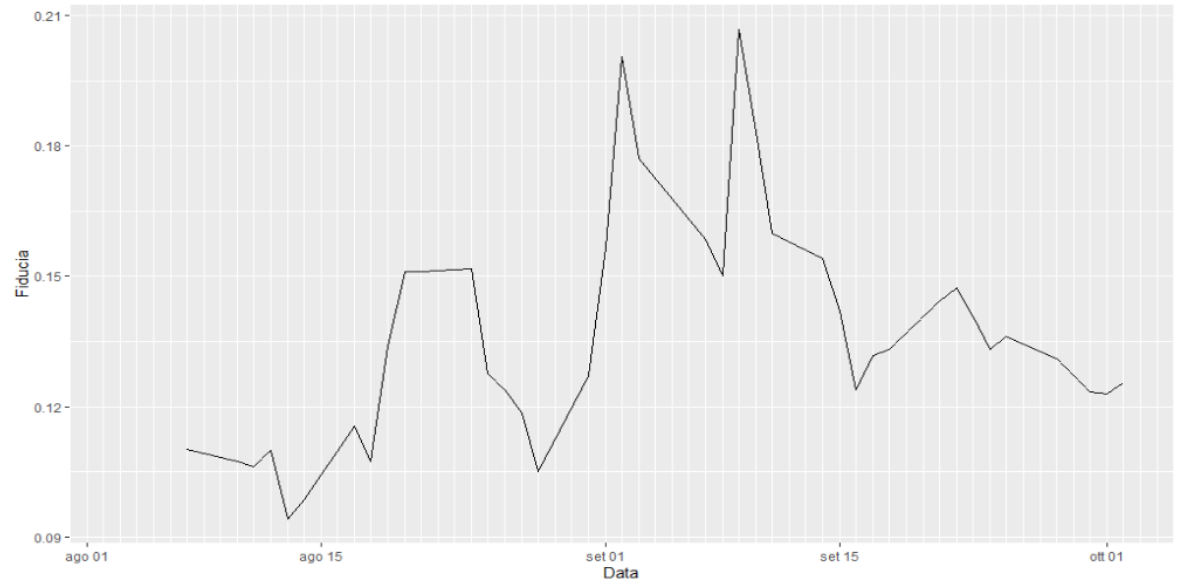

Figura B.4: Andamento sentiment relativo alla fiducia

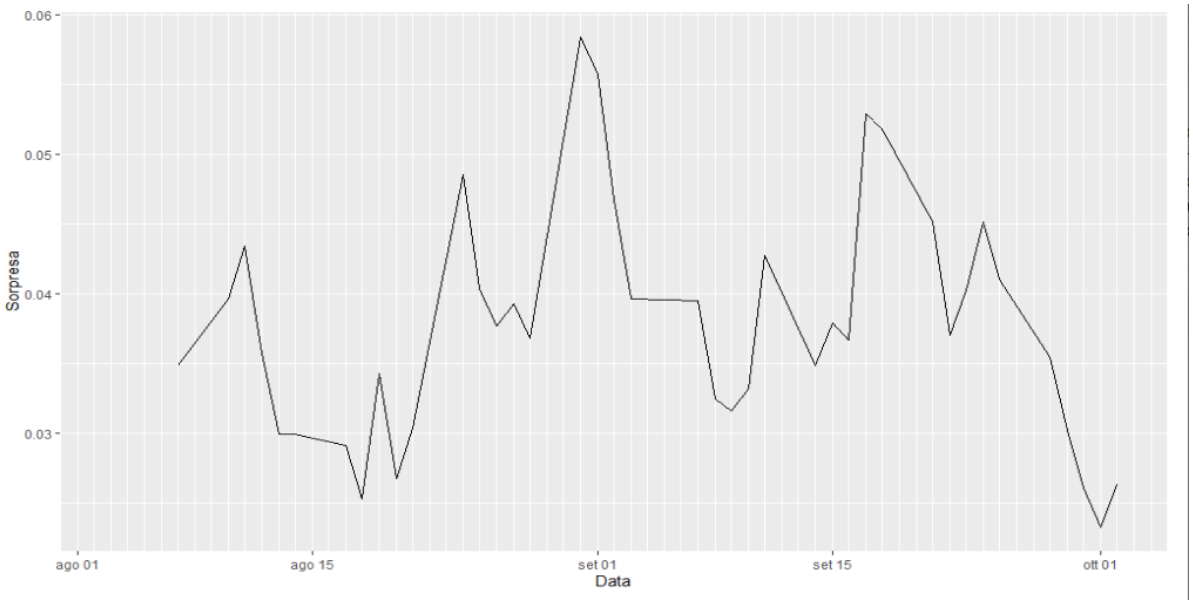

Figura B.5: Andamento sentiment relativo alla sorpresa

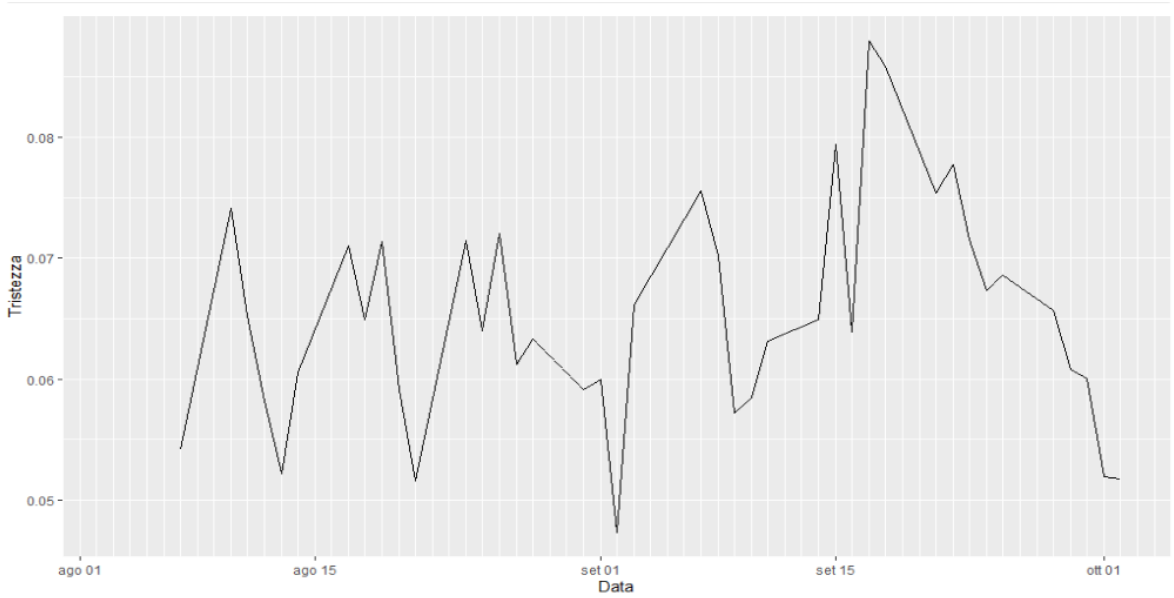

Figura B.6: Andamento sentiment relativo alla tristezza

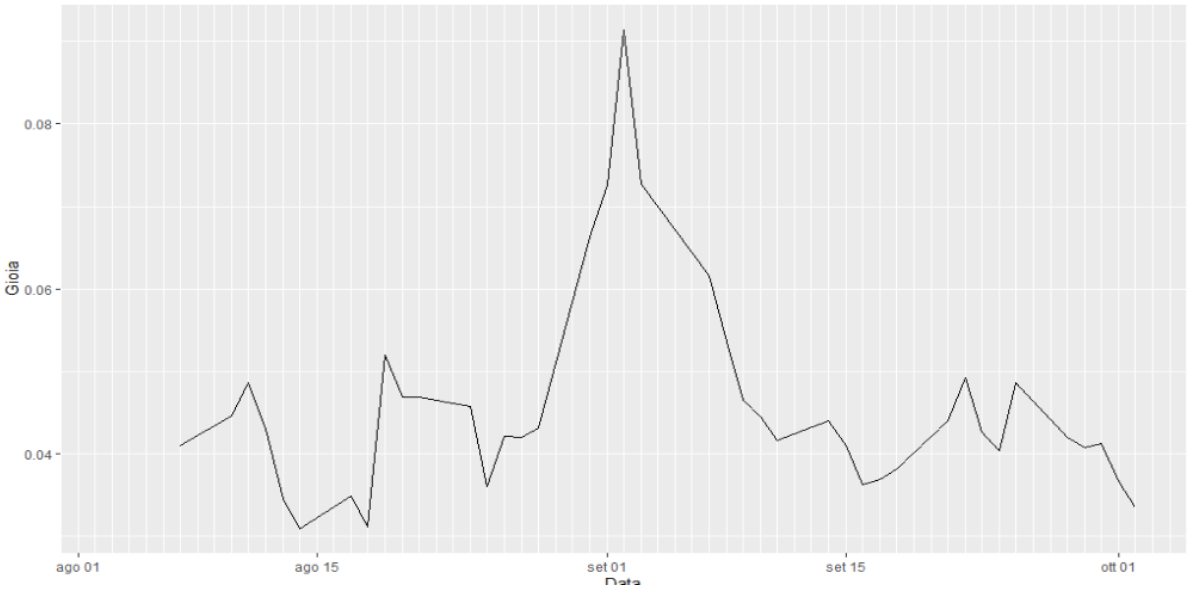

Figura B.7: Andamento sentiment relativo alla gioia

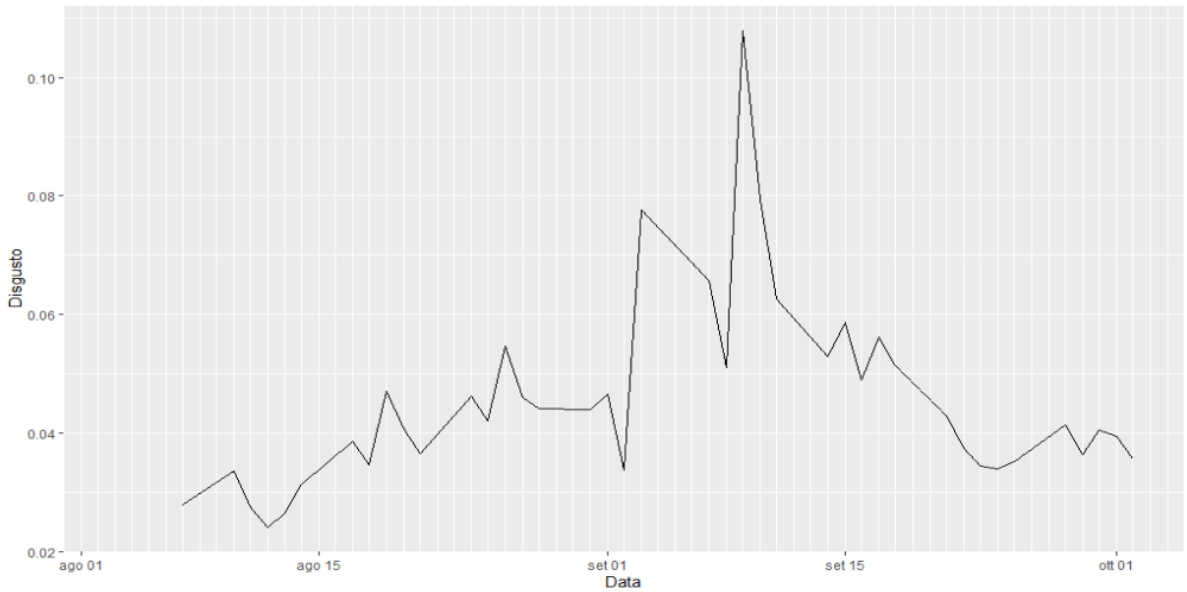

Figura B.8: Andamento sentiment relativo alla disgusto

Indice	Lag	F	Pvalue
Rabbia	1	0,2143	0,64619
Rabbia	2	0,2152	0,80750
Rabbia	3	0,6936	0,56316
Rabbia	4	0,6831	0,60979
Rabbia	5	0,9087	0,49185
Paura	1	5,9631	0,019651
Paura	2	2,5336	0,094710
Paura	3	2,0597	0,126642
Paura	4	2,7517	0,048568
Paura	5	1,7982	0,151335
Attesa	1	12,8509	0,00099
Attesa	2	4,6067	0,01720
Attesa	3	2,3017	0,09725
Attesa	4	1,3578	0,27462
Attesa	5	0,8234	0,54531

Tabella B.1: Test di Granger

Indice	Lag	F	Pvalue
Fiducia	1	1,1873	0,28312
Fiducia	2	0,4285	0,65504
Fiducia	3	0,4950	0,68851
Fiducia	4	1,2900	0,29865
Fiducia	5	0,8019	0,55937
Sorpresa	1	5,5047	0,02459
Sorpresa	2	1,0894	0,34821
Sorpresa	3	1,2591	0,30609
Sorpresa	4	0,9075	0,47359
Sorpresa	5	1,9992	0,11512
Tristezza	1	1,1100	0,29909
Tristezza	2	0,3999	0,67359
Tristezza	3	0,8159	0,49523
Tristezza	4	1,2509	0,31339
Tristezza	5	0,8138	0,55155
Gioia	1	16,9564	0,000213
Gioia	2	4,5719	0,017675
Gioia	3	2,8815	0,052202
Gioia	4	1,9915	0,124273
Gioia	5	1,7655	0,158229
Disgusto	1	0,4771	0,49416
Disgusto	2	0,0477	0,95349
Disgusto	3	0,3776	0,76983
Disgusto	4	1,0366	0,40657
Disgusto	5	0,8208	0,54698

Tabella B.2: Test di Granger

Bibliografia

[1] M. Costa L. Fanelli Gardini, A. G. Cavaliere. *Econometria*. Econometria. Franco Angeli, 2000.

[2] Robert F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007, 1982.

[3] Torben G. Andersen, Tim Bollerslev, Francis X. Diebold, and Heiko Ebens. The distribution of realized stock return volatility. *Journal of Financial Economics*, 61(1):43 – 76, 2001.

[4] Mark B. Garman and Michael J. Klass. On the estimation of security price volatilities from historical data. *The Journal of Business*, 53(1):67–78, 1980.

[5] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1 – 8, 2011.

[6] Xue Zhang, Hauke Fuehres, and Peter A. Gloor. Predicting stock market indicators through twitter "i hope it is not as bad as i fear". *Procedia - Social and Behavioral Sciences*, 26:55 – 62, 2011. The 2nd Collaborative Innovation Networks Conference - COINs2010.

[7] Steve Yang, Sheung Yin Mo, and Anqi Liu. Twitter financial community sentiment and its predictive relationship to stock market movement. *Quantitative Finance*, 15, 07 2015.

[8] V. S. Pagolu, K. N. Reddy, G. Panda, and B. Majhi. Sentiment analysis of twitter data for predicting stock market movements. In *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, pages 1345–1350, 2016.

[9] Nuno Oliveira, Paulo Cortez, and Nelson Areal. The impact of microblogging data for stock market prediction: Using twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with Applications*, 73:125 – 144, 2017.

[10] Young Bin Kim, Jun Gi Kim, Wook Kim, Jae Ho Im, Tae Hyeong Kim, Shin Jin Kang, and Chang Hun Kim. Predicting fluctuations in cryptocurrency transactions based on user comments and replies. *PLOS ONE*, 11(8):1–17, 08 2016.

[11] Jessica James, Dietmar Maringer, Vasile Palade, and Antoaneta Serguieva. Special issue of quantitative finance on 'financial data analytics'. *Quantitative Finance*, 15(10):1617–1617, 2015.

[12] Irene Russo, Francesca Frontini, and Valeria Quochi. OpeNER sentiment lexicon italian - LMF, 2016. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa.

[13] Saif Mohammad and Peter Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29, 08 2013.

[14] George Anastassiou. *Intelligent Systems: Approximation by Artificial Neural Networks*, volume 19. 01 2011.

[15] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. The MIT Press, 2016.

[16] Julian D Olden and Donald A Jackson. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological modelling*, 154(1-2):135–150, 2002.

[17] G. David Garson. Interpreting neural-network connection weights. *AI Expert*, 6(4):46–51, April 1991.

[18] Jeff Gentry. *twitteR: R Based Twitter Client*, 2015. R package version 1.1.9.

[19] Julia Silge and David Robinson. *Text Mining with R: A Tidy Approach*. O'Reilly Media, Inc., 1st edition, 2017.

[20] Matthew L. Jockers. *Syuzhet: Extract Sentiment and Plot Arcs from Text*, 2015.

[21] Matt Dancho and Davis Vaughan. *tidyquant: Tidy Quantitative Financial Analysis*, 2020. R package version 1.0.0.

[22] Charles C. Holt. Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, 20(1):5 – 10, 2004.

[23] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

[24] D. J. Bartholomew. *Operational Research Quarterly (1970-1977)*, 22(2):199–201, 1971.

[25] Unimpresa. Studio dell'associazione sul possesso delle aziende italiane. https://www.unimpresa.it/borsa-unimpresa-crolla-valore-quotate-meno-100-miliardi-in-un-anno/36314"/.

[26] BorsaItaliana. Indice ftse-mib. https://www.borsaitaliana.it/borsa/azioni/ftse-mib/lista.html?lang=en"/.

[27] Eugene F. Fama. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417, 1970.

[28] Stefan Fritsch, Frauke Guenther, and Marvin N. Wright. *neuralnet: Training of Neural Networks*, 2019. R package version 1.44.2.

[29] Martin A. Riedmiller and H. Braun. A direct adaptive method for faster backpropagation learning: the rprop algorithm. *IEEE International Conference on Neural Networks*, pages 586–591 vol.1, 1993.

[30] Rob Hyndman, George Athanasopoulos, Christoph Bergmeir, Gabriel Caceres, Leanne Chhay, Mitchell O'Hara-Wild, Fotios Petropoulos, Slava Razbash, Earo Wang, and Farah Yasmeen. *forecast: Forecasting functions for time series and linear models*, 2020. R package version 8.12.

[31] Robin John Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, Australia, 2nd edition, 2018.