UNIVERSITA' DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE ECONOMICHE ED AZIENDALI "M. FANNO"

CORSO DI LAUREA IN ECONOMIA

PROVA FINALE

"SENTIMENT INDICATORS AND ECONOMIC ACTIVITY"

RELATORE:

CH.MO/A PROF./SSA ELISA TOSETTI

LAUREANDO/A: MATTEO ZANCANARO

MATRICOLA N. 2001280

ANNO ACCADEMICO 2023 – 2024

Dichiaro di aver preso visione del "Regolamento antiplagio" approvato dal Consiglio del Dipartimento di Scienze Economiche e Aziendali e, consapevole delle conseguenze derivanti da dichiarazioni mendaci, dichiaro che il presente lavoro non è già stato sottoposto, in tutto o in parte, per il conseguimento di un titolo accademico in altre Università italiane o straniere. Dichiaro inoltre che tutte le fonti utilizzate per la realizzazione del presente lavoro, inclusi i materiali digitali, sono state correttamente citate nel corpo del testo e nella sezione 'Riferimenti bibliografici'.

I hereby declare that I have read and understood the "Anti-plagiarism rules and regulations" approved by the Council of the Department of Economics and Management and I am aware of the consequences of making false statements. I declare that this piece of work has not been previously submitted – either fully or partially – for fulfilling the requirements of an academic degree, whether in Italy or abroad. Furthermore, I declare that the references used for this work – including the digital materials – have been appropriately cited and acknowledged in the text and in the section 'References'.

Firma (signature)  ……

INDICE

Introduzione 4

1. L'importanza dell'informazione 5
   1.1 Introduzione 5
   1.2 Big data: fonti, disponibilità, caratteristiche e il loro uso in ambito economico 5
   1.3 Hard e soft data, un importante dicotomia 8
   1.4 L'utilità del text mining e le principali tecniche adottate 9

2. Analisi del sentimento: la creazione di un sentiment index 17
   2.1 Introduzione 17
   2.2 Le principali tecniche per l'analisi del sentiment 18
   2.3 Adam H. Shapiro, Moritz Sudhof e Daniel Wilson: la creazione del nuovo sentiment index 21

3. Correlazione del sentiment index con eventi storici e con i principali indicatori economici 26
   3.1 Introduzione 26
   3.2 Rstudio: correlazione tra indici 27
   3.3 I risultati ottenuti dall'analisi 31

4. Conclusione 33

Introduzione

La tesi elaborata approfondisce il tema dell'importanza dell'informazione nell'era dei Big Data e il ruolo sempre più importante del text mining e dei sentiment index nell'analisi economica, ma non solo. Il primo capitolo introduce i Big Data, esplorando le loro caratteristiche, le varie tipologie di classificazione e il loro utilizzo in ambito economico. Successivamente viene discussa la differenza tra le due principali categorie di informazione, ovvero i dati quantitativi(hard data) e quelli testuali(soft data), esaminando la crescente importanza di quest'ultimi. Il capitolo si conclude spiegando come i dati non convenzionali(soft) possano essere estratti e successivamente analizzati tramite il text mining, parlando oltre che dell'utilità di questo processo, anche delle varie tecniche e metodologie per applicarlo. Il secondo capitolo invece si concentra sulla rilevanza e sui principali approcci applicati nell'analisi del sentiment, descrivendo le principali tecniche utilizzate e le differenze tra alcuni degli strumenti informatici impiegati attualmente. Infine viene illustrato il processo di creazione del sentiment index di Shapiro, Sudhof e Wilson, discutendo le innovazioni introdotte in questo ambito ed introducendo le possibili relazioni tra l'index ed i cicli economici. Il terzo ed ultimo capitolo esamina le variazioni degli indici di sentimento creati dai tre economisti ed i principali eventi storici che si sono verificati tra il 1980 e il 2015. Concludendo viene proposta una breve parte empirica in cui, tramite l'utilizzo del linguaggio di programmazione R e della sua interfaccia Rstudio, viene analizzata la possibile correlazione tra gli indicatori di sentiment ed i principali parametri macroeconomici usati per la valutazione dell'economia degli Stati Uniti, rappresentando poi le loro linee temporali tramite il software Tableau.

1. L'importanza dell'informazione

1.1 Introduzione

Negli ultimi decenni, i rapidi progressi nella tecnologia informatica hanno rivoluzionato i metodi con cui vengono raccolti, archiviati, elaborati ed analizzati i dati. Questa evoluzione digitale ha portato alla creazione di enormi quantità di informazioni provenienti da una grande varietà di fonti, che vanno dai sensori nei dispositivi mobili alle transazioni finanziarie online, dalle interazioni sui social media alle immagini satellitari. Questi "Big Data" rappresentano una risorsa preziosa per gli economisti e i ricercatori, poiché offrono la possibilità di comprendere meglio i fenomeni economici, prevedere tendenze future e prendere decisioni informate. Nel contesto dell'economia, l'informazione è da sempre un elemento chiave per la formulazione delle politiche, la pianificazione aziendale e l'analisi dei mercati. Tuttavia, la quantità e la complessità dei dati disponibili oggi sono senza precedenti, portando a nuove sfide ed opportunità per gli studiosi. In questo contesto, il concetto di "Big Data" è diventato sempre più rilevante, suscitando un crescente interesse nel campo della macroeconomia e della previsione economica.

1.2 Big data: fonti, disponibilità, caratteristiche ed il loro utilizzo in ambito economico

Ogni giorno una miriade di informazioni vengono generate dagli utenti e, in particolar modo negli ultimi anni, sensori remoti producono costantemente sia dati strutturati che non, senza il bisogno di contatto fisico e diretto con l'oggetto o l'area monitorata. Questi dati sono noti come Big Data[1]. Originariamente il termine "Big data" si riferiva semplicemente alla vasta quantità di dati prodotti nell'era digitale ma, prendendo in considerazione che circa l'80% delle informazioni nel mondo sono "unstructured", non è sempre possibile utilizzare i tradizionali metodi di analisi. L'espressione al giorno d'oggi invece fa riferimento soprattutto, ma non solo, a tutti quei dataset di grandi dimensioni e complessità che richiedono specifiche tecniche algoritmiche per estrarne il contenuto[2]. Il grande numero di dati presenti, la grande velocità con cui sono processati e l'impossibilità di racchiuderli all'interno di ordinarie banche dati relazionali, costituiscono le caratteristiche principali dei Big Data.

[1] Nawsher Khan(2014)
[2] Dawn E.Holmes(2017)

La letteratura accademica offre una varietà di definizioni e classificazioni dei Big Data. Una delle più comuni è quella originata dal "Modello delle 3V dei Big Data" introdotto nel 2001 da Doug Laney, allora vicepresidente e service director presso l'azienda Meta Group. Le 3V indicate da Laney facevano riferimento a tre caratteristiche delle "nuove" informazioni: Volume, Velocità e Varietà. Negli anni a venire sono state aggiunte ulteriori V, ovvero Veridicità, Variabilità e Valore, portando il modello ad un totale di 6V. [3] Nello specifico[4]:

1. **Velocità**: rapidità con cui i dati sono prodotti e con cui circolano in entrata ed uscita da sistemi interconnessi;
2. **Volume**: quantità di dati raccolti ed archiviati, generati da umani, macchine e sensori;
3. **Varietà**: diversità nei dati per quanto riguarda fonti(interne ed esterne), formati(dati numerici, immagini, video, tweet ecc.) e strutture(dati strutturati, semi-strutturati e non strutturati);
4. **Veridicità**: qualità ed accuratezza dei dati raccolti, problema che si verifica data l'enorme quantità di fonti e formati con cui i dati si identificano;
5. **Variabilità**: parametro connesso alla varietà, in cui si diversifica tra dati utili ed irrilevanti ai fini dell'utilità informativa o predittiva;
6. **Valore**: legato alla scoperta di informazioni vantaggiose e alla creazione di pattern in termini di risultati sostanziali per l'organizzazione di riferimento;

Inoltre, una volta che i dati "unstructured" sono stati trasformati, possiamo distinguere tra tre tipi di Big Data principali: "Tall", "Fat" e "Huge". I Tall dataset includono poche variabili(N) ma molte osservazioni(T), con T>>N(per esempio "tick by tick" data usati per rappresentare transazioni finanziarie o in ricerche accademiche). I Fat dataset hanno invece molte variabili ma non molte osservazioni, quindi N>>T; ampi cross-sectional dataset fanno parte di questa categoria e non avendo una T abbastanza grande per poter rendere le variabili sufficientemente omogenee, non permettono l'utilizzo di modelli di stima accurati. Infine gli Huge dataset, con N e T molto grandi, sono sicuramente i più interessanti per un contesto di nowcasting(previsione) ma anche i più difficili da reperire, considerando anche la complessità legata all'elevato numero di variabili che possono provocare problematiche di non linearità, lead-lag relations(relazioni temporali tra variabili) e di micro-struttura delle informazioni(riferendosi per esempio alla frequenza con cui vengono raccolti i dati).[5]

[3] https://blog.osservatori.net/it_it/big-data-cosa-sono
[4] https://www.artera.net/it/data-science/caratteristiche-big-data/
[5] EUROSTAT REVIEW ON NATIONAL ACCOUNTS AND MACROECONOMIC INDICATORS,2017

Un ulteriore possibilità di classificazione invece ricorre ad una definizione espressa dalla divisione statistica della *United Nations Economic Commission for Europe*(UNECE), che svolge un ruolo importante nella formulazione di standard e linee guida per la raccolta, l'elaborazione e la condivisione di dati a livello internazionale, sviluppando una tassonomia dei Big Data che identifica e classifica diverse tipologie di dati in base alla loro fonte ed al loro utilizzo potenziale. Questa suddivisione aiuta a comprendere meglio le fonti e i potenziali usi dei Big Data in vari settori, facilitando così la loro gestione e analisi per applicazioni aziendali e statistiche[6]:

1) **Social Networks**(informazione di origine umana): fa riferimento alla raccolta di esperienze umane che in origine erano state documentate in libri ed opere d'arte e, col passare del tempo, in immagini audio e video. Attualmente però la maggior parte delle informazioni è digitalizzata ed archiviata ovunque, dai personal computer ai social network, con dati che spesso non vengono gestiti o sono vagamente strutturati. Fanno parte di questa categoria i dati presenti nei Social Networks, blog, e-mail ma anche le immagini, video e ricerche internet presenti nei dispositivi mobili;
2) **Traditional Business System**(dati mediati da processi): rinnovata dagli sviluppi IT dell'ultimo decennio, registra e monitora eventi di interesse come la registrazione di un cliente, la produzione di un prodotto o l'acquisizione di un ordine. I dati così raccolti attraverso processi aziendali sono strutturati e comprendono transazioni, tabelle di riferimento e relazioni, oltre ai metadati che ne definiscono il contesto. I dati aziendali tradizionali costituiscono la maggior parte di ciò che viene gestito ed elaborato dall'IT, sia nei sistemi operativi che in quelli di Business Intelligence, di solito dopo averli strutturati e memorizzati in sistemi di database relazionali;
3) **Internet of Things**(dati generati da macchine): i dati che vengono generati da macchine e sensori utilizzati per misurare e registrare situazioni ed eventi nel mondo fisico stanno diventando una parte sempre più importante delle informazioni che molte aziende memorizzano ed elaborano. La loro natura ben strutturata è adatta al trattamento informatico, ma le loro dimensioni e velocità non permettono l'utilizzo di approcci tradizionali. Gli esempi includono dati provenienti da sensori fissi (domotica, sensori meteo/inquinamento, sensori del traffico/webcam, ecc.) o sensori mobili (telefoni cellulari, auto connesse, immagini satellitari, ecc.), ma anche dati provenienti dai sistemi informatici (registri, log web, ecc.).

[6] https://statswiki.unece.org/display/bigdata2/Classification+of+Types+of+Big+Data

I molteplici tipi di Big Data risultanti sono già stati sfruttati in molti campi scientifici come climatologia, biologia, medicina e fisica applicata. Apposite aree dell'economia hanno mostrato un grande interesse nei Big Data per l'analisi aziendale, in particolare nel marketing e nella finanza. Invece, nell'economia macroeconomica convenzionale finora ci sono state applicazioni concentrate principalmente nelle aree di nowcasting/previsione.

1.3 Hard e soft data, un importante dicotomia

La distinzione principale che si fa quando si parla di informazioni è quella tra "hard" e "soft" data, ma cosa si intende con questo dualismo? Le due tipologie di dato rappresentano differenti approcci nel trattare ed interpretare informazioni, ciascuno con le proprie caratteristiche ed applicazioni.

I dati hard, o dati quantitativi, si riferiscono ad informazioni misurabili e numericamente rappresentabili. Ogni giorno abbiamo a che fare con questo tipo di data. In economia si menzionano ed utilizzano indicatori(PIL, inflazione, tasso di disoccupazione ecc.) e dati finanziari(ricavi, vendite, utile, debito ecc.), sottoponendoli ad analisi statistiche per estrapolarne tendenze ed effettuare previsioni. La principale caratteristica di questi dati è la facilità con cui possono essere raccolti, conservati, trasmessi ed elaborati, consentendo una valutazione oggettiva e precisa di fenomeni e facilitando la crescita di tecnologie come il trading quantitativo ed il lending digitale. Inoltre questo tipo di informazioni è indipendente dal contesto in cui vengono raccolte, mantenendo il loro significato nonostante l'ambito in cui sono state generate[7].

I dati soft, o dati qualitativi, sono più soggettivi e spesso descrivono caratteristiche non numericamente quantificabili. Questi includono opinioni, percezioni, emozioni ed altri fattori difficili da misurare in modo esatto. Tuttavia, i dati soft sono cruciali per comprendere aspetti qualitativi e contestuali di un fenomeno, come la soddisfazione del cliente o l'atteggiamento nei confronti di un brand. Negli ultimi anni c'è stato un crescente riconoscimento dell'importanza dei dati soft nel processo decisionale e nella previsione. Le opinioni degli utenti sui social media, le recensioni dei prodotti online e altre forme di feedback rappresentano una ricca fonte di dati soft che può essere sfruttata per informare modelli statistici o effettuare analisi del sentiment. In particolare, la creazione di sentiment indexes, o indici di sentimento, è diventata una pratica diffusa per il monitoraggio dell'opinione pubblica e per la valutazione del contesto

[7] José M.Libertiand, Mitchell A.Petersen(2018)

emotivo nei confronti di marchi o eventi. Diversamente dai dati hard quindi, questo tipo di data dipende fortemente dal contesto e dalla valutazione personale di chi li raccoglie.[8]

Di fatto, mentre i dati hard continuano ad essere fondamentali, l'uso efficace dei dati soft offre nuove prospettive ed approfondimenti per comprendere il comportamento umano ed anticipare tendenze future. La loro combinazione poi, consente di avere modelli più completi e predittivi, migliorando così la capacità di prendere decisioni informate in contesti in cui le emozioni giocano un ruolo significativo. Eppure nell'era attuale i confini tra queste due tipologie di informazioni stanno diventando sempre più sfumati. La potenza dei Big Data risiede proprio in questo, avendo la capacità di comprendere una vasta gamma di dati diversi tra loro, da quelli hard delle statistiche numeriche a quelli soft dei social media, permettendo di ottenere una visione più olistica dei fenomeni. Ad esempio, un azienda potrebbe analizzare il sentiment dei social media(dati soft) insieme ai dati di vendita(dati hard) per ottenere una comprensione più approfondita del comportamento dei clienti. D'altronde, i progressi della tecnologia stanno rendendo più facile quantificare ed analizzare i dati soft, con algoritmi di analisi del sentimento che possono trasformare post soggettivi sui social media in dati quantificabili ed algoritmi di apprendimento automatico, che possono scoprirne modelli e tendenze.[9]

Questa tesi si concentrerà prevalentemente sull'utilizzo di dati non convenzionali, la loro applicazione in ambito statistico-macroeconomico e il loro impiego in modelli di previsione, seguendo il working paper "Measuring News Sentiment", scritto da *Shapiro, Adam Hale, Moritz Sudhof, Daniel Wilson.*

1.4 L'utilità del text mining e le principali tecniche adottate

L'approccio generale usato per misurare il sentimento economico consiste nella formazione di indici di sentimento a partire da sondaggi, esempi già fatti sono quelli del *Michigan Consumer Sentiment index* e del *Conference Board's Consumer Confidence index.* Tuttavia le misure basate sull'analisi computazionale del testo stanno diventando sempre più popolari tra i ricercatori, grazie ai loro vantaggi rispetto ai sondaggi in termini di costo, portata e tempestività. I sondaggi sono per natura più costosi da condurre, spesso si basano su campioni relativamente piccoli di individui(possibili problemi di campionamento) e tendono inoltre ad essere pubblicati con una frequenza mensile, riducendo il loro valore nei momenti di svolta economica. La Computational Text Analysis(CTA), o text mining(TM), d'altro canto, offre una maggiore

[9] https://www.reportermanagement.com/blog/hard-data-vs-soft-data/

scalabilità e rapidità rispetto ai sondaggi, usando algoritmi e software appositamente progettati per poter analizzare grandi quantità di testo in tempi piuttosto brevi, consentendo una valutazione più rapida e aggiornata dell'opinione pubblica. Questo approccio consente di avere un interpretazione più accurata del sentimento, prendendo in considerazione una varietà di elementi linguistici e contestuali che possono sfuggire nei sondaggi tradizionali, grazie all'utilizzo di algoritmi che non analizzano solo le parole utilizzate, ma anche il contesto circostante: tono, sarcasmo, contesto culturale. Un altro vantaggio è la capacità di analizzare dati non strutturati provenienti da una vasta gamma di fonti(social media, recensioni online o anche articoli di giornale), offrendo un quadro più completo e diversificato delle opinioni ed "emozioni" del pubblico rispetto ai tradizionali sondaggi, limitati nel tempo e nello spazio. Tuttavia, ci sono anche svantaggi associati all'utilizzo dell'analisi computazionale del testo, ad esempio, la presenza di errori di interpretazione dovuti alla complessità del linguaggio umano e alla variabilità dei contesti che potrebbero limitarne l'accuratezza.[10]

L'evoluzione dell'informatica ha drasticamente cambiato l'approccio e gli strumenti utili legati all'analisi automatica dei testi ed alla statistica testuale che era iniziata a partire dagli anni 1960-1970, fondandosi non solo su "statistical tools" ma soprattutto sull'integrazione di quest'ultimi con strumenti informatici e linguistici, dando vita al termine Text Mining(TM). Questo tipo di processo viene definito come "la scoperta, tramite computer, di nuove informazioni precedentemente sconosciute, estraendo automaticamente dati da diverse risorse scritte".[11] Lo step fondamentale nel Text mining, da molti chiamato anche KDD(Knowledge Discovery from Data), è sicuramente la trasformazione di dati non strutturati(testi) in dati semi-strutturati. Una volta effettuata questa conversione non c'è nulla che impedisca di applicare una qualsiasi delle tecniche di analisi per classificazione, raggruppamento o previsione.[12] I "text data", a confronto con le informazioni numeriche, sono ambigui e difficili da processare. Tuttavia, nella cultura moderna, le informazioni testuali sono la forma più diffusa per lo scambio formale di documentazione. E in effetti, nell'era digitale in cui ci troviamo, siamo immersi in enormi quantità di soft data, generati in ogni singolo momento attraverso molteplici fonti: dai social media ai rapporti finanziari, dai messaggi di posta elettronica alle recensioni dei prodotti. Questa vasta mole di testo rappresenta un importante bacino di informazioni, spesso però inutilizzate a causa della loro complessità e dimensione.

[10] Shapiro,Wilson e Sudhof(2020)
[11] S.Bolasco(2005)
[12] V.Kotu, B.Deshpande(2018)

L'analisi del testo computazionale prova a superare questa sfida, fornendo strumenti e tecniche necessarie per l'estrazione del significato da grandi quantità di testo in modo efficiente. Questo settore, in continua crescita, combina elementi di linguistica computazionale, intelligenza artificiale, statistica e matematica, affermandosi come un pilastro fondamentale per tutti i data analyst. L'obiettivo che si prefissa è quello di convertire i testi in dati strutturati ed analizzabili, consentendo di scoprire pattern, trend ed ottenere insight esaurienti su una vasta gamma di argomenti. Alcuni esempi[13]:

- **Servizio clienti**: quando combinati con gli strumenti di analytics del testo i sistemi di feedback, come chatbot, sondaggi sui clienti, recensioni online e profili di social media consentono alle aziende di migliorare rapidamente la loro "esperienza cliente". Il text mining consente alle organizzazioni di individuare le criticità per i propri clienti(assegnandogli una priorità) e di rispondere a problematiche urgenti in tempo reale, migliorandone la soddisfazione;
- **Gestione del rischio**: può fornire una panoramica dettagliata sulle tendenze di settore e sui mercati finanziari, monitorando le variazioni di "emozione" degli investitori ed estraendo informazioni da whitepaper e report di analisti. Particolarmente prezioso per le istituzioni bancarie, fornendo maggiore fiducia per eventuali investimenti in determinati settori;
- **Manutenzione**: fornisce una visione completa e dettagliata del funzionamento e delle caratteristiche di prodotti e macchinari. Nel tempo ha automatizzato il processo decisionale rivelando modelli correlati a problemi e procedure di manutenzione, aiutando i professionisti a scoprire rapidamente le cause principali di eventuali malfunzionamenti;
- **Assistenza sanitaria**: le tecniche di text mining sono diventate sempre più preziose per i ricercatori in campo biomedico, in particolare per il raggruppamento delle informazioni. L'utilizzo di procedure manuali può risultare costoso ed allungare di molto i tempi d'attesa, mentre il TM fornisce un metodo di automazione per estrarre informazioni utili dalla letteratura medica.

[13] https://www.ibm.com/it-it/topics/text-mining

In Figura 1 vengono mostrate e successivamente spiegate le varie aree o campi del text mining, dividendole in recupero ed estrazione delle informazioni, Natural Language Processing(NLP) e Data Mining.[14]

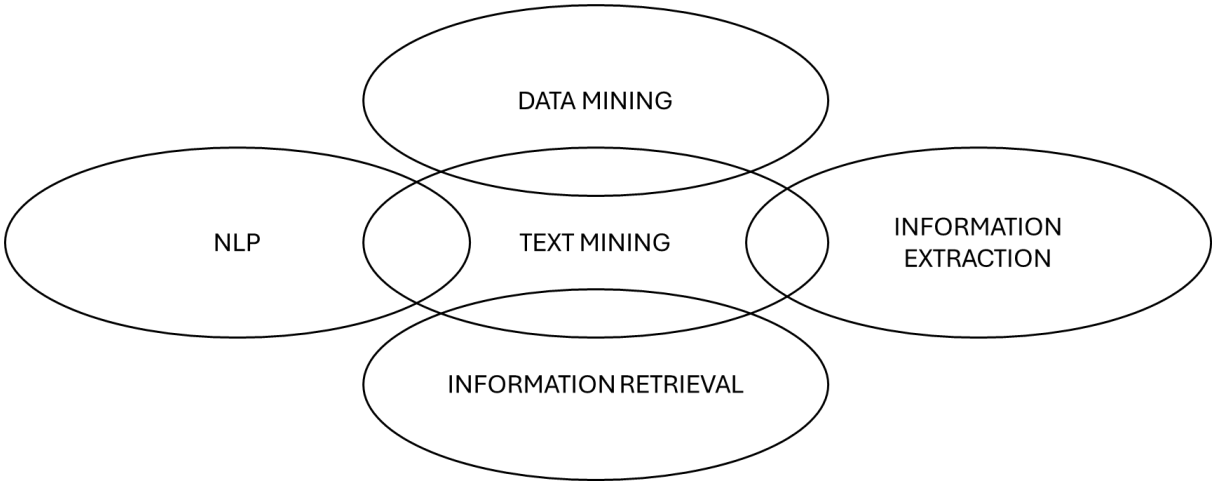

Figura-1

Queste tecniche consentono di dedurre informazioni provenienti da dati testuali non strutturati, suddividendosi in diverse sub-attività. Tuttavia, prima di poter applicare queste procedure è necessario partire dalla pre-elaborazione del testo(o data pre-processing) che consiste nella pulizia e trasformazione dei dati testuali in un formato utilizzabile nell'analisi. Questa pratica risulta fondamentale nell'elaborazione del linguaggio naturale(NLP) e generalmente coinvolge tecniche come la language identification, tokenization, part-of-speech tagging, il chunking e la syntax parsing. In particolare:

- **Langauge identification** *:* il primo passo nell'analisi del testo è identificare in quale lingua è scritto il documento. Ogni lingua ha le proprie particolarità, quindi è importante sapere con cosa abbiamo a che fare. Per quanto semplice possa sembrare, l'identificazione della lingua determina l'intero processo per ogni altra funzione di analisi del testo;
- **Chunking**: è un processo nell'analisi del linguaggio naturale che consiste nel raggruppare parole correlate in unità sintattiche più grandi, chiamate "chunk" o "frammenti". Questi chunk possono comprendere parole, frasi o parti di discorso correlate che svolgono una funzione grammaticale specifica all'interno di una frase. A differenza del POS tagging, il chunking mira ad identificare gruppi di parole che

[14] L.Kumar,P.K.Bhatia(2013)

formano unità semantiche o sintattiche più significative e a raggrupparle in chunks. Questo step permette di individuare frasi che nel complesso possano essere più significative per il risultato che si vuole ottenere;[15]

- **Syntax parsing**: implica l'analisi della struttura grammaticale delle frasi per l'identificazione delle relazioni sintattiche tra le parole. In sostanza, l'analisi sintattica si occupa di comprendere come le parole di una frase si combinano tra di loro, secondo le regole grammaticali della lingua, per formare significati complessi. L'analisi sintattica mira a creare un albero di analisi o una struttura che rappresenti le relazioni gerarchiche tra le parole in una frase. Spesso per il syntax parsing vengono utilizzati modelli speciali di apprendimento automatico senza supervisione, basati su miliardi di parole in input e su una complessa fattorizzazione della matrice, per aiutarci a comprendere la sintassi proprio come farebbe un essere umano;[16]

Una volta completata questa prima fase è possibile passare ai procedimenti rappresentati in Figura 1, che comprendono l'applicazione di algoritmi di text mining per ricavare approfonditamente i dati presenti nei testi.

Recupero delle informazioni(information retrieval, IR)

Si tratta di una estensione del recupero dei documenti, in cui essi vengono restituiti ed elaborati per condensare e prelevare specifiche informazioni richieste dall'utente. Questa fase potrebbe essere seguita da un processo di riassunto del testo che si concentra sulla query posta dall'analista, o da una fase di estrazione delle informazioni usando tecniche specifiche. I sistemi di IR aiutano a restringere il set di documenti rilevanti per un particolare scopo e, poiché il text mining implica l'applicazione di algoritmi molto complessi a grandi quantità di dati, l'IR può accelerare significativamente l'analisi riducendo il numero di documenti da analizzare. Il reperimento delle informazioni è comunemente usato nei sistemi di catalogazione di biblioteche e nei motori di ricerca popolari, come Google. Alcune comuni sub-attività sono:

- **Tokenization**: anche detta microsegmentazione del corpus, la tokenization(o tokenizzazione) è il processo che permette di scomporre i documenti di testo in tokens, ovvero singole unità di significato con cui si opera che possono essere parole, fonemi o addirittura intere frasi. Ciononostante, la tokenization viene detta language-specific,

[15] https://towardsdatascience.com/chunking-in-nlp-decoded-b4a71b2b4e24
[16] https://medium.com/@datailm/the-essential-role-of-syntactic-and-semantic-parsing-in-nlp-47b92118d9de

ovvero cambia le sue regole a seconda della lingua in cui è scritto il testo. Come sappiamo non tutte le lingue usano un alfabeto, alcune sono logografiche(es. cinese) e non presentano spazi tra le parole, necessitando quindi di tecniche di machine learning per ultimare questo processo.

- **Stemming e lemmatization:** operazioni solitamente opzionali, lo stemming serve a ridurre le parole alla loro radice(stem in inglese) mentre la lemmatization è la sua alternativa, che permette la trasformazione di una parola in forma flessa nella sua forma canonica, detta appunto lemma.

Natural Language Processing(NLP)

È considerato uno dei problemi più vecchi e complessi nel campo dell'intelligenza artificiale e tratta lo studio del linguaggio umano, ovvero come i computer possono comprendere ed elaborare la lingua scritta in modo simile a quello che facciamo noi umani. Il NLP tratta la comprensione del significato di una frase o di un documento poiché, mentre le parole sono gli elementi costitutivi del significato, la loro correlazione all'interno della struttura di una frase comporta il vero contenuto che si vuole far arrivare al lettore. Tutto questo viene svolto tramite l'utilizzo di algoritmi per la comprensione e l'analisi del linguaggio, oltre a procedure derivanti da varie discipline come l'informatica, l'intelligenza artificiale, la linguistica e la data science. Alcune comuni sub-attività sono:

- **Summarization**: fornisce una sinossi di lunghi pezzi di testo per creare un riassunto conciso e coerente del documento;
- **Part-of-Speech (PoS) tagging**: consiste nell'assegnare una categoria grammaticale (come sostantivo, verbo, aggettivo, ecc.) a ciascuna parola di una frase. L'obiettivo è comprendere la struttura sintattica di una frase ed identificare i ruoli grammaticali delle singole parole. In passato eseguito manualmente, il POS tagging ora viene effettuato utilizzando algoritmi che associano termini discreti e parti del discorso nascoste ad un insieme di etichette descrittive. Gli algoritmi utilizzati si suddividono in due gruppi distinti: basati su regole e stocastici. Un esempio di "tagger" ampliamente utilizzato è quello di E.Brill(1993), uno dei primi creati basato sull'utilizzo di regole;[17]

[17] https://en.wikipedia.org/wiki/Part-of-speech_tagging#:~:text=In%20corpus%20linguistics%2C%20part%2Dof,its%20definition%20and%20its%20context.

- **Sentiment anlaysis**: rileva i sentimenti positivi, negativi o neutrali da dati testuali, consentendo di monitorare i cambiamenti nel comportamento delle persone. Viene comunemente usato per fornire informazioni sulla percezione dei marchi, prodotti o servizi ma trova le sue applicazioni anche in aree della finanza o della macroeconomia.

Estrazione delle informazioni(Information extraction,IE)

Si tratta del processo in cui vengono automaticamente estratti dati strutturati da informazioni testuali(quindi non strutturate o semi-strutturate), memorizzando questi dati rilevanti in database per una successiva analisi dettagliata. Le operazioni di elaborazione di documenti multimediali, come l'annotazione automatica e l'estrazione di informazioni da immagini/audio/video, sono tra i migliori esempi, tra tutti l'IE di Google Search Engine. Alcune comuni sub-attività sono:

- **Feature selection**: o "selezione delle caratteristiche", è il processo di selezione delle caratteristiche importanti(dimentions) per contribuire maggiormente all'output di un modello di analisi predittiva;
- **Feature extraction**: trasformazione del testo in una rappresentazione numerica o simbolica, esempi comuni sono i metodi Bag-of-Words(BoW) o TF-IDF. Questo processo è fondamentale per trasformare il testo in un formato compatibile con gli algoritmi di machine learning, riducendone la dimensionalità e mantenendo la maggior parte delle informazioni rilevanti. Nello specifico:
    - BoW: rappresenta un testo come un insieme di parole ignorandone l'ordine e la grammatica. In questo modello, ogni documento viene considerato in quanto contenente parole, analogamente ad una borsa(bag in inglese); ciò consente una gestione di queste basata su liste, dove ogni borsa contiene determinate parole di una lista;
    - TF-IDF: il term frequency-inverse document frequency è una funzione utilizzata per misurare l'importanza di un termine rispetto ad un documento o ad una collezione di documenti. Questa funzione aumenta proporzionalmente al numero di volte che il termine è contenuto nel documento, ma cresce in maniera inversamente proporzionale con la frequenza del termine nella collezione. L'idea alla base di questo comportamento è di dare più importanza ai termini che compaiono nel documento, ma che in generale sono poco frequenti.[18]

[18] https://it.wikipedia.org/wiki/Tf-idf

- **Named-entity recognition (NER):** chiamata anche entity identification(identificazione dell'identità) o entity extraction(estrazione dell'identità), mira a trovare e categorizzare termini specifici nel testo come nomi o posizioni. Ad esempio, NER può identificare "California" come luogo e "Mary" come nome di una donna.

Data mining(DM)

Può essere descritta come una ricerca di pattern nelle informazioni che si analizzano, estraendo dati nascosti e di cui non si era a conoscenza. Gli strumenti utilizzati in questa fase possono predire comportamenti e tendenze future, consentendo alle aziende o in generale agli utenti di prendere decisioni informate. Questi strumenti cercano nei database pattern nascosti e sconosciuti, trovando informazioni critiche che gli esperti potrebbero trascurare(es. opinioni dei clienti, come tendenze di acquisto, preferenze di prodotto o sentimenti associati a marchi specifici) . L'obiettivo generale del Data Mining è estrarre informazioni da un insieme di dati e trasformarle in una struttura comprensibile per un ulteriore utilizzo, solitamente di analisi.

2. Analisi del sentimento: la creazione di un sentiment index

2.1 Introduzione

Le origini della sentiment analysis si trovano nell'analisi dell'opinione pubblica all'inizio del XX secolo e negli studi sulla soggettività dei testi condotti dalla comunità di linguistica computazionale negli anni '90. Tuttavia, la grande disponibilità di testi sul web ha portato alla vera espansione di questo tipo di analisi e, di conseguenza, il 99% degli articoli relativi all'argomento sono stati pubblicati dopo il 2004. L'analisi del sentimento è cambiata molto negli ultimi anni, inizialmente concentrata sullo studio delle recensioni di prodotti online, ora include testi provenienti da tutta la rete, soprattutto social media platform come Twitter o Facebook. Oggi questa tecnologia è utilizzata in diversi ambiti come i mercati azionari, le elezioni, la medicina, l'ingegneria del software e il cyberbullismo.[19]

Nonostante questo tipo di analisi sia utilizzata per diversi obiettivi, un possibile fine ultimo è la creazione di un indice che rappresenti il sentimento delle persone nel corso del tempo e che possa essere implementato in modelli statistici per la previsione dei dati. Alcuni esempi sono:

- **Frear and Greed Index**: creato da CNNMoney, viene utilizzato per valutare se gli investitori siano influenzati dalla paura o dall'avidità, misurandola sul mercato azionario americano. Esso si basa su una combinazione di sette diversi indicatori che misurano alcuni aspetti del comportamento del mercato azionario(slancio del mercato, la forza del prezzo delle azioni, l'ampiezza del prezzo delle azioni, le opzioni put e call, la domanda di obbligazioni spazzatura, la volatilità del mercato e la domanda di beni rifugio);[20]
- **CBOE Volatility Index** (VIX) o "Fear Index": è un indicatore primario della volatilità del mercato azionario, offre un'idea di come i professionisti della finanza percepiscano le condizioni di mercato a breve termine. Misura quindi la volatilità che, secondo gli investitori, subirà l'indice S&P 500 nei seguenti 30 giorni.[21] Si basa sulla volatilità implicita calcolata sul mercato delle opzioni, riflettendo le aspettative future circa i movimenti del prezzo delle azioni sottostanti;[22]

[19]M.V.Mäntylä , D.Graziotin, M.Kuutila(2016)

[20] https://www.finanzaonline.com/notizie/fear-greed-index-troppa-avidita-sui-mercati-possibile-inversione
[21] https://www.forbes.com/advisor/it/investire/vix-indice-volatilita/
[22] https://www.forbes.com/advisor/it/investire/vix-indice-volatilita/

- **Social Mood on Economy Index**(SME): indice sviluppato da ISTAT che utilizza campioni di tweet in lingua italiana per analizzare le percezioni dei cittadini sull'andamento dell'economia. La creazione dell'indice è stata possibile tramite l'estrazione di circa 55.000 tweet giornalieri contenenti parole specifiche, utilizzando lessici per l'assegnazione di valori positivi o negativi a ciascun termine. Inizialmente sperimentale, ora è un importante componente dell'analisi economica contemporanea in Italia.[23]

2.2 Le principali tecniche per l'analisi del sentiment

Come discusso in precedenza, l'approccio generale per la creazione di un sentiment index fa affidamento sui sondaggi, tuttavia gli autori nel loro working paper presentano una tecnica di analisi NLP per misurare il sentiment economico nel corso del tempo. Prima di discuterne comunque è giusto esaminare quali siano gli approcci più diffusi per l'analisi del sentiment e soprattutto dare una definizione chiara di cosa andremo ad analizzare.

Il sentiment in un testo(parola, espressione, frase etc.) è tipicamente espresso come un problema di classificazione ternario(positivo, negativo, neutrale) o di rating(valenza su una scala da 1 a 5) ed è la misura del tono, atteggiamento o valutazione di un argomento da parte dell'autore, indipendentemente dall'orientamento "sentimentale" del topic(per esempio un horror può risultare piacevole). La letteratura dell'analisi del sentimento sottolinea due obiettivi chiave nel caratterizzare il sentiment di un dato insieme di testo: specificità del dominio e complessità. Il dominio fa riferimento all'oggetto del corpus di testo che si vuole analizzare, del resto le parole possono avere diversi significati in domini differenti. La complessità invece riguarda tutti gli aspetti multisfaccettati di un testo, andando oltre alla semplice prevalenza di determinate parole. L'espressione del sentimento è composizionale e contestuale e la sua complessità è evidente in caratteristiche semplici come la negazione, dove una singola parola può influenzare direttamente l'orientamento sentimentale delle parole che seguono. Ad esempio, da "buono" a "non buono", così come in frasi più composizionali come "vorrei poter dire che mi è piaciuto", dove il sentimento di un'espressione è chiaramente diverso dalla somma delle sue parole. Per quantificare il sentiment in un testo vengono individuati generalmente due metodi principali: **Lexical method** e **Machine learnings techinques**.

[23] https://www.infodata.ilsole24ore.com/2019/02/23/istat-misura-leconomia-twitter-funziona-social-mood-index/ e https://www.infodata.ilsole24ore.com/2019/11/24/la-recessione-governo-conte-social-mood-degli-italiani/

Il primo approccio fa affidamento a liste di parole predefinite chiamate lexicons(in italiano lessici) dove ad ogni parola viene assegnato un punteggio per il sentimento di interesse, solitamente 1, 0 o -1 per rappresentare sentimenti rispettivamente positivi, neutri o negativi ma non si limitano a questa rappresentazione. Una tradizionale applicazione di questo metodo misura il sentimento contenuto in un testo basandosi sulla prevalenza di parole positive rispetto a quelle negative: questo tipo di metodo già menzionato viene chiamato Bag-of-Words(BOW) perché le caratteristiche contestuali di ciascuna parola, come il suo ordine all'interno del testo, la parte del discorso, la co-occorrenza con altre parole e altre caratteristiche contestuali specifiche del testo in cui appare, vengono ignorate. I recenti progressi dei metodi lessicali di analisi del sentimento si sono concentrati sull'integrazione delle caratteristiche contestuali delle parole all'interno del corpus di interesse. Alcuni lexicon specifici molto utilizzati nell'analisi del sentiment(ma non solo) sono Vader, SentiWordNet 3.0(estensione di WordNet) e AFIIN.

Vader, sviluppato da Hutto e Gilbert (2014) è un classificatore di sentimento a livello di frase che consiste in un lessico composto da diverse migliaia di parole ("unigrammi") etichettate da -4 a 4 corrispondenti al massimo negativo e al massimo positivo, oltre che in un insieme di regole euristiche che tengono conto del contesto di una parola all'interno della frase. Vader assegna un punteggio di negatività ad una frase aggregando i punteggi(di negatività) delle parole al suo interno. Il punteggio di un termine inizia con il suo valore numerico di negatività nel lessico, ma viene poi aumentato o diminuito moltiplicatamente, in base al contesto all'interno della frase. Il contesto è catturato da un insieme di semplici regole relative alla negazione, alla punteggiatura, alla capitalizzazione, all'essere precedute o seguite dalla parola "ma," ed all'essere precedute da un modificatore di grado come "molto," "estremamente," "leggermente".

SentiWordNet 3.0 è una risorsa lessicale avanzata appositamente ideata per supportare le applicazioni di classificazione del sentiment e opinion mining.[24] Questo strumento è pubblicamente disponibile ed è attualmente concesso in licenza a più di 300 gruppi di ricerca in tutto il mondo. Si tratta del risultato dell'annotazione automatica di tutti i synset(insieme di sinonimi che possono essere descriti da un'unica definizione) di WordNet, un database linguistico sviluppato dal linguista George Miller presso l'Università di Princeton che organizza, definisce e descrive i concetti rilevanti della lingua inglese.[25] Ogni synset è associato a tre punteggi numerici che indicano quanto positivi(Pos) negativi(Neg) o neutrali(Obj, in inglese Objective) siano i termini contenuti nel synset. Ciascuno dei punteggi varia

[24] Pang, Lee(2008)
[25] https://www.ittig.cnr.it/Ricerca/materiali/JurWordNet/WordNet.htm

nell'intervallo [0.0, 1.0] e la loro somma è 1 per ogni synset, garantendo che i punteggi siano normalizzati e che rappresentino proporzioni relative tra positività, negatività e neutralità. Ad esempio, il synset [estimable(J,3)], corrispondente all'italiano "può essere calcolato o stimato" ha un punteggio Obj di 1.0(e quindi Pos=Neg=0.0), mentre il synset [estimable (J,1)] corrispondente a "degno di rispetto o alta considerazione" ha un punteggio Pos di 0.75(Neg= 0.0 e Obj=0.25).[26]

AFIIN è un lexicon sviluppato da Finn Årup Nielsen tra il 2009 e il 2011 a partire da tweet relativi alla Conferenza sul Clima delle Nazioni Unite(COP15). Le parole presenti al suo interno sono state valutate utilizzando tre dimensioni di sentimento: valence, arousal e dominance(traducendo in italiano sarebbero rispettivamente valenza, eccitamento e dominanza). È stato argomentato che emozioni individuali come gioia, rabbia e paura siano punti in uno spazio tridimensionale composto dalle tre dimensioni citate sopra. Il lessico presente all'interno di AFIIN è moderno(includendo perfino slang dell'Urban Dictionary) e alle parole viene assegnato un punteggio in un intervallo [-5, 5]. L'attuale versione contiene oltre 3.300 parole. [27]

Il secondo approccio, più emergente, utilizza tecniche di Machine Learning (ML) per costruire modelli complessi e prevedere probabilisticamente il sentimento di un determinato set di testi. Il linguaggio naturale è troppo creativo e complesso, l'espressione del sentimento è troppo sfumata per essere pienamente catturata da un lessico statico o da un elenco fisso di regole euristiche.[28] Sempre più spesso gli approcci di analisi del sentimento sfruttano il ML per costruire modelli più espressivi, tipicamente stimati/addestrati su un ampio train-set di testi, contenenti una mappatura tra espressioni testuali e valutazioni del sentimento assegnate da umani. Gli approcci tramite ML possono potenzialmente identificare la miriade di caratteristiche contestuali che contribuiscono al sentimento di un testo e apprendere automaticamente il sentimet attribuito a determinate parole o intere frasi. Lo svantaggio nell'uso di questa tecnica è la richiesta di grandi train-set "etichettati" che risultano spesso dispendiosi in termini di tempo e costo. I modelli considerati da Shapiro, Sudhof e Wilson adottano tecniche chiamate embeddings, ovvero la codificazione di parole o interi documenti in vettori al fine di incorporare le conoscenze esterne sulle parole, in alcuni casi compensando la scarsità di dati d'addestramento. Il primo modello si chiama GloVe(Gloval Vector for Word Representation), sviluppato dal gruppo NLP di Stanford composto da Pennington, Socher e

[26] S.Baccianella, A.Esuli, F.Sebastiani,(2010)
[27] https://review.gale.com/2023/08/22/understanding-recent-enhancements-to-sentiment-analysis-in-gale-digital-scholar-lab/
[28] Shapiro,Wilson e Sudhof(2020)

Manning nel 2014. Si tratta di un transfer learning model, ovvero un modello pre-addestrato su una task specifica e che ha appreso le relazioni contestuali tra le parole. Questo algoritmo di apprendimento è stato progettato per creare rappresentazioni vettoriali dense(embeddings) per parole che hanno un significato simile, memorizzandone informazioni semantiche e sintattiche.[29] Essendo pre-addestrato su un corpus di testo molto ampio GloVe apprende le relazioni tra parole basandosi sulla loro co-occorrenza nei testi[30]: se le parole "orribile" e "terribile" appaiono spesso in circostanze simili, i loro vettori saranno vicini nello spazio vettoriale del modello. Tuttavia, data la noncuranza di GloVe per l'ordine delle parole ed il contesto, i tre economisti hanno considerato l'utilizzo di un modello sviluppato più recentemente, noto come BERT(Bidirectional Encoder Representations from Transformers). Sviluppato da Devlin, Chang, Lee e Toutanova(Google) nel 2018 fornisce vettori predefiniti e fissi per le parole, generando embeddings consapevoli del contesto.[31] A differenza di GloVe, BERT non elabora il testo in una sola direzione ma utilizza un approccio bidirezionale, coinvolgendo l'analisi delle sequenza di testo sia da sinistra che da destra. L'innovazione di questo modello è stata la possibilità di apprendere rappresentazioni bidirezionali con i trasformatori, componente di deep learning che elabora l'intera sequenza di dati in parallelo, "definendo" le parole in relazione alle altre presenti in una frase. Quindi, a differenza delle RNN(Reti Neurali Ricorrenti), che elaborano i dati sequenzialmente(ogni passaggio dipende dal completamento del precedente) questo utilizzo dei trasformatori permette di conoscere il significato di una determinata parola derivandolo da tutti gli altri termini presenti nel segmento, lavorando contemporaneamente su tutta la sequenza.[32]

2.3 Adam H. Shapiro, Moritz Sudhof e Daniel Wilson: la creazione del nuovo sentiment index

L'obiettivo finale di Shapiro, Sudhof e Wilson era quello di applicare un modello di analisi dei sentimenti ad un insieme di articoli di notizie economiche al fine di costruire un indice di serie temporale che catturasse il sentimento espresso dai testi scritti. I dati grezzi presi per questo processo comprendevano articoli di giornale(editoriali compresi), dal 1980 al 2015, provenienti da 16 delle principali testate giornalistiche degli Stati Uniti. Il materiale scelto è stato comprato da un archivio di articoli di giornale del servizio Lexis-Nexis(LN) e doveva soddisfare 5 criteri:

[29] https://developers.google.com/machine-learning/crash-course/embeddings/video-lecture
[30] https://nlp.stanford.edu/projects/glove/
[31] Shapiro,Wilson e Sudhof(2020)
[32] https://www.nvidia.com/en-us/glossary/bert/

1) Il Paese di riferimento doveva essere “United States”;
2) L’argomento doveva essere “Economy” o “Economic”;
3) LN non doveva aver classificato gli articoli come “Brief” , “Summary” or “Digest“;
4) Gli articoli dovevano contenere almeno 200 parole;
5) L’articolo doveva contenere almeno una delle seguenti parole: “said”, “says”, “told”, “stated”, “wrote”,“reported”.

Queste restrizioni hanno permesso agli autori di filtrare gli articoli acquisiti concentrandosi solamente su quelli relativi alle news economiche degli USA(1 e 2), evitando duplicazioni di testi(3), rinunciando ad articoli troppo brevi(4) e incentrando il campione su testi che potessero contenere citazioni o parafrasi di intervistati. Questa selezione ha consentito di rilevare i testi con contenuti più “emotivi” e ha ridotto il campione a circa 238.685 articoli.

Inizialmente gli studiosi hanno valutato l’utilizzo di diversi tipi di modelli “lexical”, abbandonando l’ipotesi di utilizzo dei modelli “machine learning”, i quali hanno avuto risultati di molto peggiori rispetto ai primi, probabilmente per la dimensione limitata del training set. I modelli lessicali sono stati costruiti tramite il processo Bag-of-Words, usando lexicon diversi e calcolando il risultato netto(net positivity o positivity) dato dalla proporzione di parole positive meno la proporzione di parole negative. In questo processo sono stati considerati 3 lexicon popolari nell’analisi del sentimento:

1) **Harvard General Inquirer (GI) Dictionary**: è uno dei primi lexicon esistenti, pensato per la lingua generale inglese e formato da 3626 parole identificate come positive, negative o neutre, anche se quest’ultime non verranno trattate nel processo di Bag-of-Words;

2) **Loughran-McDonald (LM) lexicon**, **2014 updated version**: ha una forma ridotta, contenente 2707 parole che però danno un apporto chiave all’analisi economica, provenendo principalmente da relazioni annuali 10-K, ovvero documenti annuali che le aziende quotate in borsa negli Stati Uniti devono presentare alla Securities and Exchange Commission(SEC). Per questo la positività o negatività assegnata in questo dizionario fa riferimento al significato specifico che queste parole hanno nel campo della finanza;

3) **Hu and Liu’s lexicon (HL)**, **2004 version**: è stato costruito partendo da recensioni di film online a cui sono state assegnate punteggi di negatività/positività dai revisori stessi, creando quindi un lexicon ampio, che comprende 6786 parole, le quali tuttavia non sono specifiche per il dominio economico/ finanziario.

Come ci si poteva aspettare, i "dizionari" sopra indicati si sono "comportati" in modo diverso. Per quanto il lexicon LM sia il più piccolo come parole contenute ha registrato la più alta frazione di parole uniche(il 58% delle parole non si trovano negli altri due lexicon). Al contrario, GI era ben coperto dagli altri due(69% è presente sia in LM che HL) e HL aveva il vantaggio di possedere una grande frazione(57%) di parole uniche. Questi "lessici" hanno classificato insolitamente pochi unigrammi nel corpus di testo come positivi o negativi: il dizionario GI aveva la copertura più elevata etichettando solamente il 6,4% delle parole mentre, HL e LM, coprivano rispettivamente un misero 1,3% e 0,7%. Per quanto riguarda la concordanza invece, i lessici erano raramente in totale disaccordo sulla valenza delle parole e solo una piccola percentuale dei termini in comune risultava avere punteggi opposti. In generale nessun lessico era adatto a misurare i sentimenti che si potevano trovare nelle notizie. Il LM era appropriato per il dominio economico/finanziario ma non copriva bene il corpus mentre GI e HL, anche se includono molte parole, non erano utili per analizzare argomenti economici. Gli economisti hanno valutato dunque la possibilità di combinare i lessici per migliorare l'accuratezza del modello di valutazione, prendendo in considerazione anche l'aggiunta di un ulteriore vocabolario per migliorarne le performance.

Per valutare l'accuratezza dei modelli di sentiment, gli autori hanno utilizzato come test set un corpus di 800 articoli(selezionati casualmente dal corpus principale) etichettati manualmente da un gruppo di 15 assistenti di ricerca presso al Federal Reserve Bank di San Francisco. La valutazione avveniva su una scala di cinque punti, con l'obbiettivo di catturare il tono emotivo degli articoli piuttosto che il contenuto economico.

Per confrontare i modelli di sentiment invece, gli studiosi hanno calcolato la "positività netta" di ciascun articolo, utilizzando poi quattro metriche principali per la valutazione delle prestazioni:

- **Correlazione per ranghi di Spearman**: è una misura statistica non parametrica di correlazione basata sui ranghi(posizioni relative dei dati in un insieme ordinato). Diversamente da quella di Pearson, che misura la relazione lineare tra due variabili la correlazione di Spearman valuta la relazione monotona tra due variabili. Variando tra -1 e 1, un valore di 1 indica una perfetta correlazione monotona positiva, -1 una perfetta correlazione monotona negativa, e 0 nessuna correlazione monotona;[33]
- **R^2 da una regressione OLS**: misura quanto bene la retta di regressione OLS si adatti ai dati. È la frazione della varianza campionaria della variabile dipendente spiegata dai

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

regressori(o variabili esplicative). Solitamente il suo valore è prossimo a 1 quando i regressori predicono bene il valore della variabile dipendente; mentre se è uguale a 0 significa che non lo fanno;[34]

- **pseudo-R^2 da una regressione logit ordinata**: una regressione logit ordinata viene utilizzata quando la variabile dipendente è ordinata ma non continua(ad esempio valori su scala da 1 a 5, come il sentiment). Per questo tipo di regressioni, al posto del classico R^2 usato nelle regressioni OLS, viene usato questo pseudo-R^2 per valutare quanto il modello si adatti ai dati;[35]

- **statistica Macro-F1**: misura utilizzata per valutare le prestazioni di un modello di classificazione. Comunemente usata nella valutazione delle performance nella "NLP sentiment analysis literature", determina la precisione di un modello in termini della sua capacità di classificare correttamente il testo nelle categorie discrete di interesse, in questo caso "positive" , "neutral" o "negative".

L'analisi effettuata ha dimostrato come i modelli basati sui lessici Loughran-McDonald (LM) e Hu and Liu (HL) performavano meglio rispetto al General Inquirer (GI). In particolare, i modelli LM e HL mostravano correlazioni di Spearman intorno a 0.44, rispetto allo 0.27 registrato per il modello GI. I risultati indicano che i modelli LM e HL erano più efficaci nel catturare il sentiment degli articoli rispetto a GI, probabilmente per l'utilizzo di un lessico più ampio(HL) o per un dominio più specifico(LM). Successivamente, gli autori hanno combinato i lessici esistenti cercando di ampliare il numero di parole coperte, scoprendo che il modello che combina i lessici LM e HL produceva una previsione del sentiment più accurata rispetto alle altre combinazioni(compresa quella GI + LM + HL). Questo abbinamento permetteva dunque di coprire un maggior numero di parole e di sfruttare i punti di forza di ciascun lexicon, avendo una limitata sovrapposizione nella copertura lessicale.

Infine, per migliorare l'efficienza del modello i tre economisti hanno voluto ampliare il vocabolario utilizzato per predire il sentiment. Per riuscirci hanno diviso il processo in tre fasi:

1) Inizialmente hanno assegnato una classe di sentiment(positivo, neutrale o negativo) a ciascuna frase nel corpus di 238.685 articoli utilizzando il classificatore di frasi Vader(spiegato nel capitolo prima) modificato per includere i lessici LM e HL(i più

[34] James H. Stock, Mark W.Watson(2020)
[35] Hosmer, Lemeshow, Sturdivant(2013)

performanti), fornendo una misura inziale dell'orientamento del sentiment di ogni frase;

2) In seguito hanno creato una matrice che conta la co-occorrenza di ciascuna parola con le tre classi di sentiment, permettendo loro di raccogliere dati sulla frequenza con cui ogni termine appare in frasi con un determinato sentiment;

3) Infine hanno utilizzato la Pointwise Mutual Information(PMI)[36] per ripesare questa matrice, calcolando il grado di associazione parola-classe e quantificandone la relazione. Questo ha dato vita ad un nuovo lexicon utilizzabile per calcolare il punteggio di sentiment in ciascun articolo, facendo una media dei punteggi delle parole contenute in esso e garantendo una copertura più completa per il dominio economico-finanziario.

I risultati ottenuti hanno delineato la strada da intraprendere. Integrando il nuovo lessico con i precedenti(LM + HL) e aggiungendo una regola di negazione(Vader) il modello ottenuto ha migliorato ulteriormente le performance predittive. Per valutare la significatività statistica invece, gli autori hanno utilizzato il test di Diebold-Mariano[37] che ha confermato come il modello migliorato avesse errori di previsione significativamente inferiori rispetto ai modelli a confronto.

Il "best model" è stato poi sfruttato per creare un indice temporale, calcolando i punteggi di sentiment per un vasto insieme di articoli economici e finanziari datati dal 1980 e aggregando questi punteggi in index giornalieri e mensili. L'indice mensile ha mostrato una forte correlazione con il ciclo economico e gli eventi economici chiave, oltre a correlarsi con indici di sentiment dei consumatori basati su sondaggi. Questo suggerisce che l'indice di sentiment delle notizie abbia un alto rapporto segnale-rumore(in inglese, Signal-to-Noise Ratio, SNR), rendendolo un utile strumento per analizzare le tendenze economiche.

[36] **PMI**: misura utilizzata in linguistica computazionale e teoria dell'informazione per quantificare l'associazione tra due eventi(come due parole che co-occorrono in un corpus di testo).

[37] **Test Diebold-Mariano**: procedura statistica utilizzata per confrontare la precisione predittiva di due modelli di previsione, valutando se le differenze nei valori di errore di previsione sono statisticamente significative

3. Correlazione del sentiment index con eventi storici e con i principali indicatori economici

3.1 Introduzione

L'ultimo capitolo di questa tesi è incentrato sull'analisi della possibile correlazione tra gli indicatori di sentiment creati da Shapiro, Wilson e Sudhof e alcuni dei principali indicatori economici presenti al giorno d'oggi. L'obiettivo è constatare come la variabilità delle emozioni pubbliche possano influenzare l'andamento economico di un paese, fornendoci insight importanti per valutare e prevedere la formazione di cicli economici. I sentiment index sono oramai strumenti essenziali per valutare e comprendere le percezioni, emozioni e valutazioni di consumatori ed investitori, offrendo una panoramica importante sulle aspettative future generate da notizie ed opinioni attuali. Nel concreto, un aumento del valore di un indice di fiducia può suggerire che i consumatori siano propensi alla spesa e all'investimento, stimolando quindi l'economia del Paese; al contrario, un aumento del valore di un indice di sfiducia può anticipare crisi economiche, segnalando una crescente prudenza o addirittura paura nella popolazione. Da sempre i cambiamenti emotivi condizionano le decisioni economiche, andando ad incidere sulla fiducia ed in particolar modo sulle strategie e le politiche da noi messe in atto, producendo cicli di panico o euforia che esulano dai comportamenti razionali che bisognerebbe mantenere.

Nella Figura 2 qui sotto, presa dalla pubblicazione di Shapiro e Wilson sul sito della Federal Reserve Bank of San Francisco, possiamo vedere, oltre alla correlazione tra le serie storiche di due indici da loro creati, anche come le percezioni negative reagiscano a rilevanti eventi economici, politici e sociali. Negli anni '80 l'indice di negatività ha registrato picchi evidenti, soprattutto durante il periodo 1980-1982, corrispondente alla recessione causata in buona parte dalla seconda crisi petrolifera del 1979. Nel decennio successivo si possono osservare nuove fluttuazioni dell'indice in risposta alle recessioni dei primi anni '90(dovute in parte alla guerra del Golfo) e, nonostante il periodo di crescita economica e tecnologica, esso è aumentato in occasione della crisi finanziaria russa del 1998. Successivamente, negli anni 2000, vengono rilevati aumenti molto importanti della negatività pubblica, causati dall'attentato alle torri gemelle nel 2001 e dall'invasione dell'Iraq nel 2003, entrambi eventi che hanno ridotto la fiducia dei consumatori e aumentato il debito pubblico, impattando significativamente sull'economia americana. Tuttavia, durante la crisi finanziaria del 2007-2008, uno degli eventi più significativi di quel periodo, l'indice non ha registrato picchi acuti di negatività come nel 2001 e 2003. Perché? Probabilmente la risposta si cela nelle caratteristiche degli eventi che

stiamo analizzando. L'11 settembre e l'invasione dell'Iraq sono stati eventi molto più "immediati", con un impatto diretto sulle emozioni del pubblico, riflettendo nel Negativity index il panico e l'incertezza di quel periodo. La crisi finanziaria del 2008(Lehman bankruptcy) è stato un avvenimento che, per la sua complessità, non ha permesso alla gente comune un'immediata comprensione delle sue implicazioni. Il pessimismo è aumentato gradualmente, man mano che gli effetti della crisi si propagavano nei mercati e nell'economia reale. Infine, nel quinquennio 2010-2015 l'index ha mostrato una tendenza al ribasso, con aumenti presumibilmente causati dalla crisi del debito sovrano europeo(iniziata nel 2010) e dalle numerose tensioni geopolitiche in medio oriente.

In sintesi, il Negativity Index si è rivelato un potente mezzo per monitorare le emozioni collettive in risposta ad eventi economici e politici che si sono verificati tra il 1980 e il 2015, mostrando come le opinioni negative emergano e si rafforzino durante crisi economiche, recessioni ed incertezza.

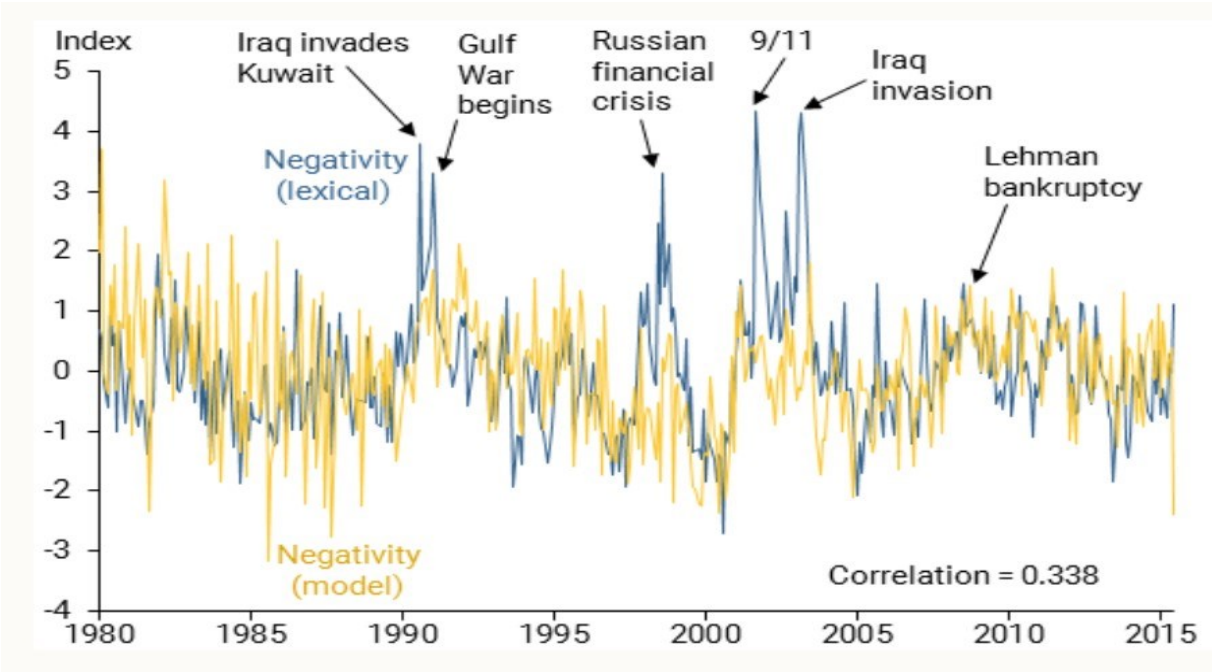

Figura-2

3.2 Rstudio: correlazione tra indici

Data la forte relazione tra i sentiment index e i principali eventi storici che tra il 1980 e il 2015 hanno fortemente influenzato l'economia americana, ho deciso di effettuare un'ulteriore analisi, esplorando la correlazione tra gli indici sviluppati da Shapiro, Wilson e Sudhof e vari indicatori

economici, prendendo spunto dal loro lavoro e cercando di comprendere come il sentiment espresso nelle notizie influenzi o sia influenzato dalle principali variabili macroeconomiche.

I quattro sentiment index che ho usato per questa analisi, creati tramite Kanjoya's model[38], rappresentano tre tipi di emozioni presenti nel corpus text analizzato e variano nel tempo:

- **Negativity(Lexical)** e **Negativity(Model)**: questi due indici misurano il livello di negatività nei testi analizzati;
- **Worried(Model):** misura il livello di preoccupazione espresso nei testi, si concentra su parole o frasi che indicano incertezza o timore riguardo ad eventi economici o finanziari;
- **Satisfied(Model):** misura il livello di soddisfazione, rilevando parole e frasi che indicano sentimenti positivi o ottimistici.

Gli indicatori economici che ho selezionato sono variabili fondamentali per catturare le performance macroeconomiche e finanziarie del Paese, essendo fortemente correlati ad oscillazioni dell'attività economica nel tempo(cicli economici), rappresentano una visione complessiva della "salute economica" degli Stati Uniti nel periodo che va dal 1980 al 2015. In particolare[39]:

- **S&P500**: è un indice che rappresenta le 500 maggiori società quotate negli Stati Uniti, ampliamente utilizzato come parametro delle performance del mercato azionario;
- **All Employees(PAYEMS)**: comunemente noto come Total Nonfarm Payroll è una misura del numero di lavoratori nell'economia degli Stati Uniti, escludendo proprietari, dipendenti domestici privati, volontari non retribuiti, lavoratori agricoli e lavoratori autonomi non incorporati. Questa misura rappresenta circa l'80 per cento dei lavoratori che contribuiscono al prodotto interno lordo (PIL);
- **Industrial production(IP, INDPRO)**: è la misura dell'output del settore industriale, riflettendo la domanda di beni e servizi nell'economia è un indicatore chiave per analizzare i cicli economici;
- **Personal Consumption Expenditures(PCEPI)**: è una misura del cambiamento nei prezzi dei beni e servizi consumati. Le sue variazioni sono note per catturare l'inflazione(o deflazione) riflettendo cosi i cambiamenti nel comportamento dei consumatori;

[38] Kanjoya's model: il modello calcola la probabilità che l'articolo esprima un sentimento particolare in base alle parole presenti in esso, creando un sentiment score che varia nel tempo.
[39] https://fred.stlouisfed.org/

- **Federal Funds Effective Rate(FEDFUNDS)**: è il tasso di interesse a breve termine con cui le banche statunitensi si prestano denaro tra loro per soddisfare requisiti di riserva imposti dalla Federal Reserve(FED). È uno dei principali strumenti di politica monetaria utilizzato dalla FED per controllare l'inflazione e stabilizzare l'economia;
- **Unemployment Rate(UNRATE)**: è il tasso di disoccupazione che misura la percentuale della forza lavoro disoccupata e attivamente in cerca di un'occupazione, fondamentale per misurare la salute del mercato del lavoro.
- **Consumer Price Index for All Urban Consumers(CPIAUCSL)**: è un indice dei prezzi al consumo per tutti i consumatori urbani. Misura la variazione media dei prezzi pagati dai consumatori urbani per un paniere fisso di beni e servizi. Anche questo è uno degli indicatori più comuni per rappresentare l'inflazione.

Per analizzare la correlazione tra questi indici, di sentiment ed economici, ho usato R, un linguaggio di programmazione utilizzato principalmente per l'analisi statistica e grafica. Sviluppato originariamente dagli statistici Ross Ihaka e Robert Gentleman è diventato uno strumento fondamentale per i data scientist e gli analisti, grazie alle sue capacità di elaborazione dati e alle sue librerie estensive. Il linguaggio R è stato utilizzato nell'ambiente di sviluppo integrato(IDE) Rstudio, appositamente creato per fornire un'interfaccia utente e numerosi strumenti che semplificano la scrittura di codice e l'analisi delle informazioni. R e Rstudio sono utilizzati in una vasta gamma di settori, dall'analisi finanziaria alla bioinformatica, grazie alla flessibilità e potenza che caratterizza la loro combinazione.

Prima di tutto ho installato e "caricato" i pacchetti necessari per eseguire la correlazione. Secondariamente ho effettuato il download, dal sito della Federal Reserve Bank of San Francisco, della serie storica dell'indice di sentimento(in formato Excel), scaricando poi tramite la funzione "getSymbols" i dati storici relativi ai vari indicatori economici descritti qui sopra. Per quest'ultime variabili ho utilizzato come fonte Yahoo Finance(solo per il S&P500) e FRED(Federal Reserve Economic Data), un database gestito dalla divisione Ricerca della Federal Reserve Bank of St.Louis che contiene più di 800.000 serie temporali economiche compilate dalla Federal Reserve, di cui molte raccolte da agenzie governative come U.S. Census e il Bureau of Labor Statistics. Successivamente ho filtrato i dati storici degli indici economici per ottenere delle serie uniformi a quelle dei sentiment indexes, partendo dal primo Gennaio 1980 fino ad arrivare al 30 giugno 2015, ultimo mese di cui si hanno i valori dell'index di Shapiro. In questa fase, avendo scaricato la serie storica dello S&P500 con dati giornalieri(non mensili come dal sito FRED) ho dovuto calcolare i rendimenti mensili dell'indice del mercato

azionario, standardizzandoli per renderli una misura più comparabile ed intuitiva tramite la formula $\frac{Last\ of\ the\ month(x) - First\ of\ the\ month(x)}{First\ of\ the\ month(x)}$.

Nello step successivo ho preparato ed integrato tutti i dati all'interno di un unico dataset, allineando le date dei vari indicatori ed eliminando eventuali informazioni mancanti, processo comunemente chiamato "dataset cleaning". Infine, tramite l'utilizzo del pacchetto e della funzione "corrplot" sono riuscito a calcolare e visualizzare la matrice contenente i valori di correlazione tra gli indici di sentiment e quelli economici. Il risultato ottenuto si può osservare qui sotto in Figura 3.

Per i valori all'interno della matrice è stato utilizzato l'indice di correlazione lineare di Bravais e Pearson, sviluppato da Karl Pearson nel 1880, la formula matematica fu derivata e pubblicata da Auguste Bravais nel 1844. Questo indice esprime un'eventuale relazione di linearità tra due variabili e ha un valore compreso tra -1 e 1, dove 1 corrisponde alla perfetta correlazione lineare, 0 ad un assenza di correlazione e -1 ad una perfetta correlazione lineare negativa. La formula utilizzata per ottenere il coefficiente è $\rho XY = \frac{\sigma XY}{\sigma X \sigma Y}$, definita come la covarianza di due variabili statistiche X ed Y divisa per il prodotto delle loro deviazioni standard.

Figura-3

3.3 I risultati ottenuti dall'analisi

Il Negativity index(lexical) presenta una correlazione negativa(-0.13) con lo S&P500, suggerendo che un aumento della negatività possa essere associato ad una diminuzione del prezzo delle azioni. Questo è credibile data la possibile riduzione della fiducia degli investitori derivante da notizie negative. Una correlazione negativa simile(-0.12) è quella osservata con il Federal Funds Rate, che potrebbe riflettere una riduzione dei tassi di interesse per stimolare l'economia in momenti di pessimismo. Tuttavia, la correlazione risulta bassa, indicando che molti altri fattori influenzano queste variabili economiche e la negatività lessicale è verosimilmente solo uno dei tanti elementi.

Il Negativity index(model) ha invece una correlazione negativa(-0.20) relativamente forte con il Total Nonfarm Payroll(PAYEMS), suggerendo che una maggiore negatività nei testi possa essere associata ad una diminuzione dell'occupazione. Questo viene confermato dalla correlazione positiva(0.38)con il tasso di disoccupazione(UNRATE), affermando l'idea che le notizie negative possano segnalare o riflettere condizioni economiche difficili. Infine, utile menzionare una correlazione negativa(-0.19) con la produzione industriale(INDPRO) mettendo in evidenza una probabile domanda più debole o una riduzione nella fiducia delle prospettive economiche in periodi sfavorevoli.

L'indice Worried(Model) invece è correlato positivamente con l'occupazione(0.14) e la produzione industriale(0.12), suggerendo che l'aumento di preoccupazione possa andare di pari passo con miglioramenti economici. Sembrerebbe un controsenso, ma potrebbe riflettere le preoccupazioni per un'economia in espansione che sta per "surriscaldarsi".

L'indice Satisfied invece presenta correlazioni relativamente forti con PAYEMS(0.32), e INDPRO(0.31), indicando che una maggiore soddisfazione nei testi è associata a miglioramenti nelle condizioni economiche. Questa affermazione è coerente con l'idea che un sentiment positivo rifletta un'economia in crescita, con aumenti nell'occupazione e nella produzione industriale. Inoltre una correlazione negativa è stata registrata con il FEDFUNDS(-0.38), suggerendo che periodi positivi possano essere collegati a tassi di interesse più bassi, che rispecchiano una stimolazione monetaria dell'economia.

Le correlazioni osservate nella matrice propongono relazioni interessanti tra gli indici di sentiment e gli indicatori economici, con negatività, preoccupazione e soddisfazione che rivelano relazioni significative con le varie componenti dell'economia. Ciononostante, la complessità di queste dinamiche richiederebbe metodi di analisi più avanzati per comprendere al meglio queste relazioni. Il coefficiente di Pearson, seppur quantificando la forza e la direzione

della relazione lineare tra due variabili, presenta delle limitazioni per l'analisi delle relazioni causa-effetto, che potrebbero essere superate utilizzando modelli VAR(Vector Autoregression) e studiando le funzioni di risposta all'impulso, presentando una visione più accurata e completa delle interazioni tra queste due tipologie di variabili nel tempo.

Infine, per una visualizzazione più intuitiva e dettagliata degli indici a confronto, ho deciso di creare i grafici delle loro serie temporali, confrontando i sentiment indexes con le variabili economiche a cui sono più correlati. I grafici permettono di visualizzare chiaramente come questi indicatori si comportino nel tempo e come le loro tendenze possano influenzarsi reciprocamente. Per realizzare questa parte di "Data visualization" ho utilizzato Tableau, un potente strumento che consente di sfruttare meglio i dati tramite una rappresentazione visiva di informazioni, rendendole facilmente comprensibili al pubblico.[40] Tableau permette quindi di creare grafici interattivi e dinamici, che facilitano l'analisi delle relazioni e consentono la sovrapposizione di più linee temporali, applicando filtri e colori per distinguere facilmente i vari indici.

Per rendere i grafici confrontabili sulla stessa scala, ho applicato una normalizzazione dei dati, ridimensionandoli in modo da averli compresi nell'intervallo [0,1], pur avendo variabili con unità di misura e scale differenti. Successivamente ho raggruppato le informazioni di sentimento ed economiche per annata(dal 1980 al 2015) calcolandone la media, ottenendo una rappresentazione grafica più semplice ed intuitiva tramite le loro linee temporali. I 4 grafici proposti nelle due pagine seguenti, uno per ogni sentiment index, confermano molte delle correlazioni osservate nella matrice, raffigurandole efficacemente.

[40] https://www.tableau.com/it-it

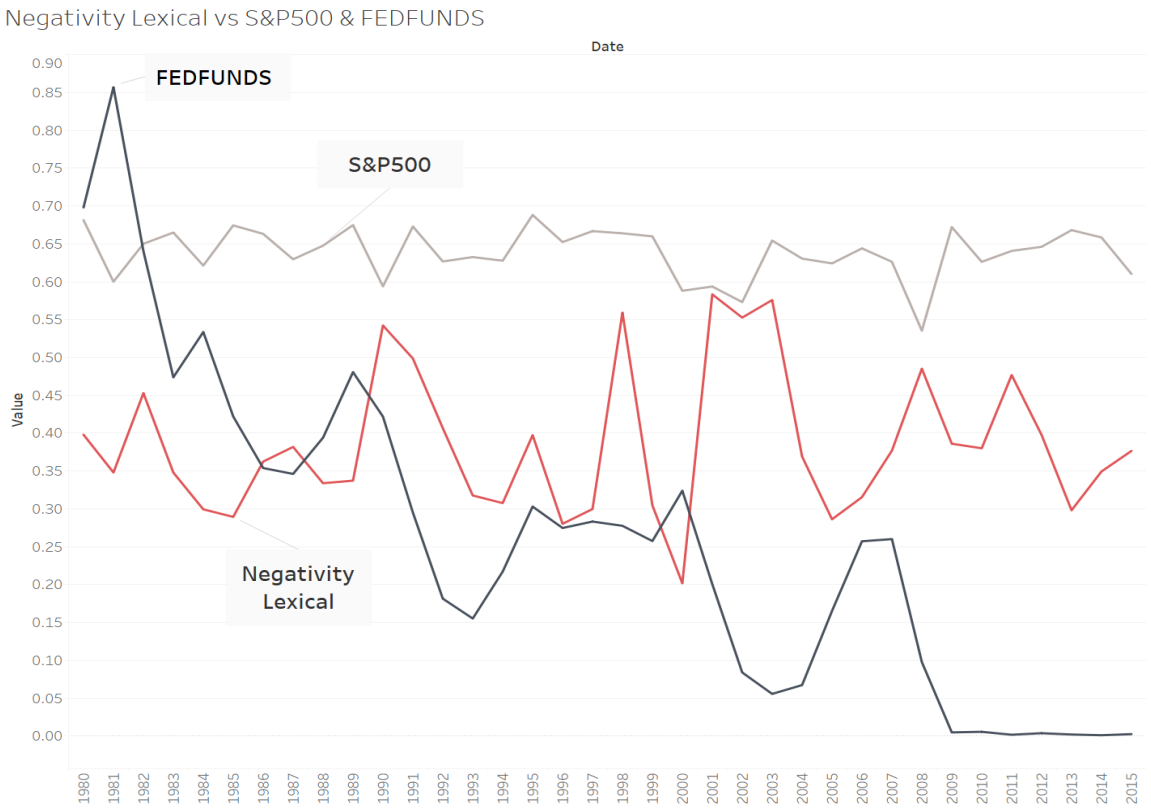

Grafico-1

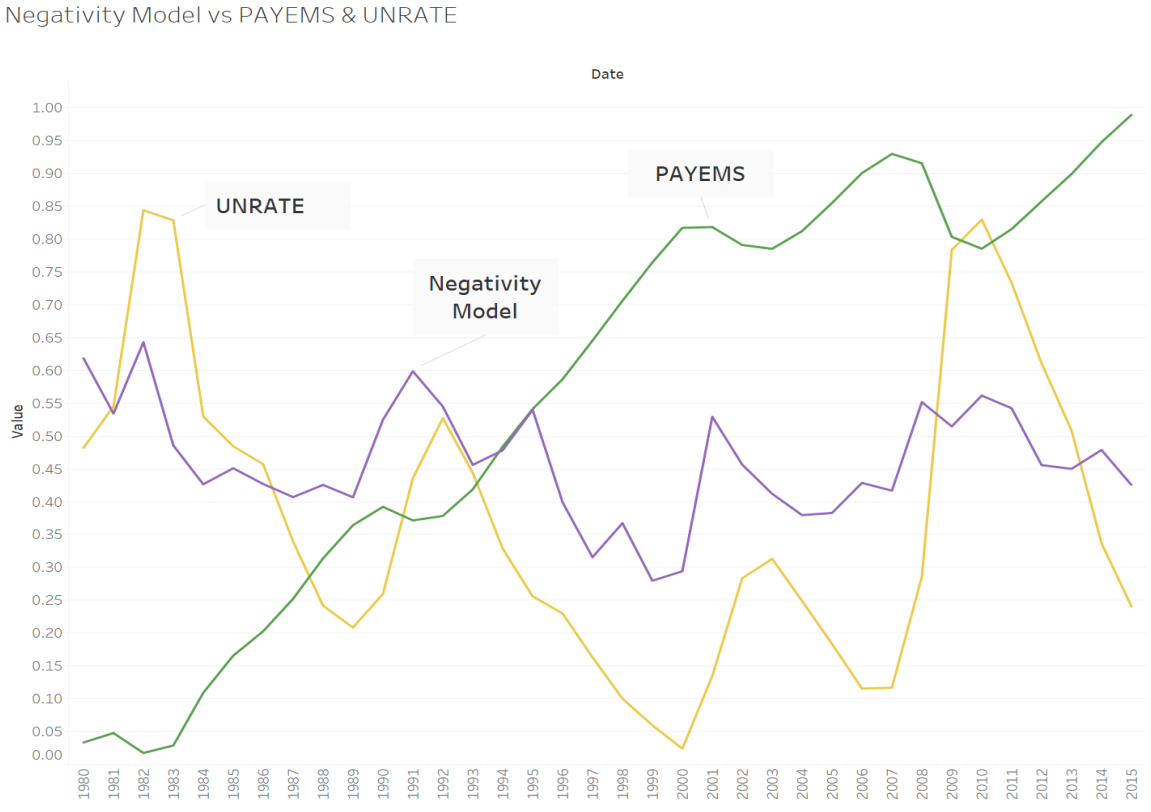

Grafico-2

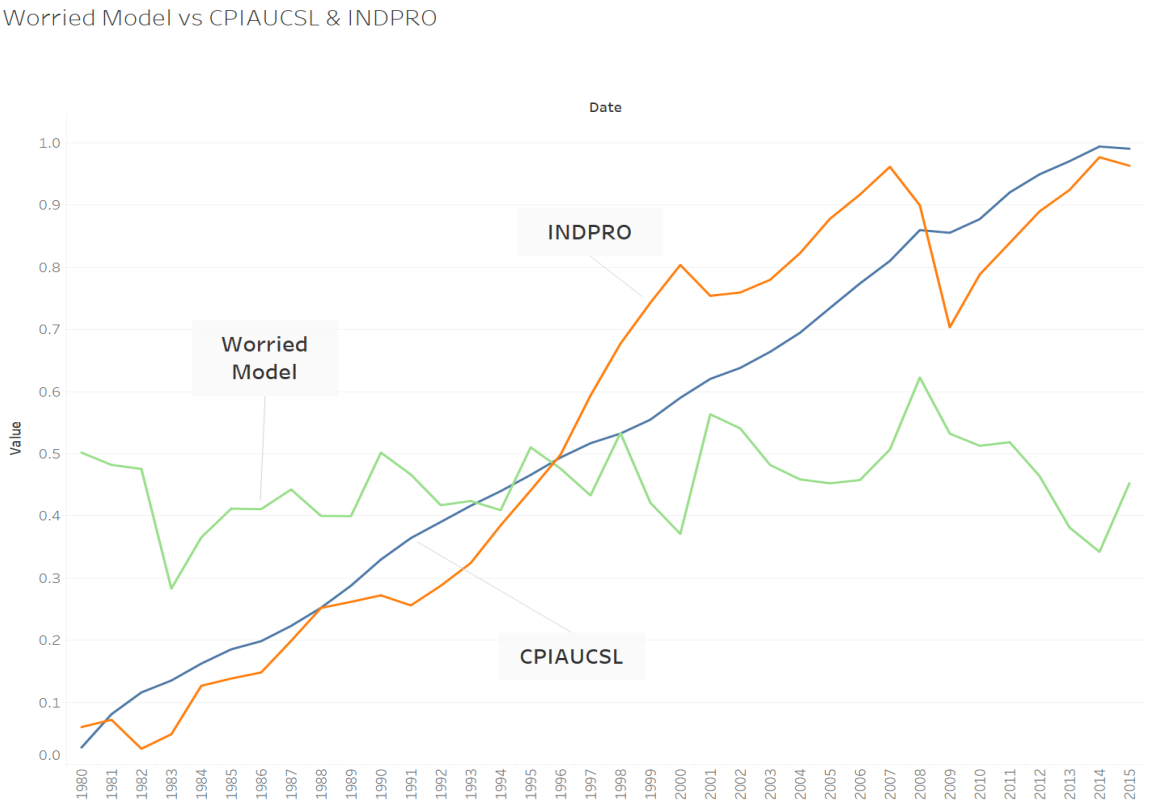

Grafico-3

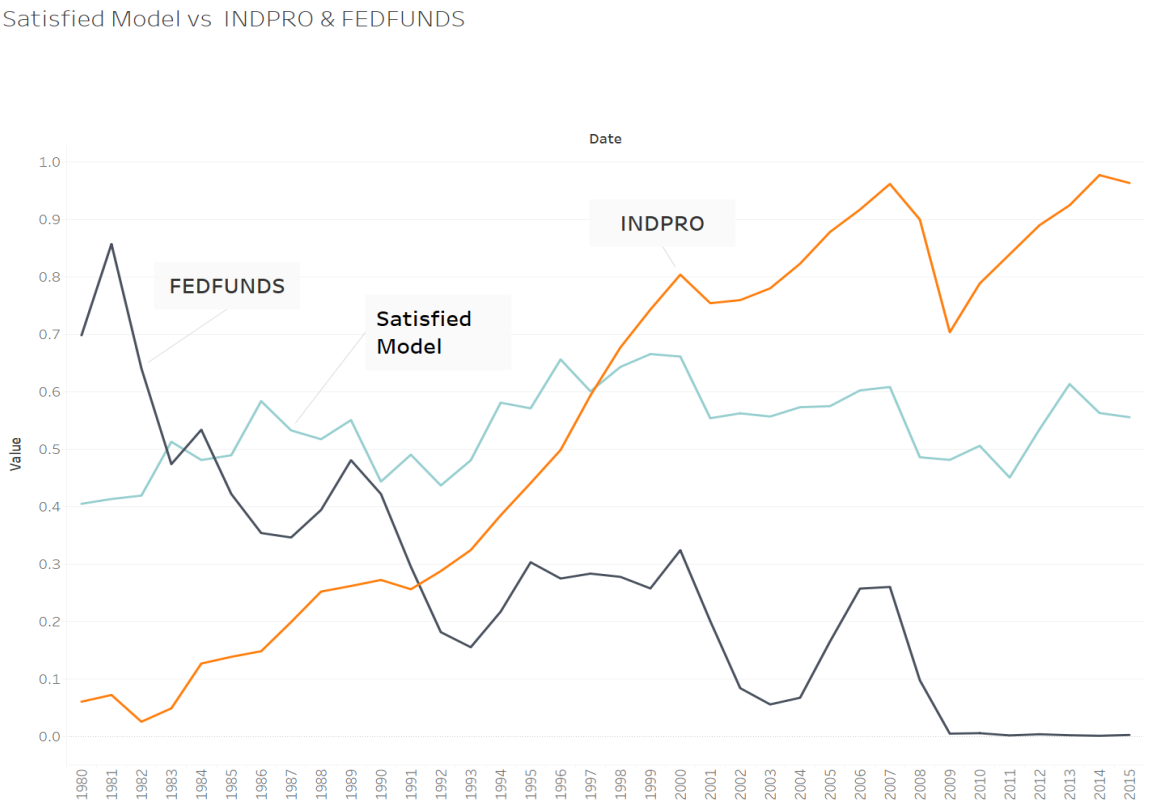

Grafico-4

4. Conclusione

Da sempre emozioni ed informazioni hanno un impatto cruciale sulla nostra vita quotidiana. John Maynard Keynes, padre della macroeconomia, con il termine “animal spirits” sottolinea come le scelte e le decisioni economiche degli investitori siano spesso istintive e basate su un’emozionalità che trascende la razionalità “esaltata” dall’economia tradizionale. Gli spiriti animali, che Keynes menziona nel suo libro "The General Theory of Employment, Interest and Money"(1936), rappresentano infatti le emozioni di sfiducia, speranza, paura e pessimismo che influenzano il processo decisionale in ambito finanziario.[41] Anche l’economista e premio Nobel(2013) Robert Shiller, intervistato da poco al Festival dell’Economia di Trento, ha detto “Da sempre i comportamenti umani sono fortemente condizionati dalle emozioni o dalle narrazioni, dalle storie raccontate da altre persone, su determinati temi” proseguendo con “Il progresso tecnologico ha sempre avuto un ruolo importante. Oggi con le nuove tecnologie le narrazioni, vere o false, riescono a diventare virali e riescono ad avere un impatto enorme sulle nostre vite, sulla società e sui mercati, determinandone di fatto il loro andamento”.[42] Questo suggerisce come, sia in passato che ora, gli economisti diano molta importanza alle emozioni, che possono modellare il comportamento di consumatori, investitori ed istituzioni, influenzando o riflettendo i cambiamenti economici nel mondo. Shiller inoltre riprende i concetti espressi in questa tesi, poiché in un era digitale come la nostra le informazioni, soprattutto quelle non convenzionali, sono diffuse e presenti in una moltitudine di “canali”. Tra social media, giornali e TV news il continuo flusso di dati ed aggiornamenti riflette il sentiment della società, rendendoci partecipi delle percezioni pubbliche che possono influenzare significativamente le nostre vite. Per questo motivo ho deciso di dare centralità al tema dell’analisi testuale, essenziale per comprendere informazioni ed emozioni provenienti dalla nostra società ed in grado di trasformarle in strumenti che ci permettano di effettuare decisioni informate, poiché in questo mondo moderno, più complesso che mai, notizie vere o false possono provocare panico o euforia tra la gente, impattando fortemente sull’economia e sul benessere collettivo.

[41] https://www.milanofinanza.it/investimenti-trading/gli-animal-spirits-di-keynes-202210170833014280
[42] https://www.ildolomiti.it/societa/2023/il-premio-nobel-shiller-le-paure-e-le-emozioni-influenzano-gli-investimenti-le-narrazioni-sono-una-delle-forze-che-guidano-leconomia

BIBLIOGRAFIA e SITOGRAFIA

NAWSHER KHAN, IBRAR YAQOOB, IHRAHIM ABAKER TARGIO HASHEM e ZAKIRA INYAT, 2014, The Scientific World Journal, *Review Article Big Data: Survey, Technologies, Opportunities, and Challenges*

DAWN E. HOLMES, 2017, Oxford University Press, *Big data: a very short introduction*

Osservatori Digital Innovation, Politecnico di Milano, *Big Data e Business Analytics, cosa sono e come sfruttarli,* https://blog.osservatori.net/it_it/big-data-cosa-sono

Artera*, "Caratteristiche dei Big Data: Volumi, Velocità, Varietà, Veridicità e Variabilità."*, https://www.artera.net/it/data-science/caratteristiche-big-data/

EUROSTAT REVIEW ON NATIONAL ACCOUNTS AND MACROECONOMIC INDICATORS, 2017

UNECE, *Classification of Types of Big Data*, https://statswiki.unece.org/display/bigdata2/Classification+of+Types+of+Big+Data

JOSÉ MARÍA LIBERTIAND E MITCHELL A. PETERSEN, 2018,National Bureau of Economic Research , *Information: Hard and Soft*

Rep order Management, 2023, https://www.reportdermanagement.com/blog/hard-data-vs-soft-data/

ADAM SHAPIRO, DANIEL WILSON E MORIZ SUDHOF, 2020, *Measuring news sentiment*

SERGIO BOLASCO, 2005, *Statistica testuale e text mining*

VIJAY KOTU AND BALA DESHPANDE,2018, *Data Science: Concepts and Practice*

IBM, *Text Mining: Concepts, Process and Applications*, https://www.ibm.com/it-it/topics/text-mining

LOKESH KUMAR, PARUL KALRA BHATIA, 2013, *Text Mining: Concepts, Process and Applications*

NIKITA BACHANI, Medium, *Towards Data Science: Chunking in NLP Decoded*, https://towardsdatascience.com/chunking-in-nlp-decoded-b4a71b2b4e24

MD KHALEEL AHAMED ,Medium, *The Essential Role of Syntactic and Semantic Parsing in NLP*, ,https://medium.com/@datailm/the-essential-role-of-syntactic-and-semantic-parsing-in-nlp-47b92118d9de

WIKIPEDIA, *Part-of-speech tagging*, https://en.wikipedia.org/wiki/Part-of-speech_tagging#:~:text=In%20corpus%20linguistics%2C%20part%2Dof,its%20definition%20and%20its%20context

WIKIPEDIA, *Tf-idf* , https://it.wikipedia.org/wiki/Tf-idf

MIKA V. MÄNTYLÄ , DANIEL GRAZIOTIN, MIIKKA KUUTILA, 2016, The Evolution of Sentiment Analysis - A Review of Research Topics, Venues, and Top Cited Papers

SIMONE BORGHI, Finanza Online, 2023, *Fear & Greed Index: avidità domina sui mercati ma forse è troppa. Segnale di possibile inversione*, https://www.finanzaonline.com/notizie/fear-greed-index-troppa-avidita-sui-mercati-possibile-inversione

ALBERTO RIVA e MIRANDA MARQUIT, Forbes, 2023, *VIX: cos'è e come funziona l'indice di volatilità*, https://www.forbes.com/advisor/it/investire/vix-indice-volatilita/

Il Sole 24 ORE, 2019, *Istat "misura" l'economia su Twitter. Come funziona il Social mood index*, https://www.infodata.ilsole24ore.com/2019/02/23/istat-misura-leconomia-twitter-funziona-social-mood-index/

Il Sole 24 ORE, 2019*, La recessione, il governo Conte e il social mood degli italiani*, https://www.infodata.ilsole24ore.com/2019/11/24/la-recessione-governo-conte-social-mood-degli-italiani/

BO PANG , LILLIAN LEE, Now Pub, 2008, *Opinion Mining and Sentiment Analysis*

Istituto di Teoria e Tecniche dell'Informazione Giuridica, *WordNet*, https://www.ittig.cnr.it/Ricerca/materiali/JurWordNet/WordNet.htm

STEFANO BACCIANELLA, ANDREA ESULI E FABRIZIO SEBASTIANI, 2010, *SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining*

SARAH L. KETCHLEY, The Gale Review, 2023, *Understanding Recent Enhancements to Sentiment Analysis in Gale Digital Scholar Lab*, https://review.gale.com/2023/08/22/understanding-recent-enhancements-to-sentiment-analysis-in-gale-digital-scholar-lab/

GoogleDevelopers, *Embeddings*, Google Machine Learning Crash Course, https://developers.google.com/machine-learning/crash-course/embeddings/video-lecture

Stanford NLP Group, *GloVe: Global Vectors for Word Representation* ,Stanford University https://nlp.stanford.edu/projects/glove/

NVIDIA, *BERT*, https://www.nvidia.com/en-us/glossary/bert/

WIKIPEDIA, *Coefficiente di correlazione per ranghi di Spearman*, https://it.wikipedia.org/wiki/Coefficiente_di_correlazione_per_ranghi_di_Spearman

JAMES H. STOCK, MARK W. WATSON, 2020, Pearson, *Introduzione all'econometria 5a edizione*

HOSMER, D. W., LEMESHOW, S., & STURDIVANT, R. X., John Wiley & Sons, 2013, *Applied Logistic Regression.*

Federal Reserve Bank of St. Louis, FRED Economic Data, https://fred.stlouisfed.org/

GIANLUCA DEFENDI, MILANO FINANZA, 2022, *Gli animal spirits di Keynes*, https://www.milanofinanza.it/investimenti-trading/gli-animal-spirits-di-keynes-202210170833014280

Il Dolomiti,2023, *Il premio Nobel Shiller:"Le paure e le emozioni influenzano gli investimenti, le narrazioni sono una delle forze che guidano l'economia",* https://www.ildolomiti.it/societa/2023/il-premio-nobel-shiller-le-paure-e-le-emozioni-influenzano-gli-investimenti-le-narrazioni-sono-una-delle-forze-che-guidano-leconomia

Federal Reserve Bank of San Francisco, sito: https://www.frbsf.org/

NUMERO DI PAROLE UTILIZZATE: 9.993 parole utilizzate(escluso frontespizio, indice e bibliografia)