



## DOEM-UNIBO: Un dataset su *Otto e mezzo*, un talk show politico italiano

**Tommaso Aicardi**  
Bocconi University

**Chloe Papadopoulou**  
University of Amsterdam

**Marco Albertini**  
University of Bologna

---

### Abstract

L'articolo presenta DOEM, un nuovo dataset longitudinale per lo studio della comunicazione politica televisiva in Italia. Il dataset raccoglie le trascrizioni complete di 2.889 episodi del talk show politico *Otto e mezzo*, trasmesso su La7 tra il 4 aprile 2011 e il 10 giugno 2025. Le trascrizioni sono state prodotte a partire dall'audio degli episodi tramite modelli di automatic speech recognition basati su Whisper-AI e successivamente sottoposte a procedure di pulizia, controllo e normalizzazione. Oltre al testo completo delle puntate, DOEM include metadati sulla data di trasmissione e sugli ospiti presenti. Il paper illustra la struttura del dataset e presenta alcuni esempi di utilizzo per l'analisi della copertura mediatica di eventi politici, dell'attenzione verso specifiche policy issues e dello stile comunicativo degli attori politici. DOEM costituisce una risorsa originale per studi di comunicazione politica, stile della leadership, agenda setting e, più in generale, promuove l'utilizzo di approcci computazionali e di "text as data" negli studi della comunicazione politica in Italia.

*Keywords:* Political discourse, Political communication, Text as data, Computational social science, Natural language processing, Political talk shows, Television, Dataset, Italian politics.

---

### 1. Introduzione al dataset

*Otto e mezzo* è un programma di approfondimento politico che va in onda sull'emittente televisiva italiana La7, dal lunedì al venerdì, alle 20.30. Ogni stagione del programma inizia a settembre e finisce tra la fine di maggio e la fine di giugno. Ciascuna puntata del programma dura circa trenta minuti. Il programma è condotto dalla giornalista Dietlinde (Lilli) Gruber. Il format consiste in un dibattito mediato tra vari ospiti (alcuni fissi o ricorrenti, altri specifici per il tema affrontato), su temi di attualità politica. Le personalità presenti all'interno del programma variano tra giornalisti, opinionisti politici e rappresentanti dei partiti politici nazionali, della società civile e del mondo accademico. All'interno del format è presente

anche un breve servizio di approfondimento, redatto dal giornalista Paolo Pagliaro.

Il dataset DOEM contiene le trascrizioni complete delle tracce audio di quasi tutti gli episodi di *Otto e mezzo* andati in onda tra il 4 aprile 2011 e il 10 giugno 2025. Il corpus viene distribuito in formato ".csv" o ".parquet" e contiene 2889 osservazioni. Ogni osservazione contiene le seguenti variabili: la data di messa in onda dell'episodio, la trascrizione completa dell'episodio e la lista degli ospiti presenti.

A scopo di esempio, di seguito si possono leggere le prime battute dell'episodio andato in onda il 18 maggio 2017:

Buonasera, benvenuti a 8.30, buonasera benvenuto Romano Prodi, grazie per aver finalmente accolto il nostro invito. Finalmente, sono venuto! Adesso poi cerchiamo di capire anche perché. Buonasera, bentornato Marco Damilano, vice direttore dell'Espresso. Buonasera. Ecco, uno dei motivi per cui Prodi è venuto è perché ha scritto un libro manifesto politico che potrebbe essere benissimo un programma di governo. Subito dopo la pubblicità.

Come mostra il frammento qui sopra, la qualità delle trascrizioni è generalmente molto elevata. Questo risultato è reso possibile grazie all'utilizzo del modello di riconoscimento vocale e di conversione audio-testo chiamato Whisper-AI (Radford et al. 2023). Non tutti gli episodi sono trascritti con la precisione di questo frammento. In alcuni casi ci sono errori di trascrizione, come ripetizioni, errori di ortografia o di punteggiatura. In casi estremi la trascrizione è totalmente mancante. Questi problemi sono stati risolti con operazioni di pulizia del corpus esposte in seguito.

Nella prossima sezione viene riassunto il processo, insieme alle finalità e alle tecniche utilizzate, che ha portato alla costruzione del dataset. In quella successiva sono presentati i risultati dell'esplorazione del dataset e vengono esposti alcuni esempi di utilizzo.

## 2. Costruzione del dataset

Il dataset nasce nel contesto di una ricerca che mirava a studiare aspetti e contenuti specifici della comunicazione politica nel medium televisivo, con un interesse particolare per l'andamento temporale della rilevanza di alcuni temi (ad esempio, povertà, immigrazione), per i nuclei tematici ad essi connessi e per il sentiment associato alla loro trattazione nella discussione politica pubblica. Ai fini del progetto si è reso necessario costruire un dataset testuale, analizzabile attraverso le tecniche del natural language processing, a partire da contenuti audio-video.

La scelta è ricaduta sul programma *Otto e mezzo* per vari motivi: 1) l'emittente La7 pubblica gran parte delle registrazioni integrali dei suoi programmi televisivi online dal 2011, offrendo una quantità di contenuti molto elevata; 2) il programma *Otto e mezzo* va in onda senza interruzioni e con lo stesso format, conduttori e durata da almeno 11 anni. Questa continuità e omogeneità sono molto utili dal punto di vista teorico, in quanto permettono di ipotizzare che le variazioni nella struttura del programma abbiano avuto poco impatto sul contenuto dello stesso. Sostanzialmente l'esistenza e la disponibilità di questa fonte di informazione consentono di sfruttare, ai fini di ricerca, la caratteristica "always on" tipica delle tracce digitali (Salganik 2017). Tale caratteristica rende possibile, ad esempio, analizzare i cambiamenti associati a eventi esogeni al programma, avvenuti nel periodo di osservazione,

quali l'emergere della pandemia legata a Covid-19 o l'esplosione della guerra in Ucraina o in Palestina, per citare solo alcuni degli eventi più noti e recenti; 3) la durata relativamente breve del programma ha permesso di limitare gli errori degli strumenti di trascrizione automatica come Whisper-AI, che presentano significative difficoltà nell'acquisizione di contenuti audio eccessivamente lunghi.

I primi passi per la creazione del dataset sono stati la raccolta delle tracce audio e la loro trascrizione con l'utilizzo di Whisper-AI. Whisper-AI è un modello di riconoscimento vocale in grado di acquisire input audio e di trascriverli. Esso è in grado di produrre documenti testuali molto accurati, comprensivi di punteggiatura, riconoscimento di nomi propri e utilizzo di lettere maiuscole. La lingua italiana è tra quelle per cui il modello raggiunge i livelli di accuratezza più elevati, con una frequenza di errore di trascrizione delle parole stimata attorno al 3-5%. In ragione della mole di documenti audio da trascrivere è stata usata una derivazione del modello originale chiamata Faster Whisper ([SYSTRAN 2023](#)), in grado di trascrivere un episodio di *Otto e mezzo* in circa due minuti, mantenendo gli stessi livelli di fedeltà. Tutto il lavoro è stato svolto su Google Colab.

Le trascrizioni generate automaticamente presentavano alcuni problemi ricorrenti che rendevano necessario un processo di pulizia del corpus prima della costruzione della release finale. Un primo problema riguardava la punteggiatura non sempre coerente. Un secondo problema era rappresentato da artefatti tecnici non sostantivi, cioè stringhe chiaramente estranee al contenuto editoriale della puntata, come riferimenti residuali a sottotitoli o piattaforme di distribuzione. Un terzo problema consisteva nelle ripetizioni spurie e nelle troncature del discorso introdotte dalla trascrizione automatica: in alcuni episodi comparivano righe duplicate, blocchi molto simili ripetuti, oppure code finali in cui una stessa frase o formula di chiusura veniva reiterata più volte. Infine, il corpus presentava anche errori residui nei nomi propri, sotto forma di grafie rumorose o incoerenti degli stessi nomi e cognomi, soprattutto per ospiti ricorrenti o figure politiche frequentemente citate. La risoluzione di questi problemi è stata effettuata combinando regole locali di correzione, procedure assistite da AI e revisione manuale mirata dei casi residui.

Per ogni episodio è stata inoltre costruita una variabile dedicata ai presenti in puntata, definita come la lista delle persone effettivamente presenti nello studio o in collegamento come ospiti o interlocutori. La costruzione di questa variabile non si è basata sull'intero testo della trascrizione, ma sull'estratto iniziale di ciascun episodio, poiché nelle puntate di *Otto e mezzo* l'apertura contiene normalmente i saluti, la presentazione degli ospiti e le prime interazioni in studio. Da questo estratto sono stati identificati i nomi delle persone effettivamente presenti in ogni puntata attraverso un prompt dedicato somministrato al modello GPT-5.4 mini. Le liste così ottenute sono state poi corrette e uniformate attraverso regole locali e controlli mirati sui casi residui più ambigui o atipici.

### 3. Esplorazione del dataset ed esempi di utilizzo

Come riportato in precedenza, il dataset contiene 2889 episodi andati in onda tra aprile 2011 e giugno 2025. Il numero degli episodi per anno presenti nel dataset varia tra i 175 e i 215. Fanno eccezione, ovviamente, l'anno iniziale e quello finale, per cui il periodo di osservazione copre solo alcuni mesi (Figura 1). La variabilità tra gli anni nel numero di episodi è data sia dal fatto che ogni stagione di *Otto e mezzo* contiene un numero diverso di episodi, sia perché nel dataset non sono presenti alcuni episodi, eliminati a causa di errori di trascrizione che non

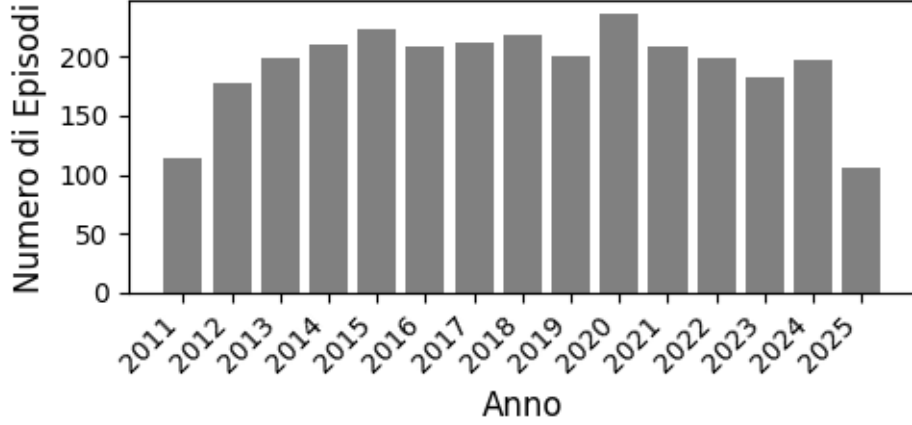


Figura 1: Numero di episodi di *Otto e mezzo* per anno.

Tabella 1: Top 10 degli ospiti più presenti nel dataset

Nome	Frequenza
Marco Travaglio	411
Andrea Scanzi	313
Beppe Severgnini	296
Massimo Giannini	268
Lucio Caracciolo	214
Alessandro Sallusti	204
Antonio Padellaro	197
Paolo Mieli	164
Lina Palmerini	157
Massimo Cacciari	156

li rendevano fruibili. Il dato degli episodi per anno mostra, tuttavia, che il dataset copre la quasi totalità delle puntate andate in onda nei 15 anni di osservazione.

Come esposto nella sezione precedente, ogni osservazione contiene la lista degli ospiti presenti. Da questa variabile possiamo estrarre informazioni come il dato sui 10 ospiti più presenti nel programma nei 15 anni osservati (Tabella 1). I principali nomi sono quelli di giornalisti tra i più noti al grande pubblico italiano, come Marco Travaglio e Andrea Scanzi del *Fatto Quotidiano*, Massimo Giannini di *Repubblica* o Beppe Severgnini del *Corriere della Sera*.

Il dataset offre molte opportunità di analisi del contesto politico italiano, della comunicazione politica e pubblica e, in generale, della comunicazione televisiva. Ad esempio, la Figura 2, prendendo in considerazione i nomi degli ultimi otto Primi Ministri del governo italiano, mostra la frequenza relativa con cui essi vengono nominati nel corso di ciascun semestre di trasmissione considerato. Si può notare, ad esempio, come Enrico Letta, Primo Ministro dal 28 aprile 2013 al 22 febbraio 2014, venga citato prevalentemente nel periodo in cui è in carica, mentre Paolo Gentiloni, Premier dal 12 dicembre 2016 al 1 giugno 2018, resti, anche nel suo periodo di premiato, una figura marginale. Diversamente, ci sono personaggi

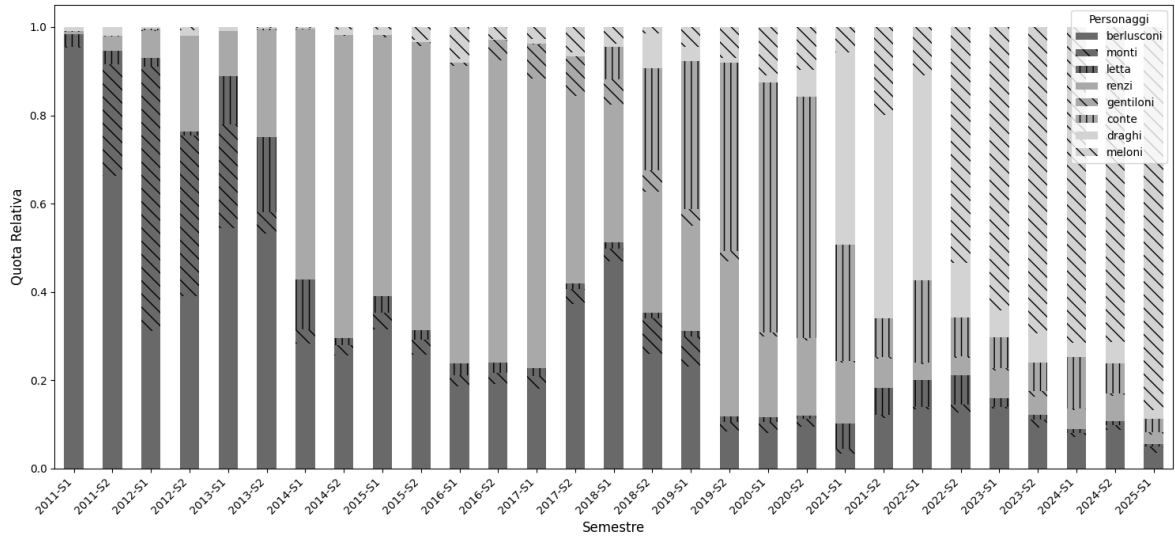


Figura 2: Frequenza dei nomi degli ultimi 8 Primi Ministri nel dataset, in termini relativi, per semestre.

politici, come ad esempio Matteo Renzi (Presidente del Consiglio dal 22 febbraio 2014 al 12 dicembre 2016), che vengono frequentemente nominati durante la trasmissione anche ben oltre la durata della loro carica di governo. Queste e altre analisi possono chiaramente informare gli studi dei fenomeni e delle dinamiche politiche in Italia, o almeno quegli studi interessati alla rappresentazione mediatica delle dinamiche, dei processi e delle discussioni politiche e pubbliche.

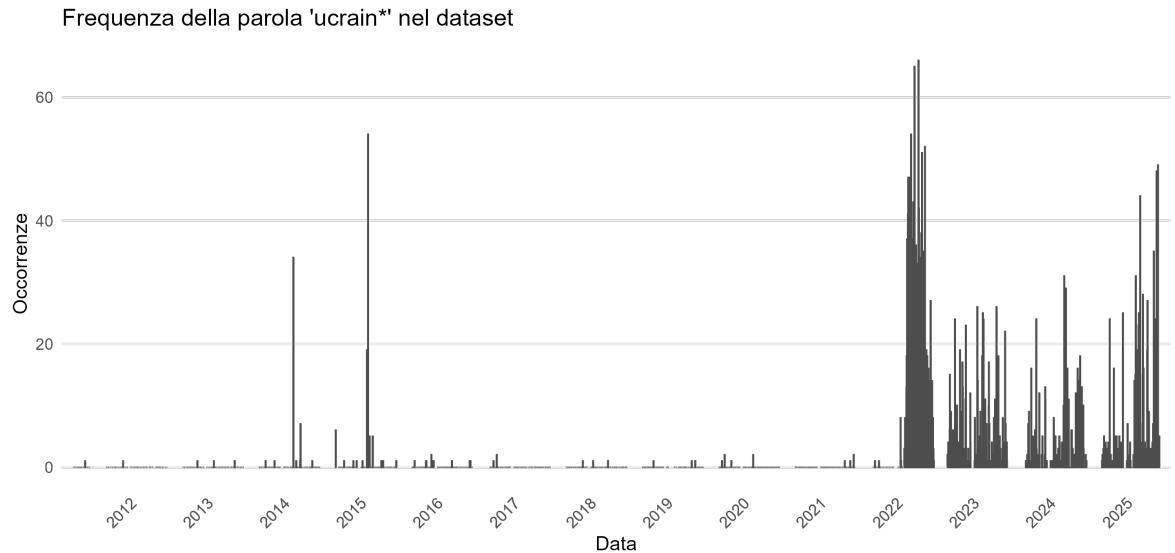


Figura 3: Frequenza delle parole aventi come radice 'ucrain\*', all'interno del corpus, tra il 2011 e il 2025.

Il dataset offre anche le risorse per analizzare, oltre ai personaggi, temi specifici. Ad esem-

pio, si può osservare la distribuzione delle occorrenze di una parola nel corso del tempo, così come, attraverso modelli di topic modelling, i temi principali associati al suo utilizzo all'interno della trasmissione. Il caso delle parole riferite all'Ucraina (Figura 3) in questo senso è interessante, in quanto permette di analizzare l'interesse mediatico relativo agli eventi avvenuti nella regione nell'ultimo decennio. Analizzando il grafico si può notare che un primo aumento dell'attenzione verso le questioni ucraine avviene a inizio 2014, in corrispondenza della puntata datata 20 febbraio 2014. In questa puntata vengono trattate estensivamente le proteste di piazza che prenderanno poi il nome di *Rivoluzione di Maidan*. A seguito di questo episodio vi è un anno di sostanziale silenzio. Tracciando altre parole, come "Crimea" o "separatisti", si ottiene un pattern simile. Ciò sembra suggerire che gli eventi bellici relativi alla guerra del Donbass non abbiano suscitato un eccessivo interesse da parte dei giornalisti e degli autori di *Otto e mezzo*. L'episodio successivo con un numero rilevante di occorrenze è quello dell'11 febbraio 2015, nel quale vengono discussi gli eventi della conferenza che il giorno successivo porterà al secondo protocollo di Minsk (11-12 febbraio 2015). La situazione cambia radicalmente a inizio 2022, quando nel programma vengono discusse e commentate le varie fasi dell'invasione russa dell'Ucraina, iniziata il 24 febbraio 2022. Durante la primavera del 2022, il tema della guerra in Ucraina viene trattato con frequenza pressoché quotidiana. Il grafico suggerisce che poi l'attenzione alle vicende dell'Ucraina cala lentamente, mantenendosi comunque su un livello di frequenza e intensità elevato.

Anno	Giudizio positivo	Giudizio negativo	Effetti positivi su povertà	Effetti nulli o negativi su povertà	Effetti positivi su occupazione	Effetti nulli o negativi su occupazione	Costi eccessivi	Assistenzialismo
2018	34	50	18	2	6	13	13	10
2019	23	25	17	2	4	7	4	5
2020	14	18	10	3	1	12	0	1
2021	14	11	14	2	0	8	0	0
2022	21	19	17	2	0	10	1	2
Totale	106	123	76	11	11	50	18	18

Tabella 2: Occorrenze dei giudizi espressi riguardo alla politica del Reddito di Cittadinanza, nel dataset, per anno, nel periodo 2018-2022.

Un altro possibile esempio di utilizzo del dataset è quello relativo all'analisi del contenuto di specifici episodi o di specifici frammenti di testo. La tabella 2 mostra i risultati di una ricerca nella quale sono stati analizzati giudizi, contenuti e sentiment relativi alle parti del corpus relative alla politica del Reddito di Cittadinanza. Nello specifico, sono stati selezionati tutti i frammenti di testo, dal 2018 al 2022, contenenti il testo "Reddito di Cittadinanza". Ogni frammento è stato letto ed etichettato manualmente in modo da classificare il giudizio generale espresso relativamente alla misura di policy e la valutazione specifica lungo alcune dimensioni rilevanti: effetti sulla povertà, effetti sull'occupazione, critiche specifiche relative a costi eccessivi o assistenzialismo. I risultati suggeriscono che la politica del Reddito di Cittadinanza, all'interno di *Otto e mezzo*, sia stata elogiata per i suoi effetti sul contrasto alla povertà, ma sia stata attaccata per la sua carenza di effetti nel contrasto alla disoccupazione. Un'analisi di questo tipo mostra come il dataset consenta di raccogliere informazioni dettagliate su un'ampia gamma di argomenti di attualità politica, incluso il giudizio su una politica pubblica che emerge dal dibattito politico televisivo.

Il dataset permette anche di dare conto delle dichiarazioni degli ospiti nel corso del tempo. A titolo di esempio, si è studiato lo stile comunicativo dell'esponente della Lega Matteo Salvini.

Sono stati selezionati tre episodi specifici in cui è presente il leader della Lega (7 marzo 2012, 3 dicembre 2014 e 15 giugno 2015). Per trovare trascrizioni in cui fosse presente Salvini sono state cercate le occorrenze della parola "salvini" all'inizio di ogni trascrizione, in quanto a inizio puntata la giornalista Lilli Gruber introduce gli ospiti presenti. In seguito, utilizzando il testo, sono state raccolte e analizzate le sue dichiarazioni.

L'analisi delle dichiarazioni raccolte attraverso il dataset suggerisce la presenza di una strategia comunicativa di Matteo Salvini centrata sull'enunciazione di liste di elementi all'interno dei propri interventi televisivi. Tra i vari esempi si può guardare alla puntata del 7 marzo 2012, dove Salvini fa un elenco di professioni che, a suo modo di vedere, stanno soffrendo molto la crisi economica:

**Salvini:** "Non bastano le parole bisogna passare ai fatti io sentivo in studio parlare di serenità un governo che ha portato serenità a scendere lo spread io non so se chi è in studio parla con la gente normale parla con i lavoratori con gli artigiani, con i taxisti con i cassi integrati"<sup>1</sup>.

Una struttura comunicativa simile, seppur su temi diversi, si può notare in altre occasioni. Ad esempio, nella puntata del 15 giugno 2015:

**Salvini:** "la Francia fa l'interesse dei francesi come la Repubblica cieca fa l'interesse dei suoi cittadini come l'Inghilterra, come l'Austria, come la Svizzera, come la Germania, come la Polonia e c'è invece che l'Italia che pensa di poter accogliere tutto il mondo".

**Salvini:** "come Lega sì siamo stati in grado di fare un passo indietro di rinunciare a qualche poltrona di rinunciare a qualche candidato sindaco o candidato governatore come in Liguria pur di arrivare al progetto di rilanciare Venezia Arezzo la Liguria Cologno-Montese".

**Salvini:** "guardi io ieri era la giornata mondiale della donazione del sangue il dono del sangue è bello perché è anonimo gratuito e volontario ma se un plasma può andare a un bianco a un giallo a un nero a un roma a un clandestino a un finlandese non lo so non mi interessa faccio del bene".

## 4. Conclusioni

L'articolo ha presentato un nuovo dataset per l'analisi della comunicazione politica, pubblica e mediatica in Italia. Il dataset DOEM è composto dalle trascrizioni di poco meno di 2900 puntate del programma di approfondimento politico *Otto e mezzo*, in onda nei giorni feriali dei mesi da settembre a giugno di ogni anno. Il dataset copre un arco temporale di 15 anni, dal 4 aprile 2011 al 10 giugno 2025. Gli strumenti utilizzati per costruire il dataset hanno permesso di ottenere un'ingente quantità di trascrizioni di elevata qualità. Le principali limitazioni del dataset sono rappresentate da errori grammaticali, errori di punteggiatura, omissione di parti del discorso e mancata o erronea identificazione degli interlocutori e degli ospiti. Ciò non permette di usare il dataset direttamente per lavori che richiedono una rappresentazione scritta

<sup>1</sup>I frammenti di testo sono presi direttamente dal dataset, senza correzioni. Eventuali errori di punteggiatura e di sintassi sono stati volutamente lasciati per mostrare appieno i limiti e le capacità del modello di trascrizione.

della lingua, e quindi, ad esempio, per studi nel campo della linguistica. (si pensi ad esempio ai dataset di news in lingua italiana costruiti manualmente, che vengono usati per fare training di modelli di analisi del linguaggio). Nonostante questi limiti, il dataset DOEM rappresenta una risorsa importante per l'analisi della comunicazione politica, pubblica e televisiva nel contesto italiano e permette soprattutto analisi che tengano in considerazione un arco temporale medio-lungo. La relativa invarianza nel tempo del format televisivo, dello stile di conduzione e della costruzione dei contenuti permette di utilizzare il corpus per analizzare sia le tendenze di lungo periodo nel dibattito politico italiano, sia le "conseguenze" di eventi esogeni al programma – quali pandemie, crisi di governo, specifici fatti di cronaca o eventi di rilevanza politica, guerre ed eventi internazionali – sulla discussione pubblica e politica italiana.

## 5. Condizioni d'utilizzo

Il dataset DOEM è scaricabile al seguente link: <https://centri.unibo.it/computational-social-science/it/dataset>. Il dataset è utilizzabile e redistribuibile secondo i termini della licenza Creative Commons Attribution-NonCommercial-ShareAlike (CC BY-NC-SA).

## Riferimenti bibliografici

- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.
- Salganik, M. J. (2017). *Bit by Bit: Social Research in the Digital Age*. Princeton University Press, Princeton, NJ.
- SYSTRAN (2023). faster-whisper. <https://github.com/SYSTRAN/faster-whisper>.

### Affiliation:

Tommaso Aicardi  
Bocconi University, Milan, Italy  
E-mail: [tommaso.aicardi2@phd.unibocconi.it](mailto:tommaso.aicardi2@phd.unibocconi.it)  
URL: <https://tommasoaicardi.github.io/My-Website/>

Chloe Papadopoulou  
University of Amsterdam, Netherlands  
E-mail: [c.papadopoulou@uva.nl](mailto:c.papadopoulou@uva.nl)

Marco Albertini  
University of Bologna, Italy  
E-mail: [marco.albertini2@unibo.it](mailto:marco.albertini2@unibo.it)

SocArXiv Website  
SocArXiv Preprints

<https://socopen.org/>  
<https://osf.io/preprints/socarxiv>

Preprint  
URL/DOI GOES HERE

Submitted: 9 giugno 2026  
Accepted: 9 giugno 2026