L'IMMAGINE DELL'IMMIGRAZIONE VEICOLATA DA X (TWITTER)

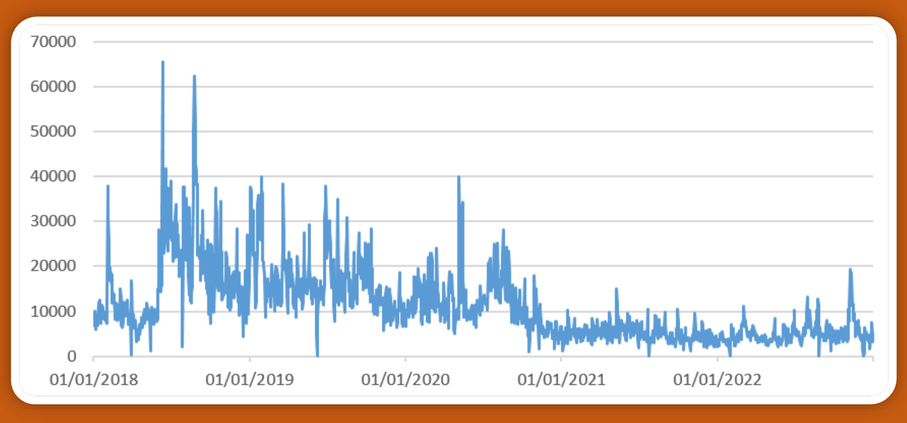
Conferenza Nazionale di Statistica

MAURO BRUNO – Istat mbruno@istat.it | **ELENA CATANESE -** Istat catanese@istat.it | **GERARDA GRIPPO -** Istat grippo@istat.it |
FRANCESCO ORTAME – Istat francesco.ortame@istat.it | **MARIA CLELIA ROMANO -** Istat romano@istat.it

Le nuove fonti di dati, in particolare le piattaforme di social network, dove i cittadini possono esprimere liberamente e pubblicamente le loro opinioni, possono rappresentare un complemento prezioso alle informazioni provenienti da fonti tradizionali. Anche per la statistica ufficiale è importante investigare gli usi possibili delle nuove fonti, al fine di arricchire la produzione statistica e aumentarne la tempestività. Questo lavoro si inserisce in un più ampio progetto di *Trusted Smart Statistics*, che mira a esplorare i contenuti veicolati da Twitter sul tema dell'immigrazione, analizzandoli mediante alcune delle principali metodologie di *machine learning*. I primi risultati mettono in evidenza l'andamento nel tempo dell'interesse per il tema, fortemente segnato dal dibattito politico, il differente contenuto semantico delle parole più frequenti e l'eterogeneità delle dimensioni messe in risalto dalla clusterizzazione.

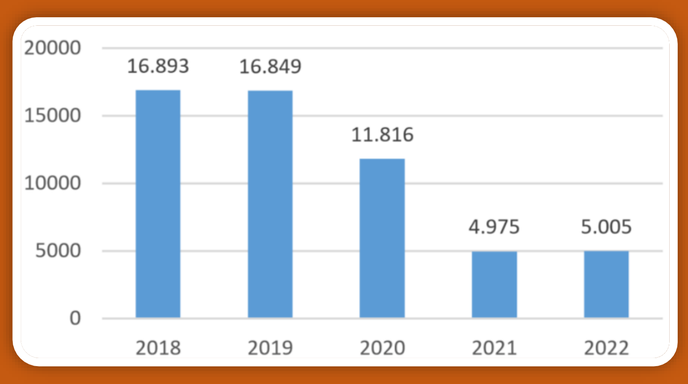
IL CORPUS DI ANALISI

Dal 2016 l'istat cattura in tempo reale campioni di tweet (ora posts) pubblici in lingua italiana su tematiche di rilevanza per la statistica ufficiale. Per lo studio di temi specifici è opportuno effettuare una selezione dei testi, in modo da individuare i contenuti di interesse e focalizzare l'analisi su di essi. In particolare per quanto riguarda il tema oggetto di studio (immigrazione) applicando un filtro di 21 *keywords* è stato ridotto significativamente il volume del corpus, in modo da analizzare solo le conversazioni pertinenti. Il periodo considerato è il quinquennio 2018-2022 e l'ammontare dei tweet, contenenti almeno una delle parole filtro utilizzate, è pari a oltre 20 milioni (20.283.772), circa il 4,4% del complesso dei tweet raccolti da Istat.



Volumi tweet giornalieri (2018-2022)

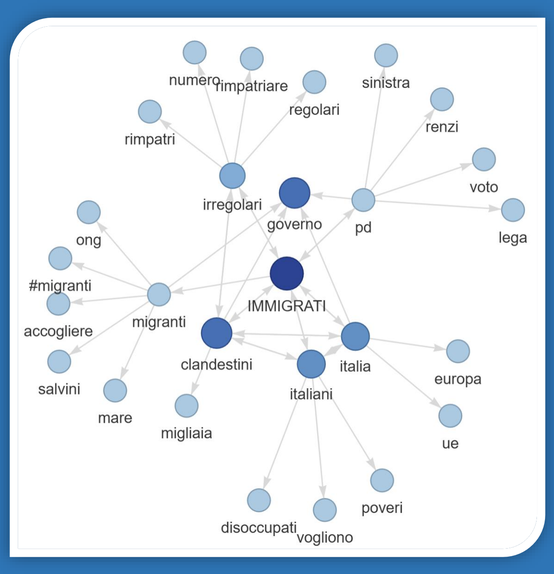
L'andamento nel tempo dei volumi è variabile e presenta un evidente trend decrescente, a indicare un calo, negli anni successivi al 2020, delle conversazioni riguardanti l'immigrazione nel dibattito su Twitter. I picchi sono perlopiù associabili a fatti di cronaca, a episodi di nuovi sbarchi o interventi normativi sul tema Le medie di tweet raccolti giornalmente per i diversi anni sono mostrate di seguito.



Media volumi tweet giornalieri per anno

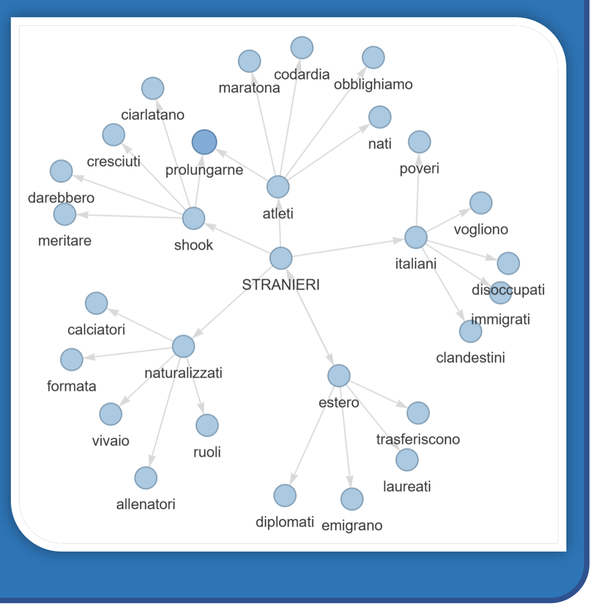
UN ESEMPIO DI WORD EMBEDDING

I word embedding sono tecniche *di machine learning* per estrarre rappresentazioni numeriche delle parole che catturano i loro significati contestuali, consentendo di identificare relazioni semantiche tra di esse. In questo lavoro, sono stati utilizzati per analizzare le principali associazioni di alcune delle *keywords* scelte per la definizione del database di analisi. I risultati mettono in luce la capacità di questa metodologia di anticipare i contenuti semantici che verranno meglio specificati nella topic analysis e di contribuire alla loro interpretazione.
A seguire, si riportano due applicazioni esemplificative della eterogeneità semantica associata a due delle *keyword* utilizzate: "immigrati" e "stranieri".



Immigrati
Focalizzando l'attenzione sul termine "immigrati", le parole più fortemente associate ad esso rimandano ad alcune delle principali problematiche nella gestione dei flussi migratori. Escludendo le parole relative a soggetti politici, il termine "immigrati" richiama perlopiù la componente clandestina dei flussi nei suoi vari aspetti: numerosità, rimpatri, accoglienza, canali di arrivo (mare, ONG). Anche l'associazione con "Italia" e "italiani" mette in risalto la critciità del tema rispetto alle problematiche dei cittadini (disoccupazione, povertà)

Stranieri
Diversamente, la parola "stranieri" si associa a termini molto eterogenei e di valore mediamente più neutro. Il tema dell'immigrazione clandestina è solo una delle componenti semantiche associate Va segnalata, infatti, l'associazione col tema dello sport, in generale, e in particolare del calcio, degli scambi universitari (Erasmus), così come della fuga di cervelli all'estero.

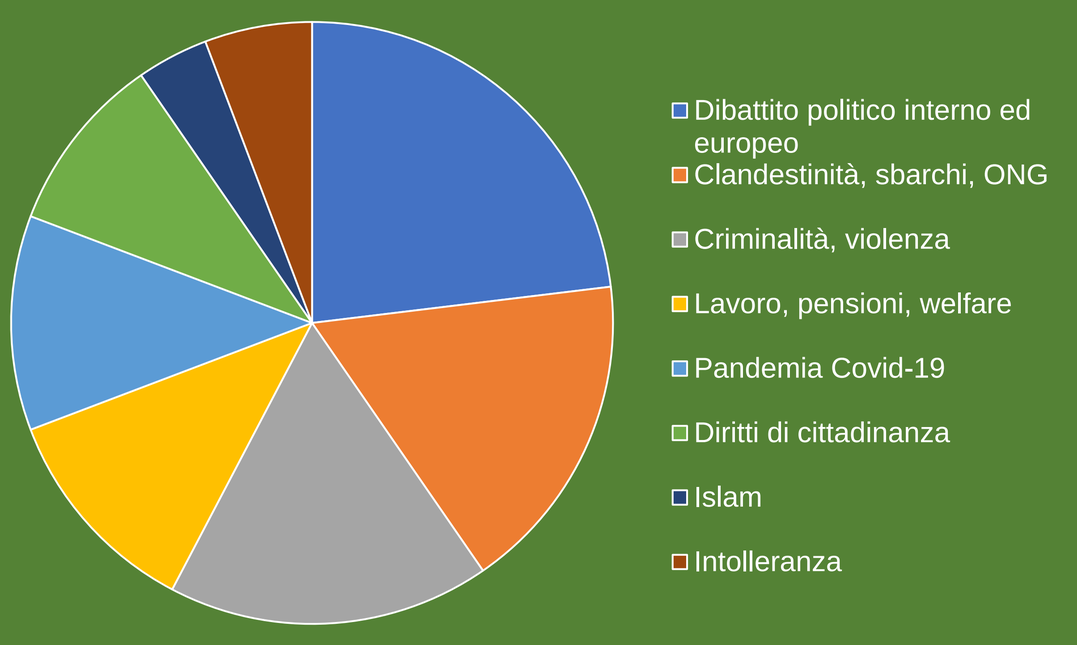


L'APPLICAZIONE DELLA TOPIC ANALYSIS

Per comprendere meglio i contenuti semantici dei tweet è stata applicata una topic analysis, che ha permesso di individuare i principali temi latenti presenti nei tweet selezionati. La tecnica utilizzata è la Latent Dirichlet Allocation (LDA), proposta da David Blei, Andrew Ng e Michael Jordan nel 2003.
A valle di analisi e ottimizzazioni, si è definito un numero di cluster (temi) pari a 75. Da questi raggruppamenti emerge una grande eterogeneità di argomenti, che riflette la complessità del dibattito pubblico. L'analisi semantica consente, tuttavia, di ricondurli ad alcuni macro-temi e di individuarne i rispettivi volumi. Il prospetto seguente sintetizza i principali macro-temi individuati, i cluster ad essi afferenti e la percentuale di tweet contenuti.

Macrotema	Cluster	% Tweet
Dibattito politico interno ed europeo	12	16,5
Clandestinità, sbarchi, ONG	9	14,3
Criminalità, violenza	9	13,1
Lavoro, pensioni, welfare	6	9,7
Pandemia Covid-19	6	8,8
Diritti di cittadinanza	5	5,9
Islam	2	3,8
Intolleranza	3	3,3
Altro*	23	24,6

*La categoria *Altro* include topic residuali e/o difficilmente interpretabili. Ciò è tipico della topic analysis, che accanto a cluster interpretabili e ben definiti tende a estrarre alcuni cluster non pertinenti.



Macro-aree non residuali

La macro-area che assorbe il numero maggiore di tweet (16.5%) riguarda il dibattito politico interno ed europeo in tema di immigrazione. I termini più caratterizzanti rimandano ai principali attori politici, alle misure politiche adottate in tema di immigrazione e alle preoccupazioni generate dall'arrivo di migranti clandestini. A titolo di esempio, si riporta, a sinistra, la rappresentazione delle parole più frequenti di un topic appartenente alla macro-area.

Dibattito politico interno e internazionale

Il secondo raggruppamento per numero di tweet (14.3%) riguarda 9 cluster caratterizzati dal tema degli sbarchi. Sebbene declinato in maniera eterogenea nei vari topic, il tema degli sbarchi ne rappresenta il filo conduttore. Parole come ONG, centro, clandestini, acque, mare, navi (Open Arms, Viking Ocean, Sea-Watch), Mediterraneo, guardia costiera, morti, accoglienza e Lampedusa caratterizzano i cluster di questa macro-area. A titolo di esempio, si riporta, a sinistra, la rappresentazione delle parole più frequenti di un topic appartenente alla macro-area.

Clandestinità, sbarchi, ONG