SOCIAL MOOD ON ECONOMY INDEX

Una misura del sentiment italiano sull'economia basata sui dati di Twitter

Milioni di italiani usano quotidianamente i social media per informarsi, esprimere il proprio stato d'animo, condividere le proprie opinioni e dibattere gli argomenti più disparati. Ciò configura i social media come uno degli strumenti più promettenti per "misurare" lo stato d'animo degli italiani e fornisce una solida motivazione per la sperimentazione di tecniche di sentiment analysis da parte dell'Istat[1].

Negli ultimi anni l'Istat ha studiato se i messaggi pubblici in lingua italiana disponibili sui social media possano essere sfruttati con successo per sviluppare indici di sentiment *specifici*, cioè strumenti statistici in grado di valutare lo stato d'animo degli italiani su argomenti o aspetti della vita ben definiti, come ad esempio la situazione economica.

A tale scopo sono state sviluppate procedure che selezionano ed elaborano unicamente messaggi il cui testo contenga almeno una parola appartenente ad un determinato *filtro*, vale a dire ad uno specifico insieme di parole chiave rilevanti. Il filtro utilizzato dall'Istat è stato predisposto da esperti di dominio con l'intento di eliminare sin dall'inizio messaggi che, qualora venissero catturati, si rivelerebbero con elevata probabilità non pertinenti rispetto all'obiettivo di produzione statistica previsto.

L'Istat rende disponibile un nuovo indice sperimentale, basato sui dati di Twitter: il Social Mood on Economy Index. L'indice fornisce misure giornaliere del sentiment italiano sull'economia, derivate da campioni di tweet pubblici in lingua italiana, catturati in tempo reale. Il periodo di osservazione coperto dall'indice parte dal 10 febbraio 2016.

Nel corso del 2021 la serie storica ha superato i 5 anni ed in virtù del rapido e crescente aggiornamento delle tecniche di sentiment analysis si è ritenuto opportuno apportare una revisione metodologica, nel seguito illustrata, volta ad aumentare la qualità ed il potere informativo dell'indice.

La serie storica è stata ricalcolata con la nuova modalità

Procedura di produzione dell'indice e metodi utilizzati

La procedura di produzione dell'indice usa la Streaming API di Twitter per raccogliere campioni di tweet che soddisfino un filtro appositamente progettato, costituito da 60 parole chiave (parole o locuzioni). Tali parole chiave sono state prevalentemente derivate dal questionario

[1] Il presente lavoro è stato realizzato nell'ambito del progetto PSN IST-02589 "Uso a fini statistici dei Big Data" inserito nel Programma Statistico Nazionale 2017-2019, e PSN IST-02807 "Statistiche con uso di fonti Big Data", inserito nel Programma Statistico Nazionale 2020-2022

dell'indagine sulla fiducia dei consumatori, una rilevazione mensile che raccoglie dati nelle prime due settimane di ciascun mese e diffonde le proprie stime entro la fine del mese. A tale proposito, occorre precisare che il fenomeno misurato dal Social Mood on Economy Index è più vasto e molto meno definito di quanto non sia la fiducia dei consumatori, la cui misura ufficiale è basata su una metodologia armonizzata e condivisa a livello europeo, nonché caratterizzata da una lunga e rilevante tradizione di utilizzo a fini di analisi congiunturale e di previsione.

A partire dalla diffusione del quarto trimestre 2021 è stato introdotto un filtro di secondo livello più specifico che esclude dal campione alcuni messaggi raccolti dal filtro di primo livello. Più in dettaglio solo alcuni tweet, a partire dal filtro precedente di 60 parole chiave, entrano nel campione. Il filtro di secondo livello contiene 115 parole chiave (parole o locuzioni) ed è un raffinamento del filtro di primo livello: ad esempio la parola "benessere" è stata declinata in più locuzioni specifiche. Si sottolinea che non sono state introdotte nuove parole e che il filtro introdotto risulta essere un sottoinsieme stretto di quello utilizzato precedentemente, pertanto si tratta di un sotto-campione di quello precedentemente utilizzato

Questo raffinamento ha permesso di individuare con maggiore probabilità tweet di contenuto specificamente economico, anche se il volume di tweet selezionati ed analizzati diminuisce sensibilmente.

La procedura di calcolo dell'indice giornaliero elabora tutti i tweet del campione (con la procedura attuale, circa 26.000 giornalieri) in un unico blocco (si veda la Figura 1). I messaggi vengono prima puliti e normalizzati, quindi analizzati con tecniche di sentiment analysis. Il metodo di sentiment analysis adottato è non supervisionato e si basa sull'uso di un *lexicon di sentiment* in lingua italiana, vale a dire un vocabolario ai cui lemmi sono associati punteggi di sentiment positivo e negativo pre-calcolati. A partire dalla diffusione del quarto trimestre 2021 il *lexicon di sentiment* in italiano è stato sostituito con uno più recente che tiene conto anche di alcuni neologismi legati alla pandemia del covid-19 e che permette di avere una copertura linguistica oltreché triplicata rispetto alla precedente.

I testi di tutti i tweet vengono confrontati con il lexicon: sulla base dei punteggi delle parole abbinate, a ciascun tweet vengono assegnati punteggi di sentiment positivo e negativo. I punteggi di sentiment dei messaggi vengono successivamente utilizzati da un algoritmo di clustering che partiziona i tweet del giorno in tre classi disgiunte: tweet negativi, tweet neutri e tweet positivi. Il valore dell'indice giornaliero viene infine ricavato applicando un'opportuna misura di tendenza centrale alla distribuzione dei punteggi di tutti i tweet. Nella precedente procedura dato che il vocabolario non forniva un punteggio di sentiment neutro, il clustering era imprescindibile per la classificazione dei tweet neutri e questi ultimi venivano esclusi dal calcolo dell'indice. I lemmi dell'attuale vocabolario hanno punteggi di sentiment positivo, negativo e neutro pertanto queste operazioni non sono necessarie. Inoltre essendosi ridotti i volumi dei tweet filtrati, quest'operazione avrebbe ulteriormente impoverito il campo d'osservazione. La formula di calcolo dell'indice è rimasta invariata.

Come passo conclusivo, l'indice giornaliero viene trasformato linearmente in modo che la sua serie storica abbia media nulla nel periodo 10 febbraio 2016 – 30 settembre 2018.

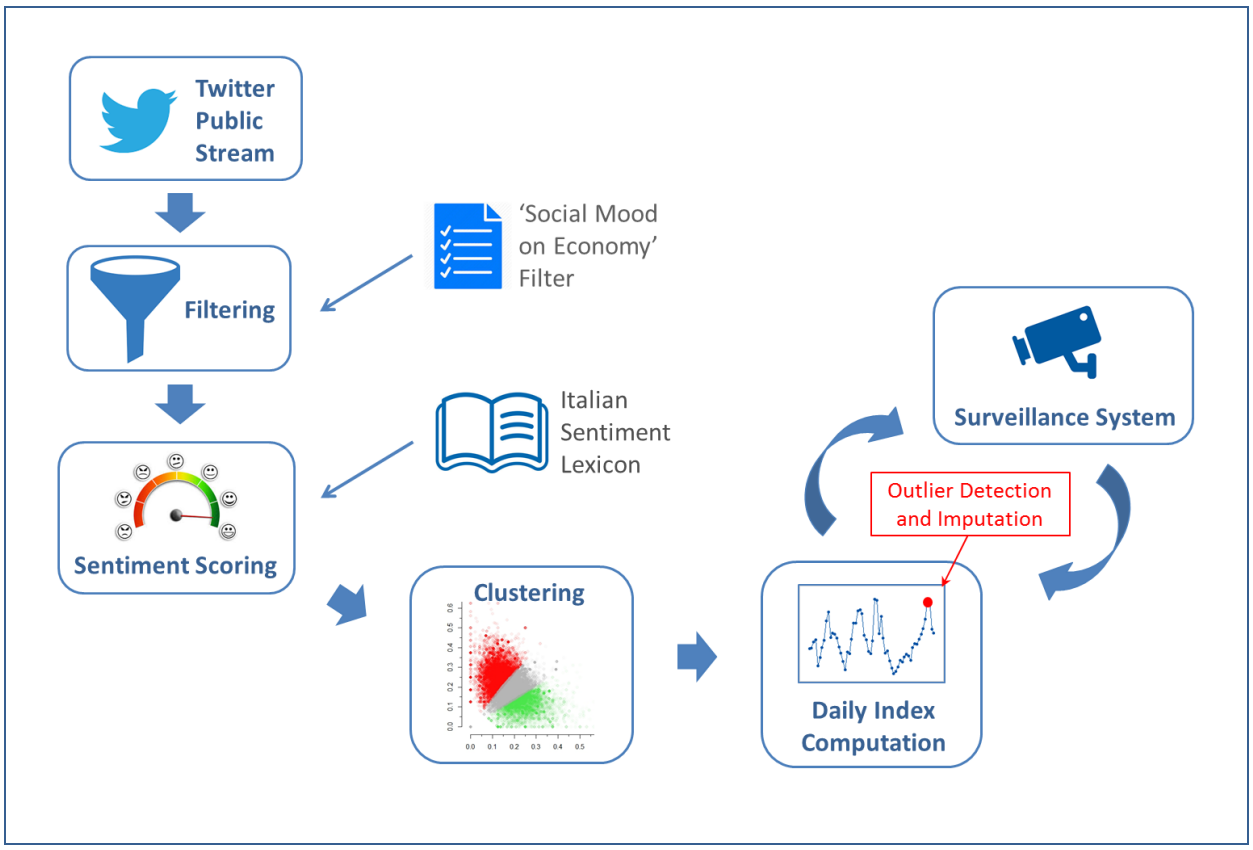

Figura 1: Schema della procedura di produzione del Social Mood on Economy Index

È stata prestata particolare attenzione a rendere l'indice robusto rispetto ad eventuali contaminazioni derivanti da tweet fuori tema che possano aver eluso il filtro. A tale scopo è stato realizzato un sistema di sorveglianza che analizza periodicamente la serie storica giornaliera dell'indice alla ricerca di valori anomali. Il sistema utilizza due metodi di outlier detection indipendenti e complementari. L'identificazione di un potenziale outlier determina la generazione automatica di report diagnostici, i quali vengono inviati a revisori esperti incaricati di stabilire se il valore segnalato costituisca effettivamente un'anomalia. Le anomalie più gravi sono tipicamente causate da tweet fuori tema che abbiano eluso il filtro e siano diventati "virali" su Twitter. I tweet virali possono essere retwittati e citati anche centinaia di migliaia di volte in un solo giorno, determinando gravi effetti distorsivi sul valore dell'indice. La procedura esclude i tweet retwittati, ma mantiene quelli citati e commentati. Nonostante ciò, alcuni valori giornalieri dell'indice giudicati anomali ed irrimediabilmente contaminati vengono sostituiti da valori interpolati.

Rappresentatività

Come accade per la maggior parte delle sorgenti Big Data, il meccanismo di generazione dei dati di Twitter non ricade sotto il diretto controllo dello statistico e non è noto. Ciò distingue radicalmente il Social Mood on Economy Index dai consueti output statistici derivati sia dalle tradizionali indagini campionarie, il cui disegno di campionamento è progettato e controllato dagli istituti di statistica, sia dalle fonti amministrative, il cui meccanismo di generazione dei dati è solitamente noto ai progettisti delle rilevazioni. Per conseguenza, non esiste – ad oggi – alcuna

metodologia rigorosa che consenta di garantire la generale validità delle informazioni statistiche derivate dai dati di Twitter. In particolare, gli utenti italiani di Twitter[2] non possono essere considerati un campione rappresentativo della popolazione italiana. Ne consegue che l'Istat non può garantire l'accuratezza di eventuali inferenze derivate dal Social Mood on Economy Index sulla popolazione italiana.

Privacy

Il Social Mood on Economy Index è alimentato unicamente da campioni di tweet pubblici[3]. La cattura dei tweet avviene senza selezionare specifici utenti di Twitter e senza tracciare l'attività di alcun utente. L'indice usa esclusivamente dati anonimizzati e non linkati: i messaggi elaborati non sono mai associati ai rispettivi autori, i quali sono ignoti all'Istat. La procedura di calcolo dell'indice elabora unicamente il contenuto testuale dei tweet raccolti. I valori giornalieri dell'indice sono il risultato dell'aggregazione di punteggi di sentiment numerici associati a decine di migliaia di messaggi. Si tratta di un processo irreversibile: in nessun caso il testo di un tweet potrà mai essere ricostruito analizzando i valori dell'indice.

Caratteristiche della serie storica

Di seguito vengono presentati i risultati del trattamento dell'indice giornaliero grezzo, allo scopo di individuare le componenti non osservabili della serie al fine di una valutazione più chiara della sua dinamica. La frequenza giornaliera del Social Mood on Economy Index ha comportato, per l'individuazione della componente stagionale, l'utilizzo di una metodologia differente dallo standard Istat che si basa sulla procedura model-based implementata in TRAMO-SEATS.

Le serie storiche giornaliere inducono ad affrontare, infatti, problematiche metodologiche diverse rispetto a quelle implementate per serie mensili o trimestrali a causa della presenza di stagionalità multiple (infrasettimanali, settimanali, mensili e/o annuali). Tuttavia l'analisi delle serie a frequenza giornaliera, rispetto alla loro aggregazione mensile, permette di conservare le caratteristiche di tempestività dell'informazione e garantisce una migliore individuazione degli effetti dei giorni lavorativi e di altri eventi giornalieri che influenzano l'andamento della serie.

Tra i recenti approcci adottati per la destagionalizzazione di serie ad alta frequenza è stata scelta la metodologia descritta in De Livera, A. M., Hyndman, R. J., Snyder, R. D. (2011) che applica modelli di *exponential smoothing* modificati rispetto ai modelli standard introdotti da Holt e Winters. La metodologia si basa sull'ipotesi che ogni serie storica sia rappresentabile come combinazione di diverse componenti: una componente che descrive il livello della serie (*level*), una per il tasso di crescita (*slope*), una che coglie i movimenti stagionali e, infine, una irregolare.

[2] Si stima che a fine 2018 gli utenti italiani di Twitter attivi su base mensile fossero circa 9 milioni, e 10 nel 2021(fonte: Agcom).

[3] I tweet *pubblici* (impostazione predefinita della piattaforma Twitter) sono immediatamente visibili a chiunque, anche a chi non possiede un account Twitter. Al contrario, i tweet *protetti* sono visibili unicamente ai follower dell'utente autore del tweet.

Per l'individuazione di tali componenti il modello è quindi rappresentato in una forma *state space* i cui parametri vengono stimati tramite la funzione *tbats* del pacchetto R *forecast*.

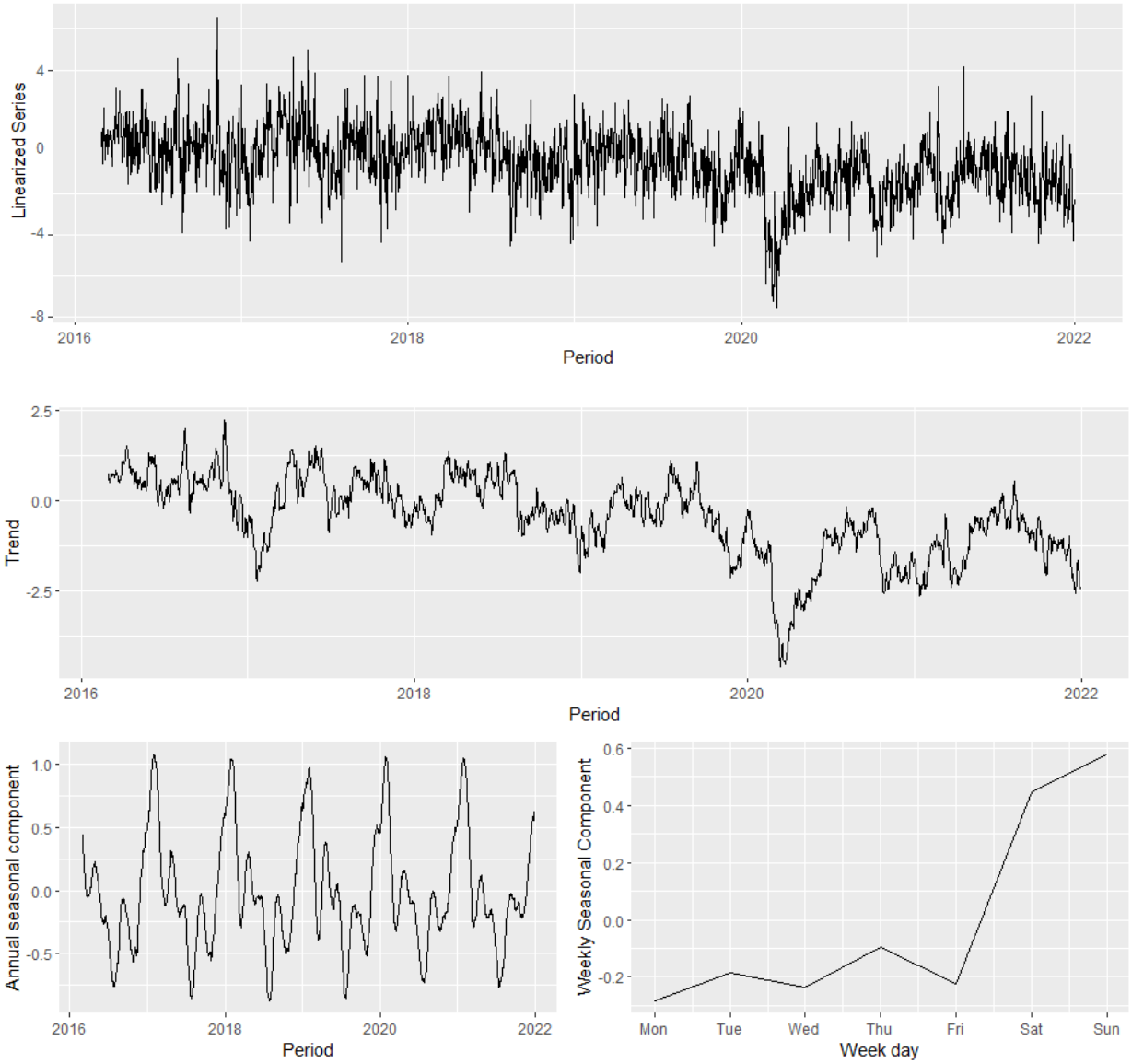

Figura 2: Decomposizione della serie storica giornaliera del Social Mood on Economy Index

In particolare, per la serie giornaliera del Social Mood on Economy Index la procedura di destagionalizzazione è stata implementata in due fasi: identificazione e stima degli effetti deterministici mediante l'introduzione di opportune variabili dummy che porta all'individuazione della serie linearizzata di cui poi nella seconda fase vengono stimate le diverse componenti. Il periodo di riferimento è dato dall'intervallo dall'1 marzo 2016 fino all'ultimo giorno disponibile (non considerando, dunque, i dati incompleti raccolti nel mese di febbraio 2016).

I risultati della destagionalizzazione sono illustrati in Figura 2, in cui sono rappresentate la componente linearizzata e di trend e le due componenti stagionali rilevate durante il processo di destagionalizzazione.

Principali revisioni osservate

Nonostante la procedura di produzione dell'indice sia rimasta invariata, le modifiche introdotte nel processo di produzione, ovvero il raffinamento del filtro e il cambio di lexicon hanno un forte impatto sulla dinamica della serie grezza giornaliera. Da un lato la modifica del campo d'osservazione dei tweet ha prodotto un cambio nell'annotazione dei punti corrispondenti ai principali picchi ed alle principali valli dell'indice, rendendo maggiormente interpretabili questi eventi da un punto di vista economico. Dall' altro lato restano osservabili vari fenomeni come quelli legati a terremoti o a particolari leggi come "opzione donna", a conferma della robustezza della procedura, nonostante le revisioni introdotte sia dal filtro sia dal lexicon. Il cambio del lexicon modifica strutturalmente la serie grezza riducendo sensibilmente la volatilità della stessa. Infine anche la dinamica sia grezza sia destagionalizzata si differenzia profondamente e gli effetti più evidenti si vedono a partire da inizio marzo 2020 ovvero l'inizio della pandemia, dove l'indice attuale mostra un significativo crollo non presente nel precedente indice.

Alla produzione ed all'analisi del Social Mood on Economy Index hanno collaborato:

- M. Bruno, E. Catanese, R. Iannaccone, A. Righi, M. Scannapieco, P. Testa, L. Valentino, D. Zardetto, D. Zurlo.