TRENDING TOPIC E SENTIMENT ANALYSIS NEL MERCATO IMMOBILIARE. UN'APPLICAZIONE SUI TITOLI DEGLI ARTICOLI DI GIORNALI DI SETTORE[1]

DI ERIKA GHIRALDO, IRENE CESARANO**, LUIGI CIPOLLINI****

1 Introduzione

In questi ultimi anni, l'utilizzo di dati di fonte non strutturata per l'analisi economica è sempre più frequente. Ciò è imputabile principalmente alla disponibilità dei *big data*, informazioni con livelli elevati in termini di dimensioni, di velocità di aggiornamento e di valore, e ai miglioramenti nelle tecniche che consentono l'accesso e l'elaborazione di questi dati. In questo contributo, viene utilizzata una tipologia di dato non strutturato, il testo, al fine di mostrare come l'analisi di questi ulteriori "dati" possa fornire un utile contributo alla comprensione di alcuni fenomeni economici. Delimitando il perimetro al mercato immobiliare, si mostra come l'elaborazione delle informazioni tratte dagli articoli di settore possa contribuire a far emergere i temi di maggior interesse per il settore stesso in un certo periodo e, inoltre, come dal testo si possa giungere ad un indicatore utile per comprendere il clima di fiducia in un dato momento e, ancora, dedurre le emozioni che emergono, ovvero le emozioni che influenzano la nostra percezione della realtà.

Per l'analisi sono raccolti, attraverso metodologie di *web scraping*, dati testuali da diverse fonti. I dati testuali estrapolati sono i titoli e le brevi descrizioni di articoli pubblicati su riviste dedicate al settore immobiliare e in sezioni specifiche dei quotidiani nazionali. Con lo scopo di effettuare da un lato, l'analisi del testo utile ad individuare i temi di tendenza "*trending topic"* di un certo periodo e dall'altro costruire un indice del *sentiment,* è stata creata una base dati di titoli e descrizioni degli articoli. L'impiego di dati testuali, in particolare di testo, tratto da articoli di quotidiani e/o riviste, è già ampiamente presente nella letteratura economica, a titolo di esempio si richiama il contributo italiano di Apriliano et al. (2021) e il lavoro di Barbaglia et al. (2022) nel contesto europeo. Nell'ambito del mercato immobiliare, l'uso di dati testuali per la costruzione di indicatori di fiducia risulta crescente negli ultimi anni; tra gli altri si rimanda alle proposte di Clive (2016), Soo (2018) e Zhu et al. (2022), ma non si è venuti a conoscenza di un'applicazione per il contesto italiano. In questo lavoro, si intendono esplorare le potenzialità di un'analisi di questo tipo per il mercato immobiliare italiano.

Dopo la raccolta delle informazioni, a valle di una fase di trattamento preliminare del dato, per renderlo adatto all'elaborazione, l'analisi del testo restituisce oltre all'individuazione dei temi di maggior tendenza anche una fotografia delle connessioni tra i temi trattati negli articoli.

Per *la sentiment analysis* si è fatto ricorso al dizionario già disponibile dei termini in lingua italiana, integrato da alcune parole che appaiano frequente nei testi analizzati, ma non presenti nel dizionario. Si tratta nella maggior parte dei casi di vocaboli specifici del settore immobiliare ai quali non è associato alcun punteggio, ma che evidentemente non possono essere trascurati nell'analisi che qui si conduce. Per queste parole si è assegnato il punteggio di una parola presente nel dizionario che avesse un significato analogo.

Al termine dell'elaborazione si è costruito un indicatore che fornisce giornalmente un valore della fiducia estratta dagli articoli che riportano le principali notizie del mercato immobiliare italiano.

* CAPO UFFICIO STATISTICHE E STUDI SUL MERCATO IMMOBILIARE - DIREZIONE CENTRALE SERVIZI ESTIMATIVI E OSSERVATORIO MERCATO IMMOBILIARE.
** FUNZIONARIA UFFICIO STATISTICHE E STUDI SUL MERCATO IMMOBILIARE - DIREZIONE CENTRALE SERVIZI ESTIMATIVI E OSSERVATORIO MERCATO IMMOBILIARE.
*** FUNZIONARIO DIREZIONE GENERALE DI STATISTICA E ANALISI ORGANIZZATIVA - MINISTERO DELLA GIUSTIZIA

[1] L'articolo è una sintesi di un lavoro di ricerca svolto nell'ambito del corso di "Diploma di esperto in Data Science" organizzato dalla Scuola Nazionale dell'Amministrazione che gli autori hanno frequentato e conseguito il relativo diploma nel corso del 2023.

Il lavoro è strutturato in un primo paragrafo introduttivo, un secondo paragrafo dedicato ai richiami di letteratura, un terzo paragrafo che riassume la tecnica di raccolta dei dati, e i paragrafi tre e quattro in cui sono riportati i risultati dell'applicazione del *text mining* e della *sentiment analysis*. Infine, nell'ultimo paragrafo sono riportate delle brevi conclusioni e un cenno a possibili futuri sviluppi.

2 Richiami di letteratura

Nei contributi del 2005 e del 2017, Shiller ha mostrato come il racconto della situazione economica, fatto dagli specialisti del settore o dai social media, abbia un'influenza non trascurabile nelle variazioni che si osservano nelle variabili economiche. Lo stesso Shiller incoraggia l'uso dei *big data* e di metodi che consentono di accedere più facilmente alle informazioni economiche. In questo ventennio si è sviluppato un filone di letteratura che ha cercato di comprendere come poter utilizzare le informazioni tratte dal "racconto" dell'andamento dell'economia a fini di previsione e in generale all'interno di modelli economici. Nel 2012, Choi e Varian si chiedono se gli strumenti di *Google come Google Trends e Google Insights for Search* possono aiutare a prevedere l'attività economica e la loro risposta è nella delimitazione del perimetro della previsione: gli strumenti sono utili se si intende fare una previsione per il presente e non per il futuro e se utilizzati in combinazione con altri indicatori disponibili.

Gli studi sul ruolo del racconto dell'economia si focalizzano in particolare nell'impatto che il clima di fiducia degli agenti economici gioca nel mercato e nella misurazione della fiducia o, come definito nella letteratura anglosassone, del *sentiment*. L'importanza del clima di fiducia nei mercati finanziari è un concetto ormai consolidato (Barberis et al., 1998 e Shleifer e Vishny,1997) e gli studi più recenti si focalizzano sulle tecniche utili a misurare la fiducia (Azar e Lo, 2016, Sun et al.,2016) e ancora l'importanza della "*narrative economics*" nella misurazione del *sentiment* è sottolineata, tra gli altri, nei lavori di Akerlof e Shiller (2009), Johnson e Tuckett (2017), Mordhorst e Schwarzkopf (2017).

Per la misurazione della fiducia sono utilizzate diverse tecniche, tra le quali si trovano quelle basate su variabili di mercato che possono essere considerate *proxy* della fiducia e quelle basate sui risultati di indagini condotte con questionari. Nel tempo, la disponibilità dei *big data* e l'accesso a tecnologie avanzate di analisi dei dati ha fatto crescere l'utilizzo della misurazione del *sentiment* attraverso l'analisi dei testi tratti da articoli su giornali, riviste, canali, *social network*, ovvero dal "racconto economico" (Antweiler e Frank, 2004, Checkley et al., 2017, Nguyen et al., 2015). Alcuni studi dimostrano che gli indicatori di fiducia calcolati sulla base di dati testuali migliorano le *performance* dei modelli di previsione rispetto agli indicatori calcolati su dati da questionari e da variabili *proxy* (Zhou, 2018).

Nel mercato immobiliare l'uso di dati testuali non è così diffuso come negli altri settori dell'economia, tuttavia, è in crescita soprattutto nella costruzione di indicatori di *sentiment* legati al *real estate*. Il fatto che il clima di fiducia influenzi le decisioni di acquisto di immobili è un concetto abbastanza consolidato, intorno al quale c'è ampio accordo in letteratura, e a tal proposito si cita, oltre al lavoro di Case & Shiller (2003), anche il contributo di Jin et al. (2014) che evidenzia che la parte irrazionale degli agenti nel mercato immobiliare influenza l'andamento dei prezzi. In molti studi viene dato risalto a quanto emerge dai *social media* che rappresentano quella che viene definita *"wisdom of the crowd"* ovvero la "saggezza della folla" e in questo senso i ricercatori analizzano la relazione tra quanto emerge dai forum e l'andamento dei rendimenti (Das & Chen, 2007) e più recentemente dai tweet (Sprenger et al., 2014 e ancora Ranco et al., 2015). L'uso di dati testuali tratti dal "racconto economico", raccolti su quotidiani e riviste, viene trattato sotto diversi aspetti che attingono al tema del *sentiment*. L'articolo di Ruscheinsky et al. (2018) analizza il grado di copertura mediatica del mercato immobiliare ed esamina la

relazione tra i dati tratti da notizie e il boom dei prezzi delle case nel Regno Unito. Ancora, nel contributo di Beracha et al. (2019) si esamina se e in che misura il *sentiment* basato sulle notizie può aiutare a prevedere l'andamento del mercato immobiliare commerciale negli Stati Uniti. Tecniche di analisi testuale sono utilizzate anche a supporto dei modelli edonici come nel lavoro di Goncharov & Natkhov (2020) dove è analizzato il contenuto degli annunci di vendita di immobili nella città di Mosca. I *token* unici, unigrammi, o le frasi (bigrammi) sono inseriti come variabili predittori nei modelli di regressione per la spiegazione del prezzo.

3 Data Collection

Per realizzare gli obiettivi stabiliti di individuazione dei *"trending topic"* e *"sentiment analysis",* si è dapprima proceduto con la raccolta dei dati utili. Come detto in premessa, la fonte informativa di interesse è costituita dai titoli e dalla breve descrizione degli articoli di riviste e quotidiani che trattano temi inerenti al mercato immobiliare. Si è stabilito di attingere a dati disponibili in rete attraverso l'implementazione di un'attività di *web scraping*.
In una prima fase si è proceduto con l'identificazione delle riviste e dei giornali da cui estrapolare i dati testuali. La scelta è caduta su 14 testate online: Affari immobiliari, Arsenale 1104, Corriere della Sera, Dove.it, Forbes, Quotidiano immobiliare, Il Sole 24 ore, Monitor Immobiliare, Notiziario immobiliare, Notizie Mondo Immobiliare, pmi.it, Qui Finanza, La Repubblica e TGcom24. Alcune di queste sono riviste di settore, mentre in altri casi si tratta di quotidiani generalisti; in quest'ultimo caso sono state prese in considerazione solamente le sezioni dedicate al mercato immobiliare. Inoltre, i dati sono stati tratti anche dal sito "Intopic.it" che è un raccoglitore automatico di notizie (cronaca, politica, attualità italiana ed internazionale, etc.). In questo specifico caso, gli articoli estrapolati sono stati preventivamente filtrati per categoria (immobiliare), facendo in modo da non avere duplicazioni rispetto agli articoli estratti dalle altre testate. Il ricorso ad Intopic ha permesso di ampliare il bacino di fonti dati, portando la platea finale a oltre 200 testate. Inoltre, il ricorso ad Intopic ha permesso di poter accedere ad articoli pubblicati anche per annualità piuttosto risalenti nel tempo. Nella maggior parte dei casi, infatti, le testate non rendono più disponibili gli articoli e l'accesso è quindi limitato ad una finestra temporale ristretta.
Una breve analisi dei siti e una riflessione sul tipo di informazioni che sarebbe stato utile trarre in *output*, ha condotto ad optare per l'implementazione di uno *scraping* di tipo specifico[2] a sfavore di uno generico[3]. La modalità specifica consente un maggior controllo dei dati estratti e quindi una conseguente maggiore qualità del dato stesso. Ciò facilita le successive fasi di analisi dei dati.
Preliminarmente all'attività di *scraping* vero e proprio, si è provveduto ad un'analisi preventiva delle testate, andando ad indentificare la struttura generale del sito, cercando di stabilirne la forma (blog), il modo in cui gli articoli vengono pubblicati, la struttura degli URLs e la presenza o meno di *feed* RSS[4]. Per quanto riguarda lo *scraping* vero e proprio, sono state utilizzate tecniche e strumenti diversi a seconda della struttura del *sito target.* Questo, sia per adattarsi con più facilità alle specifiche necessità, sia in un'ottica che potremmo definire un "esercizio di stile" per testare i punti di forza e di debolezza di questi strumenti.
I primissimi tentativi di *scraping* su questi siti sono stati effettuati ricorrendo all'utilizzo di un *plug-in* di *Chrome* denominato *Instant Data Scraper*. Questo strumento si è rivelato particolarmente intuitivo, permettendo di estrarre i dati di interesse in maniera veloce e senza necessità di scrivere un codice personalizzato. Questo

[2] Il *web scraping* è specifico quando sia la struttura che il contenuto dei siti da cui estrarre le informazioni sono noti e gli algoritmi devono solo replicare il comportamento di un umano che visita la pagina e salvare le informazioni di interesse.
[3] Il *web scraping* è generico quando non si conosce a priori la struttura e il contenuto del sito. In questo caso si scarica tutto il contenuto che sarà successivamente processato in modo da recuperare le sole informazioni di interesse eliminando le parti non utili.
[4] Un feed RSS (Really Simple Syndication) è un formato di dati standard utilizzato per pubblicare contenuti aggiornati da un sito *web* o blog. È un meccanismo che permette agli utenti di ricevere automaticamente gli aggiornamenti di un sito *web* senza dover visitare manualmente il sito.

strumento ha funzionato molto bene per tutti quei siti in cui i post erano elencati uno sotto l'altro e tutte le informazioni di interesse erano rese disponibili già in questa forma. In altri casi, laddove una o più informazioni non erano direttamente accessibili a livello più alto ma bisognava entrare in ciascun singolo articolo (ad esempio per recuperare la data di pubblicazione), lo strumento ha mostrato qualche carenza. Un altro motivo di debolezza di questo approccio è dovuto al fatto che la procedura non può essere automatizzata, ovvero c'è sempre bisogno dell'intervento umano per fare uno *scraping* di aggiornamento.
Nei casi in cui nel sito *web* erano disponibili *feed* RSS si è deciso di trarne vantaggio, semplificando di molto le operazioni di *scraping*. I file *xml* presi in considerazione contengono tutte le informazioni necessarie per la nostra analisi. Il sito del Sole24Ore mette addirittura a disposizione una serie di RSS suddivisi per categoria e sottocategoria. Per la lettura automatizzata di questi *file* si è fatto ricorso all'utilizzo di *PowerQuery* uno strumento sviluppato da Microsoft (integrato sia in Excel sia in PowerBI) che consente di estrarre, trasformare e caricare dati (ETL) da diverse origini di dati. Questo strumento ha permesso una facile automazione dell'intero processo.
In altri casi l'operazione di *scraping* ha visto coinvolti strumenti quali Google Sheet. In particolare, è stata utilizzata la funzione IMPORTXML di Google Sheets che consente di estrarre dati da pagine *web* tramite query XML secondo il linguaggio di interrogazione XPath[5]. È stato dunque possibile utilizzare questa funzione per accedere a dati strutturati presenti nelle pagine *web* e, in particolare, a tutte le informazioni a noi necessarie (titolo, *link*, autore, data di pubblicazione, etc.).
Infine, nei casi di *scraping* più complessi o privi di soluzioni immediate si è fatto ricorso a *Python e a R*. In *Python* sono state utilizzate essenzialmente le librerie *"requests"* e *"BeautifulSoup*", mentre in *R* la libreria "*rvest*". Questo approccio ha conferito una maggiore flessibilità nell'attività di *scraping*, consentendo un pieno controllo sulle richieste HTTP, la manipolazione del DOM e la gestione dei dati. Inoltre, sia *Python* che *R* consentono di scrivere script di *scraping* completamente automatizzati. Pertanto, l'aggiornamento delle query e del *dataset* è pressoché automatico. Un possibile sviluppo potrebbe essere quello di elaborare un codice in *Python* e/o in *R* che estragga in automatico i dati testuali di interesse dai diversi siti. Tuttavia, poiché la differente struttura dei siti utili, richiede un codice specifico per ciascuno di essi la scrittura risulta piuttosto lunga. È questo il motivo per cui in questo lavoro non si è proceduto in questa direzione.
Per rendere i dati elaborabili, dopo aver concluso lo *scraping* da ciascun sito, si è provveduto ad una preliminare e veloce pulizia dei dati e alla costruzione di un unico *dataset* contenente tutti i record con l'ulteriore indicazione della fonte da cui sono tratti. Per questa operazione di pulizia e di accodamento è stato utilizzato *R* in ambiente *RStudio*.

4 Trending Topic

La parola inglese *"topic"* significa argomento, tema, discussione ed è divenuta di utilizzo molto comune nella lingua italiana con l'espansione dell'uso dei forum, delle *chat* in internet e negli ultimi anni con il diffondersi dei *social media*. Il "*topic*" è l'argomento principale della discussione ed è solitamente definito e indicato dal primo utente che inizia l'interazione su quel tema. Più in generale il *topic* sintetizza in modo chiaro, semplice e univoco il tema della discussione. In senso più ampio il "*topic*" va inteso anche come l'insieme delle parole o degli argomenti connessi al tema indicato dal *topic*.

[5] XPath è un linguaggio di interrogazione utilizzato per navigare ed estrarre informazioni da documenti XML. Questo linguaggio permette di individuare specifici elementi o attributi all'interno di un documento XML (o del DOM di una pagina html) utilizzando percorsi di navigazione basati sulla struttura dell'albero XML

Il "*topic model*" appartiene alla famiglia delle tecniche del *Natural Language Processing* [6] e la sua applicazione consente di estrarre in modo automatico gli argomenti principali presenti in una raccolta di documenti o testi. Si tratta di un approccio non supervisionato che non necessita quindi di un *dataset* etichettato.
In letteratura, i due principali metodi descritti e utilizzati per realizzare approcci *Topic Modelling* sono:

- la *Latent Semantic Analysis* (LSA)
- la *Latent Dirichlet Allocation* (LDA)

In questa elaborazione si è optato per l'applicazione del metodo LDA. I modelli tematici sono anche definiti modelli tematici probabilistici, che si riferiscono ad algoritmi statistici per scoprire le strutture semantiche latenti di un ampio corpo testuale. Data la disponibilità di grandi quantità di dati testuali, i modelli tematici possono aiutare a organizzare grandi raccolte di testo non strutturato, offrendo approfondimenti nella comprensione. Essi rappresentano un modello statistico utilizzato per scoprire argomenti più o meno astratti in una determinata selezione di documenti. I modelli tematici sono particolarmente comuni nel *text mining* per portare alla luce strutture semantiche nascoste nei dati testuali. Gli argomenti possono essere concepiti come reti di termini di collocazione che, a causa della co-occorrenza tra documenti, si può presumere che si riferiscano allo stesso dominio semantico (o argomento). Ciò presuppone che, se un documento riguarda un determinato argomento, ci si aspetta che le parole correlate a quell'argomento appaiano nel documento più spesso che nei documenti che trattano altri argomenti. Ad esempio, cane e osso apparirà più spesso nei documenti sui cani mentre gatto e miagolio appariranno nei documenti sui gatti. Termini come, "tuttavia" o simili, appariranno più o meno allo stesso modo in entrambi.
Per l'applicazione del *topic model* è stato utilizzato l'algoritmo di riferimento in *R* per il "*topicmodels*" che è la libreria LDA (Blei, Ng e Jordan 2003).
Nella fase di *scraping*, effettuata in un'unica data[7], sono stati raccolti circa 3 mila articoli, più del 60% dei quali riferiti a notizie pubblicate nel 2023, in particolare nei mesi di maggio e giugno, mentre la restante parte degli articoli è connessa a pubblicazioni di anni precedenti a partire dal 2006. Nelle analisi che saranno condotte nel seguito dell'elaborato, si farà generalmente riferimento alle informazioni del 2023 o degli ultimi mesi disponibili. In ogni caso, per ciascuna elaborazione sarà specificato il periodo selezionato[8].
L'unità di analisi elementare è rappresentata dal singolo titolo e dalla breve descrizione dell'articolo. Con l'ausilio delle librerie disponibili in *R*, si è proceduto alla lettura e al *pre-processing d*ei dati testuali raccolti.
Il *dataset* è stato considerato come un *corpus* unico e, su questo si è proceduto con le successive fasi dell'analisi. In primo luogo, si è effettuata la pulizia del testo, riducendo tutto in minuscolo, eliminando eventuali simboli, numeri e url. Inoltre, sono state eliminate le cosiddette *stopword*[9], quei termini che non rappresentano informazione rilevante per l'individuazione e la descrizione dei *topic*. In questa applicazione si è presa a riferimento una lista di parole da lessico italiano, ma è stata necessaria anche una fase di individuazione di parole da escludere dal *corpus* per il contesto specifico di analisi (nomi ricorrenti di autori dei testi, parole chiave del contesto di analisi, periodi ripetuti nei titoli di alcune testate, etc..) sviluppata con un approccio di tipo "manuale". Successivamente, si è effettuata la t*okenizzazione* ovvero si è applicato il processo che trasforma il documento in una

[6] Con Natural Language Processing si fa riferimento a quel campo di ricerca che ha l'obiettivo di sviluppare algoritmi in grado di analizzare, rappresentare e quindi "comprendere" il linguaggio naturale, scritto o parlato, in modo simile agli esseri umani.
[7] Dopo le opportune analisi che hanno condotto a stabilire quali tecniche fossero più adeguate, lo *scraping* è stato effettuato in un'unica soluzione il 14 luglio 2023.
[8] La scelta del periodo selezionato, ovviamente, ha un discreto rilievo perché gli argomenti su un tema possono avere una particolare concentrazione in alcuni periodi (anche brevi) e non in altri.
[9] Si tratta di parole comuni che appaiono nel testo con una frequenza molto elevata e che aiutano a costruire le frasi ma non hanno un significato di per sé come, ad esempio, gli articoli ("il", "lo", "la", "dei" etc.) o le congiunzioni ("e") e altre. Questi termini sono inseriti in una stop list che viene presa a riferimento per effettuare la rimozione delle parole non significative.

lista di *token*. Il risultato è un oggetto "*tokens"* su cui si è effettuata un'analisi delle frequenze, una *network analysis* e da ultimo sono estratti i *topic*. Il *pre-processing* può essere tuttavia differente a seconda degli obiettivi della ricerca e quindi necessita di una certa flessibilità che va decisa a seconda degli scopi dell'analisi, ad esempio, per la *sentiment analysis* si fa riferimento ad elenchi di parole etichettati (*il lexicon*) come positivi o negativi, senza applicare uno *stemming*[10] preventivo. Nel nostro studio abbiamo scelto di non applicare la procedura di *stemming* in nessuna fase dell'analisi testuale, al fine di analizzare con maggiore dettaglio le co-occorrenze tra *token,* e soprattutto in quanto la finalità dello studio tende a concentrarsi sulla semantica, sui legami logici tra *token,* piuttosto che sul semplice conteggio delle occorrenze.

4.1 Risultati dell'applicazione

L'analisi di frequenza, effettuata solo sugli articoli relativi ai mesi del 2023 (da gennaio a luglio), evidenzia i *token* maggiormente ricorrenti nel periodo. Emerge come la discussione per l'imposta sugli immobili sia stata assolutamente dominante, con il termine IMU che risulta quello più ricorrente negli articoli analizzati. A questo stesso tema si lega, in molti casi anche il *token* "pagare" come emergerà in modo più evidente dall'esame dei *bi-gram* ovvero dell'unione di due termini e dalla successiva *network analysis*. Di rilievo anche il termine "Milano" molto presente negli articoli trattati. Per il mercato immobiliare Milano è infatti un centro di attenzione, la città attrae da sola una quota elevata di investimenti immobiliari ed è un punto di riferimento per il mondo della finanza immobiliare. È una città in cui la dinamica del *real estate* può essere considerata un *driver* dell'intero mercato nazionale.

Nei primi mesi del 2023, molta attenzione è stata dedicata al tema dei prezzi degli immobili, al mercato della locazione con particolare interesse al valore dei canoni. Non meno importante l'argomento "mutui" riguardo soprattutto alla crescita sostenuta dei tassi applicati alle nuove erogazioni. Un crescente interesse è stato rivolto anche alla direttiva europea *green* e all'impatto che questa avrà nel mercato immobiliare. La sintesi grafica di quanto emerso dal *text mining* è riportata in Figura 1 dove si mostra il *wordcloud*[11] dei *token* la cui frequenza risulta non inferiore a 50 e l'istogramma di frequenza delle *top word*.

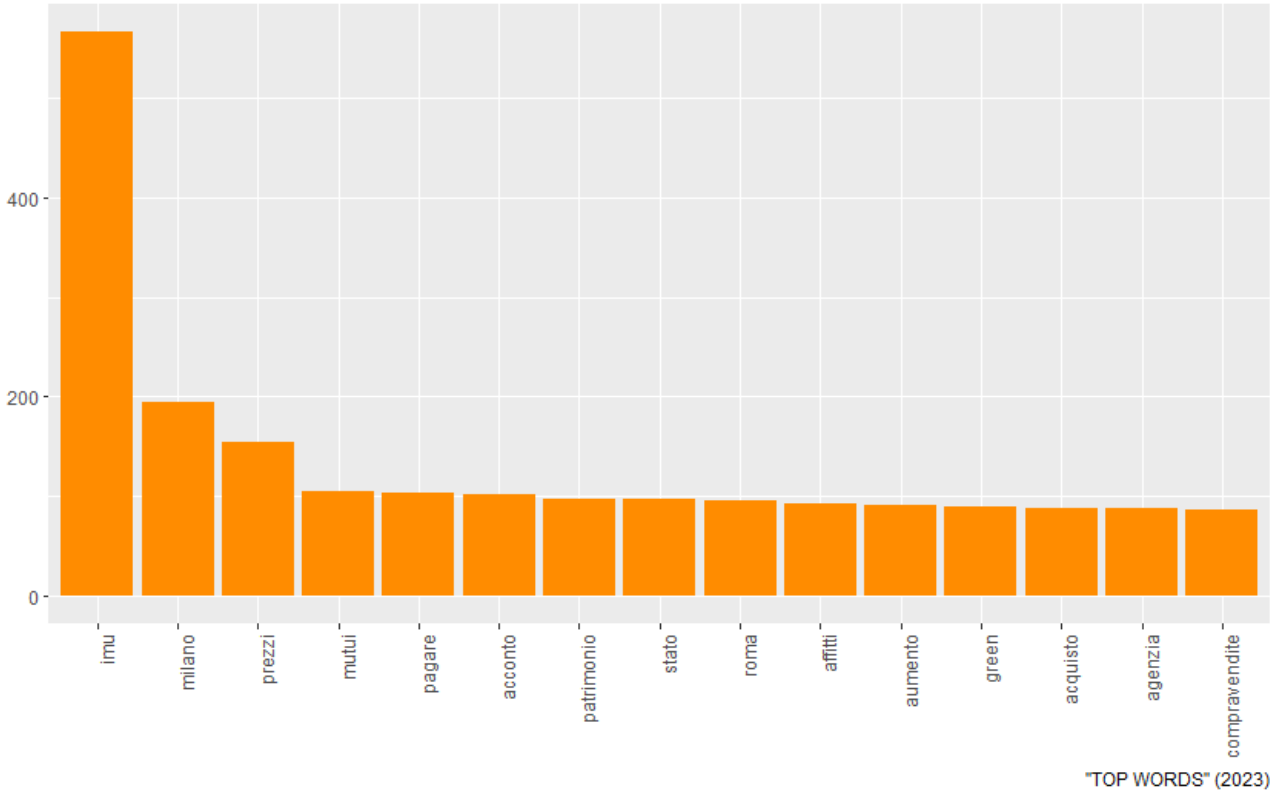
[10] Lo *stemming* è un'operazione che consiste nella riduzione della forma flessa di una parola alla sua forma radice, detta tema. Il processo di *stemming*, generalmente applicato prima della fase di modellazione è una normalizzazione del testo. Con questa procedura di raggruppano termini correlati, ad esempio si possono aggregare tutte le varie coniugazioni di un verbo.

[11] Il *wordcloud* è un grafico che mostra le parole usate in un testo. Ciascuna parola ha dimensioni proporzionale alla frequenza con cui appare nel testo.

Figura 1 - *Wordcloud* e distribuzione delle frequenze dei *token* (articoli gennaio - luglio 2023)

Fonte: Elaborazioni degli autori

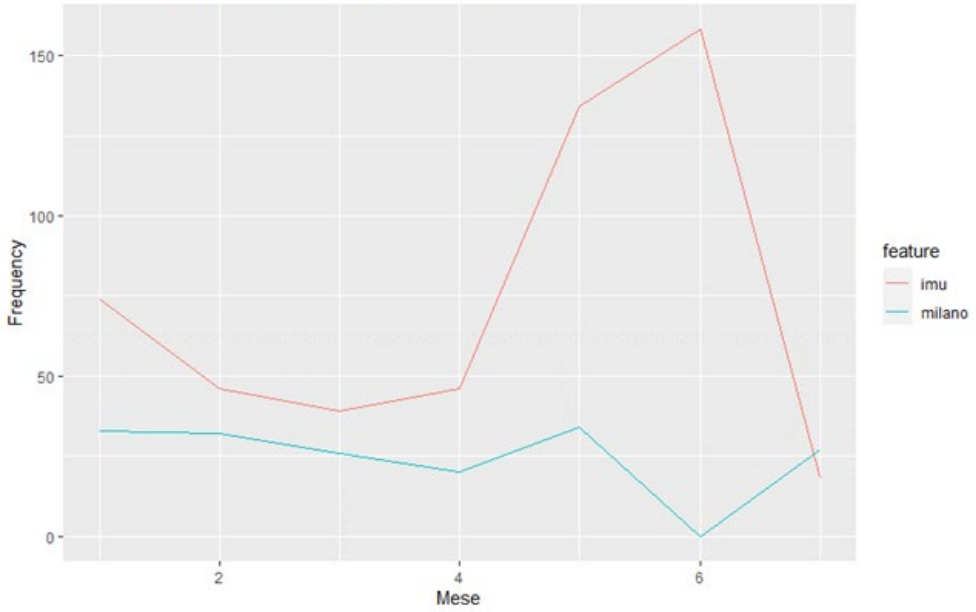
Figura 2: Frequenze delle "*top word*" - (articoli gennaio - luglio 2023)



Fonte: Elaborazioni degli autori

Di interesse risulta anche la distribuzione mensile dei *token* maggiormente presenti. In particolare, considerando i primi due *token*, “IMU” e “Milano” l’analisi evidenzia le specificità e le diversità dei due temi. L’interesse per l’imposta sugli immobili, l’”IMU” appunto, è elevato nei mesi di gennaio, maggio e giugno, in prossimità delle scadenze stabilite per il pagamento. Per contro, l’interesse verso “Milano”, termine da intendersi come detto, in senso ampio come mercato immobiliare di Milano e il complesso delle attività legate al *real estate* di questa città, resta più o meno costante in tutti i mesi del 2023. Quanto detto ben si evince dal grafico in Figura 3 che riporta le frequenze mensili distintamente per i due *token*.

Figura 3: Frequenze mensili dei token "IMU" e "milano"



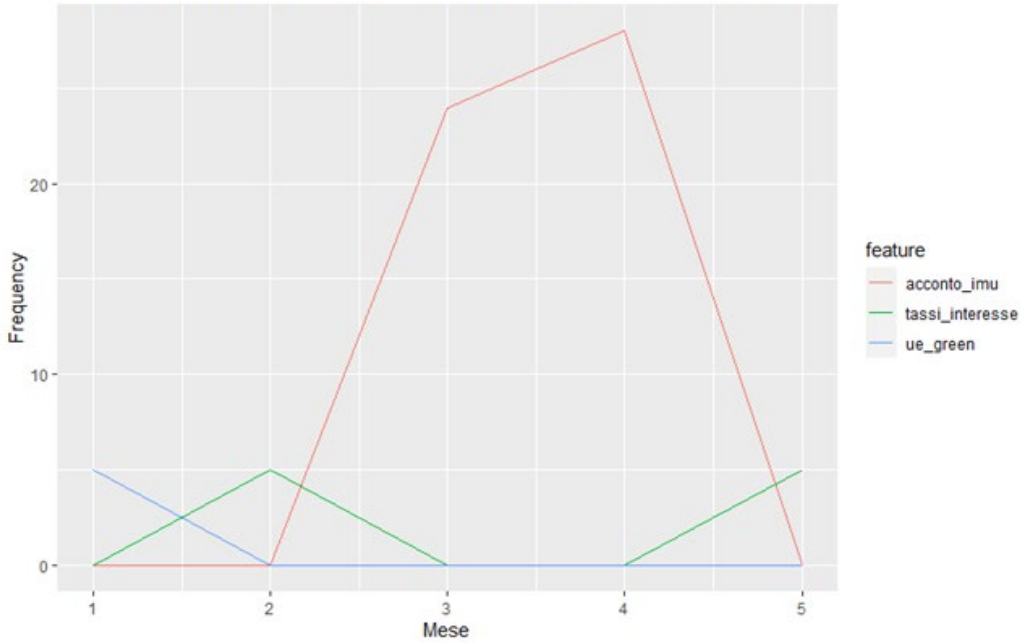
Fonte: Elaborazioni degli autori

È utile anche analizzare l’occorrenza dei *“bi-gram”*, definiti come tutti le possibili combinazioni di due *token* (anche sprovvisti di significato semantico) ottenibili in ciascun articolo. Di seguito si visualizza il *wordcloud* ottenuto dal conteggio delle occorrenze dei *bi-gram* la cui frequenza è superiore a 50. Anche dall’analisi dei termini congiunti emerge chiaramente che il tema dell’IMU inteso come scadenza imminente del tributo (‘acconto-IMU’, ‘pagare-IMU’, ‘esenzione-IMU’) è quello maggiormente presente. Si evidenziano comunque come temi discussi anche quello degli ‘affitti-brevi’, ‘contratto-locazione’ e ‘tassi-interesse’.
Nel dettaglio mensile il grafico di Figura 5 mostra che mentre per “acconto IMU” e “ue green” si ha una concentrazione evidente di riferimenti in un periodo specifico, il tema dei tassi di interesse è trattato in due periodi distinti, aprile e luglio con la stessa intensità, mai raggiunta nella restante parte dell’anno.

Figura 4: Wordcloud dei bi-grammi (articoli gennaio - luglio 2023)

Fonte: Elaborazioni degli autori

Figura 5: Frequenze mensili dei bi-gram "acconto_IMU", "tassi_interesse" e "ue_green"[12]



Fonte: Elaborazioni degli autori

[12] Si precisa che il bi-gram "ue_green" compare solo nel mese di gennaio e nei successivi ha una frequenza nulla.

Con l'analisi descritta in precedenza, si sono rilevati gli argomenti di tendenza, i *topic,* dei primi sette mesi del 2023 e selezionando i 5 *token* più frequenti in ciascuno di essi se ne mostra l'evidenza nel grafico di Figura 6. Tra tutti i *topic* individuati, si è scelto di selezionare, in modo empirico con una valutazione soggettiva dei risultati dell'analisi, i primi 8 *topic* in ordine di importanza. A tal proposito, si precisa che l'estrazione del numero di componenti è stata verificata anche attraverso l'applicazione, tra le diverse disponibili, delle due metriche di Cao et al.[13] (2009) e Deveaud[14] (2014) che restituivano un numero di componenti pari a 5. Tuttavia, fatti diversi tentativi, si è reputato in modo qualitativo, basandosi sulle conoscenze del fenomeno in indagine, che il numero di componenti adeguato da considerare fosse 8.
Le barre si riferiscono all'indicatore che misura la probabilità di ciascun termine di essere generato dall'argomento (indice "beta").

Figura 6: Diagramma dei *topic* individuati (articoli gennaio - luglio 2023)

Fonte: Elaborazioni degli autori

Si fornisce di seguito un'interpretazione dei *topic* individuati provando ad assegnare a ciascuno un'etichetta come di seguito:

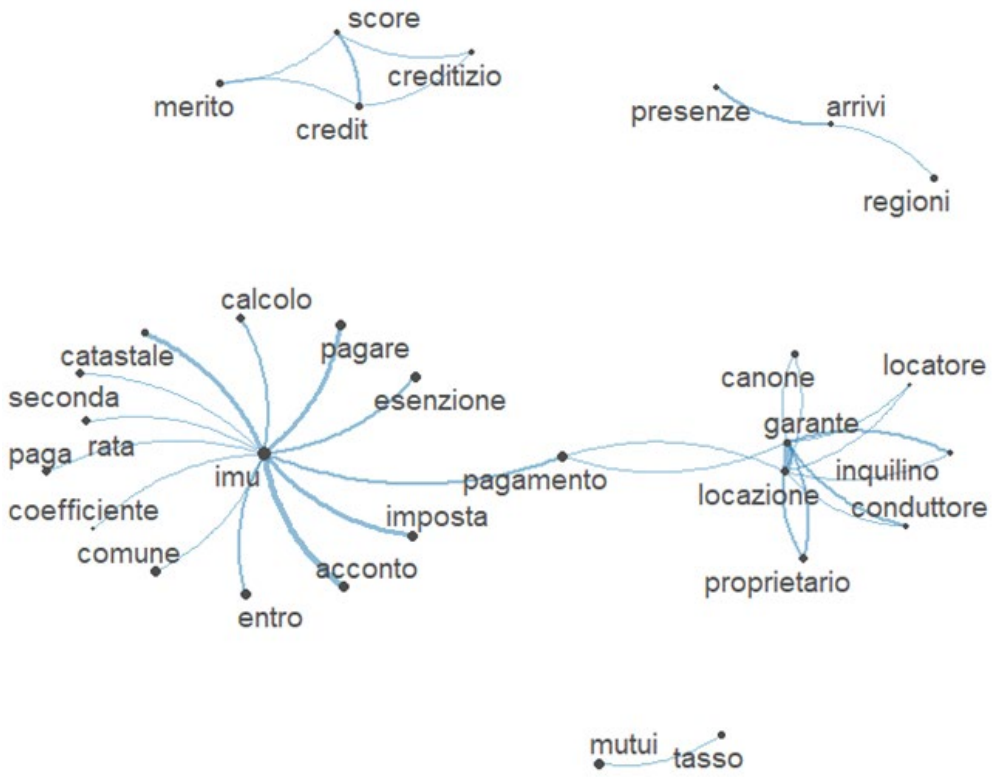
[13] La tecnica seleziona il modello migliore sulla base della densità dei *clusters* (*topic*). Si calcola per ciascun modello la distanza tra i *topic* tramite la similarità media del coseno (r), che ci permette di definire la densità di un *topic* rispetto agli altri. Il valore migliore di K è raggiunto quando r tra *topic* è al minimo, ovvero quando la distanza è massima.

[14] La metrica proposta sfrutta la divergenza Jensen-Shannon per determinare quel valore di K che massimizza la distanza tra ciascuna coppia di topics.

- “IMU” è l’argomento maggiormente di tendenza ed è rappresentato dal *topic 6* dove compaiono i termini associati acconto, scadenza, pagare, imposta. IMU è presente anche nel *topic 5* dove compare anche esenzione;
- “mercato” è il termine con cui si possono sintetizzare *i topic* 3, 4 e 8. In particolare, il *topic 3* rappresenta il tema della dinamica dei prezzi, mentre nel *topic 8* sono rappresentati i temi inerenti le compravendite, gli affitti e i mutui intesi come andamento dei volumi e dei valori. Infine, nel *topic 4* sono rappresentati le città di Milano, Napoli intese come mercati immobiliari di rilievo;
- il *topic 2* può essere sintetizzato con il termine “investimenti” essendo individuato dalla stessa parola “investimenti” e dal termine “Milano” che, come visto, rappresenta una realtà specifica che influenza e impatta l’intero mercato immobiliare;
- “UE” è l’etichetta che si adatta al *topic 7* con contenuti che riguardano sia la direttiva *green* sia i tassi dei mutui fissati dalla BCE;
- il *topic 1* rappresenta un argomento divenuto di tendenza nei giorni successivi al decesso dell’ex presidente del consiglio italiano Silvio Berlusconi dove dal punto di vista immobiliare si è discusso del suo patrimonio immobiliare e dell’eredità.

In ultimo, l’analisi testuale è stata integrata con una *network analysis*. Quest’ultima, si basa sulla creazione di una matrice di co-occorrenze tratte dal corpus, costruita contando il numero di collegamenti tra *i token*. I risultati possono essere rappresentati come un grafo, con i nodi che indicano i singoli *token* e gli archi che rappresentano le connessioni, potendo in tal modo tracciare ogni concetto in relazione agli altri e visualizzare come le diverse parole sono collegate in una specifica narrazione. Il grafico permette di individuare i collegamenti maggiormente presenti ed è altresì possibile visualizzare la rete con diversi livelli di profondità.
I risultati applicati ai testi qui esaminati sono riportati nella rete di Figura 7 che mostra i termini che presentano un elevato grado di connessione con gli altri. Le parole più connesse sono “IMU”, “locazione” e “credito”. L’”Imu” presenta molte connessioni con tutti i termini attinenti al pagamento dell’imposta (“acconto”, “rata”, etc), ma anche “locazione” risulta connessa a canone, pagare. Da questa analisi emerge il tema del turismo e quindi degli immobili utilizzati per questo scopo.

Figura 7: Diagramma dei *network* dei termini maggiormente associati



Fonte: Elaborazioni degli autori

5 Sentiment Analysis

Per la *sentiment analysis* sono stati esplorati due approcci: la costruzione di un indicatore di fiducia e l'analisi delle emozioni.

Per la costruzione dell'indicatore di *sentiment* si è assegnato un punteggio ad ogni articolo composto dal titolo e da un breve estratto del contenuto, il cosiddetto *snippe*t. Per l'assegnazione del punteggio si è utilizzato un vocabolario, l'*Italian Sentiment Lexicon*, i cui lemmi sono associati a punteggi di *sentiment* positivi e negativi. In particolare, si è utilizzato il lessico Sentix proposto da Basile e Nissim (2013)[15] che allinea diverse risorse lessicali esistenti e indipendenti (WordNet, MultiWordNet, BabelNet, SentiWordNet). La metodologia di assegnazione del punteggio di polarità a ciascun lemma è stata testata in uno studio sperimentale basato su algoritmi di classificazione automatica su dati estratti da *Twitter*, per la prima volta sulla lingua italiana. *Sentix* contiene molti lemmi duplicati e consiste di circa 75 mila lemmi complessivi di cui circa 42 mila unici. Per garantire risultati univoci e riproducibili, si è deduplicato Sentix e calcolata la media dei punteggi (positivi e negativi) dei lemmi duplicati.

Inoltre, il vocabolario è stato integrato con alcuni termini che non potevano essere trascurati per gli obiettivi di questo studio. I punteggi assegnati sono stati generati a partire da termini «simili», così ad esempio al termine "superbonus" è stato associato lo stesso punteggio assegnato in Sentix al termine "incentivo", alla parola "" è

[15] Per un approfondimento si può far riferimento al contributo di Basile e Nissim del 2013, "Sentiment analysis on Italian tweets".

stato assegnato lo stesso punteggio associato a “tassa” e ancora a “mutuo” è stato associato il punteggio di “debito”.

L’indice del *sentiment* è calcolato a partire dal calcolo di polarità e intensità per ciascun articolo, a partire dai punteggi “negativi” (n) e “positivi” (p) medi ottenuti per i termini presenti nel vocabolario.

Dalla coppia di coordinate (p, n) appartenenti allo spazio R^2 si passa alle coordinate polari (i, Θ), in cui ogni punto è espresso da un angolo, Θ e da una distanza da un punto detto “polo”, i), da cui si ricavano due grandezze aggiuntive, la polarità *ω*, e l’intensità, *i*, specificati come segue:

$$\omega = 1 - 4\,\Theta/\pi,$$

$$\text{con } \omega \in [-1, 1]\,;\ \Theta = 2\,\text{arcotg}\left(\frac{n}{\sqrt{(p^2 + n^2\,) + p}}\right) \text{ e } \pi \text{ costante matematica}$$

$$i = \sqrt{p^2 + n^2} \quad per\ i \in [0, 1]$$

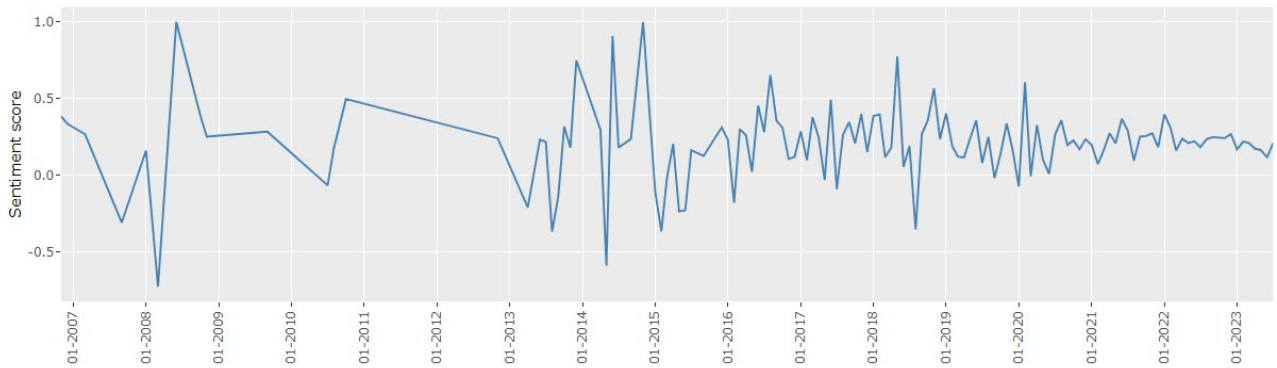
L’indice di *sentiment*, S per un certo periodo j, è allora dato dalla media delle polarità pesate con l’intensità:

$$S = \overline{\omega_j} = \sum_a i_a\,\omega_a \Big/ \sum_a i_a \quad con\ a = 1, \dots, k$$

Con a = singolo articolo e k = numero complessivo degli articoli del periodo j[16].

In Figura 8 è mostrato l’andamento dell’indice S per tutte le annualità sulle quali si hanno a disposizione informazioni. Ad eccezione di qualche picco negativo, S si mantiene sempre in campo positivo con oscillazioni più o meno ampie. Considerando l’ultimo periodo, dal 2022 a luglio 2023, l’indicatore risulta decrescente e dal valore 0.4 del 2007 è sceso a circa 0,1 a giugno 2023 e si è rialzato seppur di poco a luglio 2023.

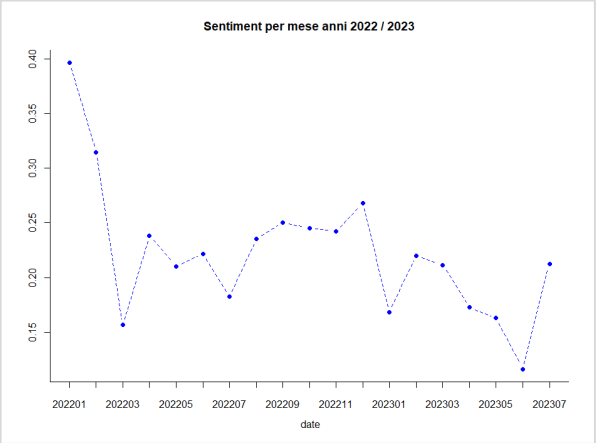
Figura 8: Andamento mensile del *sentiment* dal 2007 a luglio 2023

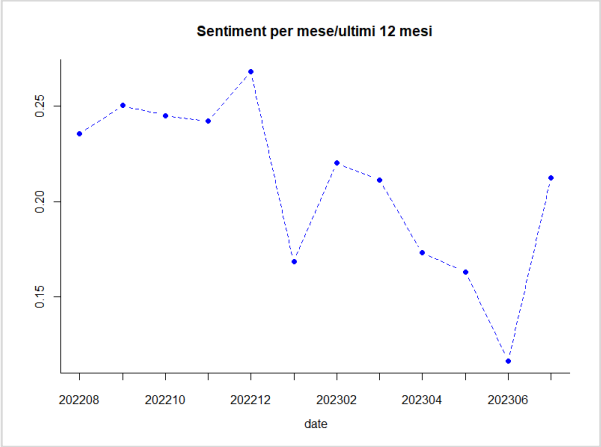


Fonte: Elaborazioni degli autori

[16] La metodologia replica il processo di elaborazione del Social Mood Economy Index di Istat, Indice sperimentale dell’Istituto Nazionale di Statistica. IL metodo è consultabile nelle note della 13ma conferenza nazionale di Statistica del 4-5-6 luglio 2018 (Zardetto) ai seguente link:
D. Zardetto, Using Twitter data for the Social Mood on Economy Index | PPT (slideshare.net)
Social Mood on Economy Index (istat.it)

Figura 9: Andamento mensile del *sentiment* dal 2022 a luglio 2023 (grafico di sx) e dettaglio da agosto 2022 a luglio 2023





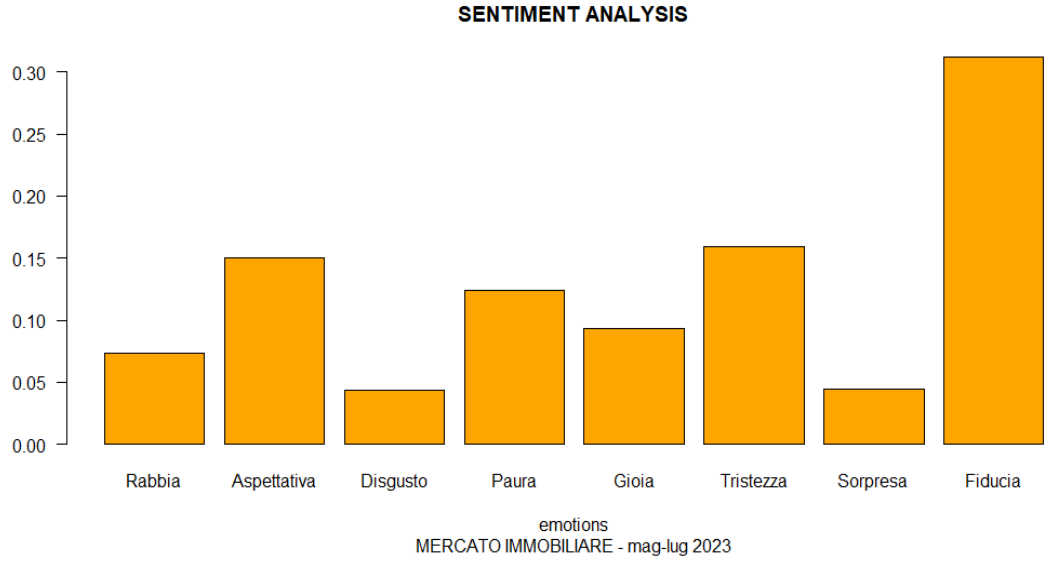
Fonte: Elaborazioni degli autori

L'analisi relativa al *sentiment* è stata integrata anche con un'analisi delle emozioni che possono trarsi dai testi disponibili. Per questa elaborazione si è utilizzato il lessico *NRC Word-Emotion Association Lexicon*, sviluppato da Saif M. Mohammad, uno scienziato del *National Research Council* Canada (NRC), presente nel pacchetto *syuzhet*[17] di *R*. Tra i diversi lessici disponibili nella *library* si è optato per il NRC in quanto è l'unico dei quattro che attualmente può essere utilizzato con testi non inglesi. Questo lessico, oltre a comprendere valori di *sentiment* positivi e negativi prevede anche otto categorie emotive. Il *set* di dati che forma il lessico è stato annotato manualmente utilizzando la tecnica Maximum Difference Scaling, o MaxDiff, per determinare gli insiemi di parole con valori negativi o positivi più elevati rispetto ad altre parole, una sorta di classifica dell'intensità del sentimento delle parole. Questo lessico che contiene 14.182 unigrammi (parole) classificati come positivi o negativi, classifica anche la connessione di una parola a varie emozioni: rabbia, aspettativa, disgusto, paura, gioia, tristezza, sorpresa e fiducia.

Utilizzando la traduzione automatica, è possibile che ci siano degli errori o che alcune sfumature linguistiche non siano del tutto comprese. Tuttavia, in questa fase dello studio l'obiettivo è avere un'idea generale di quali emozioni siano maggiormente presenti nel testo. Nel grafico a barre di Figura 10 si riporta la distribuzione delle emozioni associate agli articoli disponibili da maggio a luglio 2023. L'emozione che ha la maggiore frequenza è la fiducia che si presenta nel 30% dei testi analizzati, seguono con livelli inferiori le emozioni legate all'attesa e alla paura.

[17] Il pacchetto R Syuzhet è stato rilasciato nel 2015 da Matthew Jockers:
https://cran.r-project.org/web/packages/syuzhet/syuzhet.pdf
Il pacchetto è stato progettato per consentire di scegliere tra quattro possibili lessici di *sentiment*: Bing, Afinn, Stanford e NRC Word-Emotion Association Lexicon.

Figura 10: Distribuzione del peso per emozione (articoli da maggio a luglio 2023)



Fonte: Elaborazioni degli autori

Concentrandosi sulle 4 tra le 8 emozioni che risultano maggiormente rappresentate si visualizzano attraverso il *wordcloud* le *top words*[18] associate a ciascuna emozione: aspettativa, fiducia, tristezza, paura.

Figura 11: *Wordcloud* organizzato rispetto alle quattro emozioni più frequenti

Fonte: Elaborazioni degli autori

[18] Si precisa che in questa elaborazione il termine IMU non presente nel dizionario utilizzato è stato inserito come tassa.

6 Conclusioni

In questo contributo sono stati estratti da diversi siti *web*, con tecniche di *web scraping*, titoli e *snippet* di articoli apparsi su riviste di settore e sezioni di quotidiani dedicati al mercato immobiliare. La base dati testuale costruita è stata trattata con tecniche *text mining* per estrarre i *trending topic* e per costruire un indicatore della fiducia degli operatori nel mercato immobiliare, nonché individuare l'emozione dominante in certo arco temporale.
L'utilizzo nell'analisi economica di dati testuali tratti da quotidiani e riviste è già sperimentato sia in ambito internazionale sia nel contesto italiano. Nel comparto immobiliare l'integrazione della fonte dati di tipo testuale a quelle più tradizionali è una pratica abbastanza diffusa nel mondo anglosassone, ma nel panorama italiano non si è a conoscenza di specifici utilizzi di questa tipologia di dati. In questo senso, il lavoro che qui si è presentato rappresenta un primo esperimento con il quale si è voluta dare evidenza soprattutto delle potenzialità che un'analisi di questo tipo può offrire.
Si sottolinea, in particolare, come la possibilità di estrarre, collezionare, analizzare il linguaggio naturale costituisca un avanzamento impensabile fino a pochi decenni fa. Inoltre, la disponibilità di *software* facili da utilizzare offre la possibilità di poter applicare tecniche sofisticate. Sintetizzare velocemente l'umore degli operatori economici in un certo settore offre un vantaggio informativo notevole che aiuta a comprendere in tempo reale l'evoluzione immediata di un fenomeno e a fare previsioni per un arco temporale più lungo. Nel mercato immobiliare comprendere in anticipo le decisioni di acquisto/vendita o concedere/prendere in locazione un immobile è un'informazione rilevante per fare previsioni e più in generale per supportare scelte di politica economica.
L'analisi dei *trending topic* ha fatto emergere alcuni temi "caldi" del primo semestre 2023 quali l'attenzione verso il pagamento dell'imposta sugli immobili, IMU, l'interesse per i contenuti e gli impatti della direttiva europea "green", mutui e locazioni come argomenti di rilievo per l'analisi del mercato immobiliare, Milano come città che più di ogni altra è l'espressione a tutto tondo del mercato immobiliare italiano e ancora la curiosità verso le proprietà immobiliari dell'ex presidente del consiglio Berlusconi a seguito della sua morte.
Si è costruito l'indicatore di *sentiment* che può aiutare a fare previsioni sull'andamento dei volumi e dei prezzi nel mercato. Per meglio comprendere la capacità informativa dell'indicatore sarebbe utile comparare l'andamento di S con quello delle serie dei volumi di compravendita nazionali, del numero di immobili locati e con l'indice dei prezzi. Al momento questa analisi non è stata condotta, ma si tratta di uno sviluppo che realisticamente può essere effettuato. Inoltre, non essendo state testate, in questo studio, le capacità di *nowcasting* e *forecasting* dell'indicatore non si è in grado di fornire una misura dell'effettivo ausilio a fini di previsione.
Oltre ad ulteriori sviluppi l'analisi meriterebbe anche miglioramenti e affinamenti. Sarebbe utile un'integrazione specifica dei dizionari utilizzati o la creazione di vocabolari specifici con termini usuali nel mercato immobiliare. Si tratta, comunque, di una prima analisi che ha inteso esplorare le potenzialità dell'utilizzo di dati testuali nell'ambito del comparto immobiliare.
In questo contributo si è voluto evidenziare come l'integrazione dell'analisi con dati non convenzionali per tipologia e fonte possa arricchire e offrire nuovi indicatori per le analisi del mercato immobiliare.
Vale la pena soffermarsi, seppur brevemente, su alcuni punti di attenzione dell'analisi svolta.
In primo luogo, vanno tenuti in giusta considerazione gli elementi di complessità dell'analisi svolta. La semplice analisi descrittiva dei lemmi non è in grado di far emergere il significato che questi assumono all'interno dei testi, moltissimi sono i termini che possono avere accezione diversa a seconda del contesto di riferimento ed inoltre intervengono spesso i modificatori di intensità, che possono rafforzare o diminuire la portata di certi termini fino a modificare o invertire il significato di un termine letto isolandolo dai termini che lo seguono o lo precedono. Gli algoritmi semantici non sempre riescono a considerare tutti questi fattori e a bilanciare le connotazioni positive,

negative, neutre, per interpretare correttamente l'attitudine di chi scrive rispetto ad una scala di polarità impostata. In quest'ottica, l'analisi basata sulle *emotion* garantisce un maggiore livello di dettaglio poiché consente di andare oltre i limiti di bilanciamento delle connotazioni positive e negative. Le *emotion* riescono a riconoscere l'emozione specifica espressa nel testo, garantendo un risultato dell'analisi più preciso.
Nel caso dei dati qui utilizzati, articoli di pubblicistica, vi è anche l'interposizione dello stile e delle idee del giornalista che riporta la notizia. Non è di secondaria importanza la scelta dei termini del modo in cui sono espresse le notizie che sono le informazioni di base sulle quali si poggia sia il *text mining* sia la *sentiment analysis*. Va detto che se da un lato c'è uno scrivente che cerca di comunicare un certo messaggio, dall'altro ci sono lettori diversi che interpreteranno lo stesso messaggio secondo le proprie convinzioni, conoscenze, credenze, competenze.
Non in ultimo va evidenziato il cambio epistemologico che l'uso di queste fonti dati non convenzionali porta con sé. Sebbene nel contributo non si sia dato esplicito spazio al tema, emerge che vi è un passaggio da un approccio che si fonda su una teoria, *theory-driven*, ad un'analisi *data-driven*. Nella ricerca teorica si parte dallo sviluppo di un'ipotesi testata con l'ausilio dell'analisi dei dati e traendo conclusioni teoriche basate sui risultati. Diversamente l'approccio *data-driven* è di tipo esplorativo e parte dall'analisi dei dati per estrarre approfondimenti scientificamente interessanti. Due percorsi diversi che, come suggerito da alcuni studi, nell'era dei *big data* dovrebbero rafforzarsi a vicenda (Maass et al, 2018) e non essere esclusivi. Questo ultimo approccio è quello che probabilmente può essere validamente perseguito anche nelle analisi del *sentiment* nel mercato immobiliare.

Bibliografia

- Akerlof, G. A., & Shiller, R. J. (2009). *Animal spirits: How human psychology drives the economy, and why it matters for global capitalism*. Princeton University Press.
- Antweiler, W., & Frank, M. Z. (2004). *Is all that talk just noise? The information content of Internet stock message boards*. The Journal of Finance, 59(3), 1259–1294.
- Aprigliano, V., Emiliozzi, S., Guaitoli, G., Luciani, A., Marcucci, J., Monteforte, L. (2021). *The power of text-based indicators in forecasting the Italian economic activity*. Temi di discussione (Economic working papers) 1321, Bank of Italy, Economic Research and International Relations Area
- Azar, P. D., & Lo, A. W. (2016). *The wisdom of Twitter crowds: predicting stock market reactions to FOMC meetings via Twitter feeds*. Journal of Portfolio Management Special QES Issue 2016, 42(5), 123–134.
- Barbaglia, L., Consoli, S., Manzan, S. (2022). *Forecasting GDP in Europe with Textual Data*. Available at SSRN*: https://ssrn.com/abstract=3898680 or http://dx.doi.org/10.2139/ssrn.3898680*
- Barberis, N., Shleifer, A., & Vishny, R. (1998). *A model of investor sentiment.* Journal of Financial Economics, 49(3), 307–343.
- Basile V., Nissim M. (2013). *Sentiment analysis on Italian tweets. In Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107, Atlanta, Georgia. Association for Computational Linguistics.
- Beracha, E., Hausler, J., Lang, M. (2018). *On the Relationship between Market Sentiment and Commercial Real Estate Performance* – A Textual Analysis Examination. Analysis Examination, Journal of Real Estate Research, 41:4, 605-638, DOI: 10.22300/0896-5803.41.4.605.
- Blei, D. M., Ng, A. and Jordan, M. (2003). *Latent Dirichlet allocation. Journal of Machine Learning Research*
- Cao J., Xia, T., Li, J., Zhang, Y., Tang, S. (2009). *A density-based method for adaptive LDA model selection. Neurocomputing*, https://doi.org/10.1016/j.neucom.2008.06.011.
- Case, K. E., & Shiller, R. J. (2003). *Is there a bubble in the housing market?* Brookings Papers on Economic Activity, 2, 299–342
- Checkley, M. S., Higón, D. A., & Alles, H. (2017). *The hasty wisdom of the mob: How market sentiment predicts stock market behavior.* Expert Systems with Applications, 77, 256–263.
- Choi, H., & Varian, H. (2012). *Predicting the present with Google Trends*. Economic record, 88, 2-9.
- Das, S. R., & Chen, M. Y. (2007). *Yahoo! for Amazon: Sentiment extraction from small talk on the Web*. Management Science, 53(9), 1375–1388.
- Deveaud, R., Sanjuan, E., Bellot, P. (2014). *Accurate and Effective Latent Concept Modeling for Ad Hoc Information Retrieval.* Document numérique. 17. 10.3166/dn.17.1.61-84.
- Goncharov, G., Natkhov, T. (2020). *Textual Analysis of Pricing in the Moscow Residential Real Estate Market.* Higher School of Economics Economic Journal. 24. 101-116. 10.17323/1813-8691-2020-24-1-101-116
- Jin, C., Soydemir, G., & Tidwell, A. (2014). *The US housing market and the pricing of risk: Fundamental analysis and market sentiment*. Journal of Real Estate Research, 36(2), 187–219
- Johnson S. G. B., & Tuckett D. (2017). *Narrative decision-making in investment choices: How investors use news about company performance*. SSRN preprint. https:// doi. org/ 10. 2139/ ssrn. 30374 63.

- Maass, W., Parsons, J., P., Sandeep, S., Veda C.; and Woo, C. (2018) *"Data-Driven Meets Theory-Driven Research in the Era of Big Data: Opportunities and Challenges for Information Systems Research,"* Journal of the Association for Information Systems, 19(12), DOI: 10.17705/1jais.00526.
- Mordhorst, M., & Schwarzkopf, S. (2017). *Theorising narrative in business history*. Business History, 59(8), 1155–1175.
- Nguyen, T. H., Shirai, K., & Velcin, J. (2015). *Sentiment analysis on social media for stock movement prediction*. Expert Systems with Applications, 42(24), 9603–9611.
- Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M., & Mozetič, I. (2015). *The effects of Twitter sentiment on stock price returns*. Plos One, 10(9), e138441.
- Ruscheinsky, J., Lang, M., Schäfers, W. (2018). *Real estate media sentiment through textual analysis.* Journal of Property Investment & Finance. 36. 10.1108/JPIF-07-2017-0050.
- Shiller, R. J. (2005). *Irrational exuberance*. Princeton University Press.
- Shiller, R. J. (2017). *Narrative economics*. The American Economic Review, 107(4), 967–1004.
- Shleifer, A., & Vishny, R. W. (1997). *The limits of arbitrage*. The Journal of Finance, 52(1), 35–55.
- Soo, Cindy. (2018). *Quantifying Sentiment with News Media across Local Housing Markets.* Review of Financial Studies. 31. 3689-3719. 10.1093/rfs/hhy036
- Sprenger, T. O., Sandner, P. G., Tumasjan, A., & Welpe, I. M. (2014). *News or noise? Using Twitter to identify and understand company-specific news flow*. Journal of Business Finance & Accounting, 41(7–8), 791–830.
- Sun, L., Najand, M., & Shen, J. (2016). *Stock return predictability and investor sentiment: A high-frequency perspective.* Journal of Banking & Finance, 73, 147–164.
- Walker, Clive B. (2016). *The direction of media influence: Real-estate news and the stock market.* Journal of Behavioral and Experimental Finance, Elsevier, vol. 10(C), pages 20-31.
- Zhou, G. (2018). *Measuring investor sentiment.* Annual Review of Financial Economics, 10(1), 239–259.
- Zhu, Enwei & Wu, Jing & Liu, Hongyu & Li, Keyang. (2022). *A Sentiment Index of the Housing Market in China: Text Mining of Narratives on Social Media.* The Journal of Real Estate Finance and Economics. 66. 10.1007/s11146-022-09900-5