Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in
Scienze Statistiche

Studio comparativo tra testi costituzionali: relazione tra costituzione e democrazia

Relatore:
Prof. Andrea Sciandra

Laureanda:
Agnese Carroli
Matricola N. 2018894

Anno accademico 2023/2024

Indice

Introduzione

Secondo il report annuale di Amnesty International [1], il 2023 è stato un anno in cui il mondo ha fatto grossi passi indietro rispetto alla promessa dei diritti umani universali del 1948. Nell'anno del 75° anniversario della Dichiarazione Universale dei Diritti Umani, sembra traballare l'impegno che sancisce i diritti inalienabili di ogni essere umano, senza distinzioni di razza, sesso, religione, ideologia politica e caratteristiche culturali. Nonostante il mondo non sia mai stato così ricco, il numero di persone che vivono in un contesto democratico, in senso ampio, è regredito ai livelli del 1985 (*Democracy Report 2024*, V-Dem Institute [2]). Anche secondo l'Economist Intelligence Unit il punteggio globale medio della democrazia è in decrescita continua; soltanto l'8% della popolazione vive in una democrazia completa mentre quasi il 40% vive sotto un regime autoritario, percentuale aumentata negli ultimi anni. Nel relativo report annuale *Democracy Index 2023 - Age of conflict* redatto dell'Economist Intelligence Unit [3], il 2023 è stato definito come *l'anno del conflitto* per la crescente incidenza di conflitti violenti e per l'incapacità dei governi democratici nel gestire i conflitti politici e sociali all'interno dei rispettivi paesi.

Col fine di comprendere meglio lo scenario attuale, si vogliono conoscere e analizzare le radici della forma di governo che realmente permane in ciascuno Stato. L'unico strumento che accomuna tutti gli Stati come espressione del proprio popolo, o una sua rappresentanza, è il documento costituzionale. Essendo la dichiarazione diretta che ciascuno Stato possiede per descriversi e per organizzare la vita dei suoi cittadini, la costituzione di fatto rappresenta un'importante fonte di informazioni per studiare le radici del potere di ciascuno Stato. É auspicabile pensare che ogni costituzione venga approvata dalla maggioranza del popolo, tuttavia non è realistico. Ci si interroga quindi sulla effettiva validità del testo costituzionale per studiare la distribuzione della manifestazione delle diverse forme di governo nel mondo. La gestione e il mantenimento del potere all'interno di uno Stato è un equilibrio difficile da gestire, le cui radici possono risiedere all'interno della costituzione per un paese democratico. Lo squilibrio che porta a un'espressione autoritaria è frutto di molteplici elementi ma è di interesse comprendere se questo si riflette anche a li-

vello costituzionale. Al netto delle tempistiche che l'operazione di aggiornamento del testo comporta o dell'emanazione di un nuovo documento costituzionale, ci si interroga su quanto il contenuto testuale coincida con la manifestazione effettiva.

L'obiettivo preliminare della ricerca è quello di confrontare i diversi testi costituzionali per far emergere similitudini e divergenze in termini di struttura e contenuto. Per rendere possibile il confronto si considera la traduzione in lingua inglese di tutti i testi costituzionali, resa disponibile grazie al lavoro dell'organizzazione noprofit Compative Constitutions Project [5]. Assumendo che la perdita di contenuto derivante dall'operazione di traduzione sia minima, si analizzano i testi per valutare la discrepanza tra ciò che viene dichiarato a livello costituzionale rispetto a ciò che ne traspare da fonte esterne al paese. Per quantificare l'espressione della forma di governo è stato utilizzato l'indice di democrazia proposto dall'Economist Intelligence Unit nel report annuale del 2023. Il principale obiettivo della ricerca è quello di valutare se la forma di governo vigente di uno Stato sia il riflesso del contenuto del rispettivo testo costituzionale.

Il presente studio risulta essere composto come segue. Il primo capitolo contestualizza l'obiettivo della ricerca, andando a descrivere nel dettaglio il concetto di costituzione e l'indice di democrazia utilizzato. Il Capitolo 2 racchiude le operazioni di costruzione del dataset e le analisi esplorative sul corpus di testi. Nel Capitolo 3 si esplora la composizione dei testi costituzionali in termini di contenuto tramite tecniche di machine learning non supervisionato. Gli ultimi due capitoli sono dedicati alla relazione tra testo costituzionale e indice di democrazia. In particolare, vengono adattati dei modelli di regressione con l'obiettivo di spiegare e prevedere l'indice di democrazia tramite la composizione del testo costituzionale. Nel Capitolo 4, viene utilizzato un approccio che si basa sulla distribuzione della frequenza delle parole all'interno di ciascun testo. Mentre nel Capitolo 5 si utilizzano tecniche di elaborazione del linguaggio naturale per estrapolare e quantificare il contenuto dei testi costituzionali.

Capitolo 1

Costituzione: strumento di democrazia?

La costituzione è il principale documento che descrivere i principi base di uno Stato, il funzionamento e l'organizzazione di un governo e i diritti fondamentali dei suoi cittadini. Una costituzione si può definire come una sorta di *legge principale* o *legge madre*: un documento più generale a cui le leggi ordinarie si devono attenere.

Non esiste una definizione unica e universale sul contenuto e la natura di una costituzione, questo varia notevolmente tra i diversi Stati. Tuttavia, secondo il report *What Is a Constitution? Principles and Concepts* di International IDEA [4], si può affermare che:

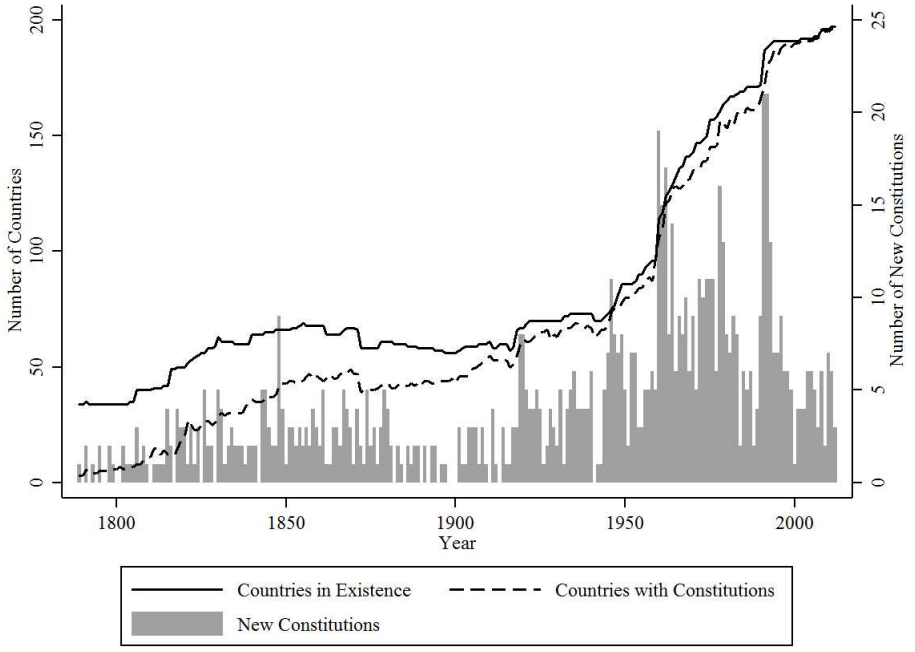
- si riferisce a tutti i membri di uno Stato, anche a coloro che legiferano le leggi;
- descrive la struttura e le operazioni delle istituzioni del governo, i principi politici e i diritti dei cittadini;
- è basata su una validità pubblica estesa;
- contiene leggi più difficili da modificare rispetto a leggi ordinarie (es. tramite 2/3 della maggioranza o un referendum popolare);
- include i criteri riconosciuti internazionalmente per un sistema democratico in termini di rappresentazione e diritti umani.

Una costituzione è, dunque, un insieme di regole che ciascuno Stato definisce e mette in vigore, approvate dalla maggioranza del popolo o da un organo costituente istituito dai cittadini stessi o dal parlamento. In una costituzione si possono definire i confini politici, la natura e l'autorità della comunità politica, l'identità e i valori

su cui si basa la nazione, i diritti e i doveri dei cittadini, le regole delle diverse istituzioni politiche, ecc.

In *Figura 1.1* è riportata la distribuzione del numero di paesi che hanno una costituzione rispetto al numero di paesi esistenti. Si osserva che il numero è in aumento e che le due distribuzioni sono molto simili, soprattutto in riferimento agli anni più recenti. Questo rimarca il fatto che ogni Stato indipendente per poter funzionare ha bisogno di una legge principale a cui far riferimento e la volontà di avere una propria costituzione deriva dai cittadini stessi.

Figura 1.1: Distribuzione del numero di costituzioni rispetto al numero di Stati esistenti e distribuzione delle nuove costituzioni dal 1800 ad oggi - Fonte: [5]



La più antica costituzione ancora in vigore è sicuramente quella del Regno Unito che risale al 1215, modificata più volte successivamente. Tuttavia, il caso del Regno Unito è del tutto eccezionale in quanto è una delle poche *costituzioni non codificate* ovvero scritte parzialmente. Infatti, il testo non è messo per iscritto in un singolo documento costituzionale bensì le regole fondamentali sono definite dal susseguirsi di statuti e trattati. Altri esempi di *costituzioni non codificate* sono quelle relative ai paesi di San Marino, Nuova Zelanda, Canada, Svezia e Israele[1]. Ad eccezione del Regno Unito, la più antica costituzione considerata propriamente scritta è quella

[1] ognuno di questi sei paesi riflette una situazione specifica in relazione alla propria storia di formazione del governo e dello Stato, si fa riferimento al sito di World Population Review [6]

degli Stati Uniti d'America emanata per la prima volta nel 1789 e successivamente aggiornata più volte. In riferimento agli ultimi dati disponibili, le costituzioni più recenti fanno riferimento all'anno 2023 in cui i paesi di Mali e Chad hanno approvato tramite un referendum popolare i nuovi testi costituzionali [7]. La scrittura o la riscrittura di un testo costituzionale può avvenire per molteplici motivi ed è strettamente collegato alla storia del Paese in tutte le sue sfaccettature. Inoltre, tali documenti possono subire variazione quindi anche le costituzioni più vecchie possono essere modificate di pari passo allo sviluppo del relativo popolo.

Il processo di stesura di una costituzione può richiedere alle volte diversi anni ma la modalità specifica è diversa da Stato a Stato. Nella maggior parte dei casi viene eletto dai cittadini un gruppo di persone a numero variabile (in Kenya il gruppo era formato da 9 esperti, in Spagna è stata scritta da 7 membri parlamentari, in Tunisia sono state elette 200 persone per istituire un'assemblea) istituito specificatamente per la composizione del testo. Indipendentemente dal numero di persone da cui è composta tale assemblea, questa dovrebbe comprendere al suo interno oppositori rispetto al governo in atto, uomini e donne, giovani e anziani, minoranze etiche, culturali e linguistiche.

1.1 La struttura di una costituzione

Rispetto a quanto già detto, non esiste una struttura univoca di un testo costituzionale poiché viene scritto da una rappresentanza di un popolo e quindi rispecchierà specificatamente le caratteristiche del singolo Stato. Nell'articolo *What Is a Constitution? Principles and Concepts* [4] viene presentato lo schema di una struttura tipica, ripotato qui di seguito:

1. Preambolo: dichiarazione dei motivi e dei principi di base della costituzione messa in vigore, alle volte sono presenti anche riferimenti ad eventi storici importanti, valori e identità della nazione;

2. Preliminari: dichiarazione di sovranità o principi base del governo; il nome del territorio dello Stato; cittadinanza e diritto di voto; ideologia dello Stato, valori e obiettivi;

3. Diritti fondamentali: lista dei diritti, includendo la loro applicabilità, limiti, sospensioni o restrizioni durante un'emergenza di Stato;

4. Diritti sociali ed economici;

5. Parlamento o legislatura: struttura, composizione, privilegi, procedure, etc;

6. Capo di Stato: metodo di elezione, poteri, durata del mandato;
7. Governo (in un parlamento o un sistema semi-presideziale): regole di formazione del governo, responsabilità e poteri;
8. Sistema giudiziario: come è strutturato, nomine giudiziarie, indipendenza del sistema giudiziario dal sistema politico, pubblici ministeri;
9. Governo sub-nazionale: struttura e definizione dei poteri locali o regionali;
10. Referedum: disposizione e validità dei referendum popolari;
11. Integrità degli organi di governo (commissione elettorale, difensore civico, istituto di revisione, etc);
12. Settore sicurezza: comandante e conformazione della sicurezza nazionale, eventuali restrizioni al potere militare;
13. Particolarità: disposizioni particolari per gruppi specifici quali ad esempio minoranze, leggi linguistiche, istituzioni particolari e proprie del singolo Stato;
14. Procedure di modifica: disposizioni transitorie e modalità di modifica del testo costituzionale.

Il testo costituzionale si può quindi considerare un documento politico, culturale e sociale che riflette la rappresentazione di uno Stato realizzata dallo Stato stesso attraverso i suoi cittadini, o meglio, una sua rappresentanza. Alla luce di ciò la costituzione si può considerare come un elemento distintivo e caratterizzante la storia, la cultura, i valori di un singolo Stato. Allo stesso tempo, si osserva che confrontare costituzioni di paesi diversi non significa necessariamente paragonarle negli stessi contenuti e che ciascuno Stato possiede le proprie specificità.

1.2 Relazione tra costituzione e forma di governo

La costituzione risulta uno strumento voluto dai cittadini e per i cittadini; auspicabilmente dalla maggioranza del popolo sebbene non sia escluso il suo contrario. Con l'obiettivo di voler conoscere e studiare uno Stato, l'analisi della rispettiva costituzione può essere una buona base di partenza ma non è necessariamente una fonte affidabile per fotografare la situazione odierna e aggiornata della relativa società. La discrepanza tra ciò che viene dichiarato a livello costituzionale da ciò che ne

traspare da fonti terze è la direzione di questa ricerca. É logico aspettarsi che questa differenza possa essere causata indirettamente: le costituzioni sono di per sé leggi rigide e che vengono modificate più difficilmente rispetto a leggi ordinarie, non possono quindi cogliere in tempi brevi un eventuale cambiamento della società e del governo. Allo stesso tempo, uno Stato coinvolto in eventi quali guerre o configurazioni specifiche che stravolgono lo scenario ordinario contribuirà a tali divergenze. Al netto delle diverse possibili configurazioni in cui uno Stato può trovarsi in un momento specifico, ci si interroga sull'affidabilità del contenuto di ciascun testo costituzionale. Ad esempio, la protezione dei diritti umani fondamentali è ampiamente accettata a livello internazionale e incarnata a livello costituzionale da tutti i paesi, nonché dalla Carta delle Nazioni Unite di cui tutti gli Stati sono membri, con qualche piccola eccezione. Nonostante tali principi siano costituzionalmente diffusi ed approvati all'unisono a livello globale, purtroppo non è verosimile pensare che siano realmente garantiti.

Tra le tante caratteristiche che si potrebbero utilizzare per descrivere uno Stato vi è la tipologia di forma di governo, ovvero il modo in cui ciascuno Stato decide di esercitare il proprio potere. Nello specifico, si vuole confrontare il testo costituzionale con un indice che descrive il grado di democrazia di uno Stato. La costituzione per definizione è uno strumento democratico che rappresenta lo Stato stesso in maniera soggettiva e vuole essere confrontata con una fonte esterna che esprima anch'essa se e quanto il medesimo paese è considerato democratico.

1.2.1 Indice di democrazia

Secondo l'Enciclopedia italiana Treccani, per democrazia si intende: *forma di governo che si basa sulla sovranità popolare e garantisce a ogni cittadino la partecipazione in piena uguaglianza all'esercizio del potere pubblico*. Tuttavia, vi è un dibattito molto acceso sulla definizione di tale concetto e di conseguenza anche su come misurarla. Secondo alcuni esperti il concetto di democrazia è dicotomico: o uno Stato è democratico o non lo è, senza alcune mezze misure. Per la maggioranza, invece, è possibile parlare di grado di democrazia, in cui si verifica la presenza di specifici elementi democratici. Alcuni esempi di caratteristiche democratiche sono: il consenso della maggioranza nell'approvazione di nuove leggi, l'esistenza di elezioni libere ed eque, la protezione delle minoranze e il rispetto dei diritti umani fondamentali [3]. Una volta definiti quali aspetti sono di rilievo per la definizione di un paese democratico, si costruisce un indice espresso attraverso una scala numerica limitata: il valore minimo significa assenza di democrazia e il valore massimo

equivale al grado di massima espressione democratica.

Prendendo come riferimento il report annuale del Economist Intelligence Unit (EIU) [3], viene proposta la definizione di indice di democrazia come unione di cinque differenti macro aree: processo elettorale e pluralismo, libertà civili, funzionamento del governo, partecipazione politica e cultura politica. Ciascuna di queste categorie viene misurata su una scala da 0 a 10 e l'indice complessivo è la media dei rispettivi cinque punteggi semplici. Gli indici di ciascuna categoria si basano sulla somma di semplici punteggi di una sequenza di domande; le opzioni di risposta sono sempre tre e corrispondono a 0, 0.5 o 1 punto. L'indice finale è espresso quindi su una scala da 0 a 10 la cui interpretazione può essere sintetizzata nel seguente modo:

- *democrazia completa*: punteggio superiore ad 8;
- *democrazia imperfetta*: punteggio superiore a 6 e inferiore o uguale a 8;
- *regime ibrido*: punteggio superiore a 4 e inferiore o uguale a 6;
- *regime autoritario*: punteggio inferiore o uguale a 4.

Una democrazia completa è una forma di governo in cui le libertà politiche e civili sono rispettate, il funzionamento del governo è soddisfacente, la magistratura è indipendente e vi è un sostegno della cultura politica favorevole alla democrazia. Con democrazie imperfette si intendono quei paesi in cui vi sono elezioni libere ed eque e le libertà civili sono rispettate, tuttavia vi sono alcuni punti deboli tra cui ad esempio una scarsa partecipazione politica, una cultura politica non sviluppata e/o problemi legati alla gestione del governo. I regimi ibridi, invece, sono caratterizzati da irregolarità importanti sul sistema elettorale, può essere comune la pressione del governo su partiti e candidati dell'opposizione, la corruzione tende ad essere diffusa e lo stato di diritto è debole. Infine, i regimi autoritari sono paesi in cui il pluralismo politico è assente o fortemente circoscritto, le elezioni, se si verificano, non sono né libere né eque, non esiste una magistratura indipendente, è presente la censura e in molti casi si tratta di vere e proprie dittature.

Per una spiegazione più dettagliata si fa riferimento all'appendice del report *Democracy Index 2023 - Age of conflict* [3], in cui è presente anche l'intero set di domande utilizzato per il calcolo dell'indice.

Nella *Figura 1.2* sottostante è riportato l'andamento dell'indice di democrazia globale dal 2006 al 2023. Si osserva un continuo declino, toccando nel 2023 il pun-

teggio più basso degli ultimi 17 anni pari a 5.23. Nel 2023 gli Stati che presentano un valore più alto sono in ordine Norvegia (9.81), Nuova Zelanda (9.61) e Islanda (9.45), mentre gli Stati peggiori risultano essere Afghanistan (0.26), Myanmar (0.85) e Corea del Nord (1.08). Andando ad analizzare la serie temporale delle sotto categorie di cui è composto l'indice di democrazia, si evidenzia che la componente di partecipazione politica è in forte e continuo aumento. Tuttavia, le restanti quattro macroaree sono in declino, in maniera più decisiva per il processo elettorale e le libertà civili, come si osserva in *Figura 1.3*.

Figura 1.2: Serie storica della media globale del Democracy Index dal 2006 al 2023

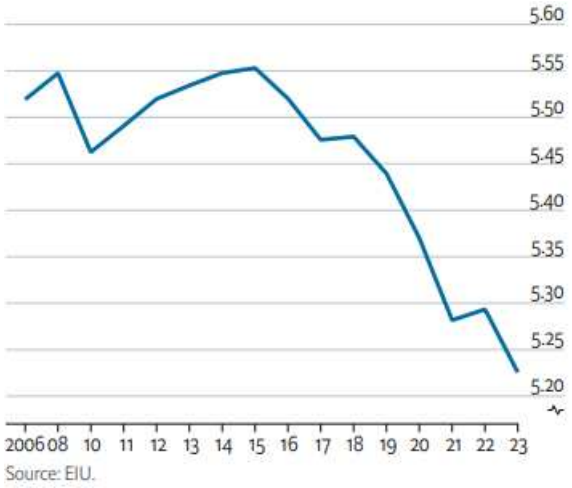
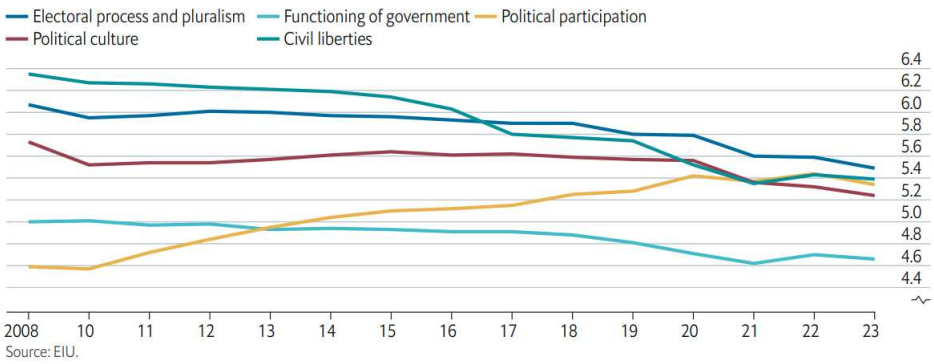


Figura 1.3: Serie storica del Democracy Index suddiviso in macroaree 2008-2023



Secondo la presente definizione di democrazia, quasi la metà della popolazione globale vive in uno Stato democratico ma soltanto il 7.8% risiede in una *democrazia completa*, in calo rispetto al 2015 che ammontava allo 8.9%. Un'altra informazione

non trascurabile è che il 39.4% della popolazione mondiale vive sotto un regime autoritario.

Figura 1.4: Democracy Index 2023 per categorie

	No. of countries	% of countries	% of world population
Full democracies	24	14.4	7.8
Flawed democracies	50	29.9	37.6
Hybrid regimes	34	20.4	15.2
Authoritarian regimes	59	35.3	39.4

Note. "World" population refers to the total population of the 167 countries covered by the Index. Since this excludes only micro states, this is nearly equal to the entire estimated world population.

Source: EIU.

1.3 Domanda di ricerca

In una prima fase, questo studio si concentra sul confronto tra i diversi testi costituzionali, analizzandone la struttura e il contenuto. L'obiettivo è quello di esplorare la distribuzione delle parole in termini di lunghezza e frequenza, per individuare somiglianze tra costituzioni provenienti da storie e contesti differenti. Un altro aspetto di interesse è quello di valutare se il corpus di testi condivida una composizione comune in termini di contenuto. Non essendoci un prototipo universale di costituzione a cui gli Stati devono attenersi, non è detto che l'informazione contenuta sia paragonabile direttamente.

Dopo aver analizzato la composizione e la struttura dei testi costituzionali, si vuole conoscere se questa rispecchi effettivamente la forma di governo vigente in ciascun paese. L'obiettivo principale è quello di mettere il relazione il testo costituzionale con la reale forma di governo, osservata attraverso fonti esterne al paese stesso. Al di là del fatto che ogni costituzione possa rappresentare l'espressione libera di un popolo o, al contrario, il volere di chi detiene il potere, questo studio si propone di valutare se la costituzione rifletta davvero la realtà del paese o se vi sia una discrepanza tra quanto dichiarato e quanto osservato. Per rispondere a questa domanda, vengono utilizzati due approcci distinti di analisi.

Il primo approccio si basa sulla distribuzione della presenza e della frequenza delle principali parole nei testi costituzionali. Nello specifico, il contenuto del testo viene semplificato attraverso le parole più ricorrenti e dalla distribuzione di queste frequenze si delinea una rappresentazione della composizione costituzionale. Questo metodo offre un'interpretazione diretta: si indaga se la frequenza di alcune parole specifiche possa fungere da buon predittore della forma di governo effettivamente in vigore nel relativo Stato. Il secondo approccio, invece, utilizza

modelli linguistici di grandi dimensioni per estrapolare il significato complessivo di ciascuna costituzione, esprimendolo attraverso un sistema di coordinate numeriche multidimensionali. Sebbene questo metodo perda un'interpretazione diretta del modello, ci si avvicina alla possibilità di cogliere il significato profondo di ciascun testo costituzionale e metterlo ancora una volta in relazione alla forma di governo vigente all'interno del paese.

Capitolo 2

Struttura del dataset e analisi descrittive

I dati provengono dal progetto *Constitute* dell'organizzazione noprofit *Comparative Constitutions Project* che colleziona e analizza i testi di tutte le costituzioni di ciascun paese del mondo. Nell'omonimo sito *Constitute Project* [5] sono state raccolte 232 costituzioni tutte tradotte in lingua inglese; le costituzioni sono catalogate come: in vigore, storiche e bozze. Per ciascun documento testuale è indicato l'anno di emanazione ed eventualmente, se presente, l'anno di revisione del documento rispettivamente aggiornato e/o l'anno di riemanazione (se uno Stato ha emanato un nuova costituzione integralmente). Inoltre, alcune costituzioni sono segnalate come *successivamente modificate*; gli emendamenti recenti delle relative costituzioni non sono ancora stati aggiornati nel testo presentato. All'interno del sito i testi sono disponibili interamente anche in spagnolo e in arabo. É bene tenere in considerazione che utilizzare i testi tradotti in lingua inglese significa considerare che una buona parte dei testi ha subito un processo di traduzione da terzi.

Per il seguente studio sono stati considerati soltanto i testi delle costituzioni in vigore, ad eccezione di Guinea (2010) e Mali (1992) che vengono classificate come storiche ma che non hanno un secondo testo disponibile in vigore. Per tali documenti verrà posta particolare attenzione in fase di analisi qualora dovesse emergere un loro comportamento anomalo. Si considerano quindi 195 costituzioni differenti di rispettivamente 195 Stati indipendenti. Rispetto ai 193 Stati membri dell'ONU si aggiungono anche le costituzioni di Palestina e Taiwan, classificati come *membri osservatori* assieme al Vaticano dall'organizzazione intergovernativa [8].

Si compone un dataset di 195 osservazioni e con le seguenti variabili, presenti sempre all'interno del sito:

- *paese*: nome dello Stato;

- *emanazione*: anno di emanazione della costituzione, si considera l'anno della prima emanazione, anche qualora dovesse essere presente una data di riemanazione;

- *revisione*: anno dell'ultima revisione, se non è presente alcuna revisione si considera la modalità 0;

- *testo*: testo integrale della costituzione.

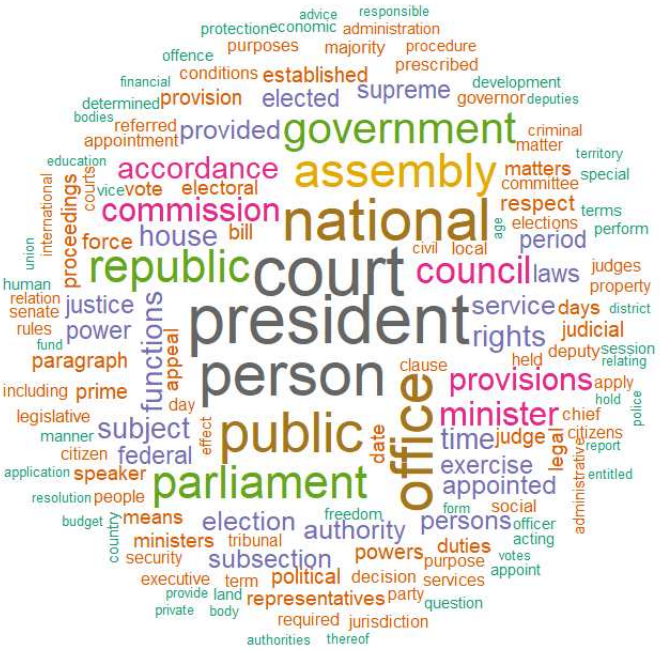
2.1 Pre processing

I dati sopra descritti sono stati scaricati tramite webscraping alla pagina del progetto *Constitute Project* [5]. Tramite il programma R Core Team (2024) [9] e con l'utilizzo della libreria *rvest* è stato possibile ottenere tramite la tecnica di scraping le variabili di interesse. La prima operazione è stata effettuata sulla pagina principale che fa riferimento all'elenco di tutte le osservazioni in cui sono state scaricate le variabili *paese*, *emanazione* e *revisione*. Per quanto riguarda la variabile *testo* si effettua la medesima operazione per ciascuna pagina web corrispondente a ciascuna costituzione. Tramite lo strumento SelectorGadget è stato possibile selezionare soltanto gli elementi HTML di interesse rispettivamente di ogni pagina web. In questo modo è stato ottenuto il testo di ciascuna costituzione ripulito rispettivamente da titoli, sottotitoli, indici, ecc e considerato come unica stringa alfanumerica.

2.2 Analisi descrittive

Per ciascuno dei testi è stata ottenuta la rispettiva lunghezza calcolando il numero di parole di cui sono composti; la lunghezza media dei tesi del corpus è pari a 23579 parole. Le costituzioni più lunghe sono quelle dei paesi del Regno Unito (che conta 218910 parole) e Nuova Zelanda (con 174132 parole), con un netto distacco rispetto a tutte le altre. Tuttavia, si osserva che i testi di questi due paesi fanno parte delle *costituzioni non codificate* per cui all'interno del file considerato come costituzione sono trascritte una serie di trattati e statuti raccolti nel tempo. La costituzione vera e propria più lunga risulta essere quella dell'India con 95773 parole, meno della metà della lunghezza di quella del Regno Unito. Il testo costituzionale più corto è quello della Libia con sole 2893 parole, a seguire Monaco (con 4106 parole) e Islanda (con 4339).

Figura 2.1: Worldcloud delle 150 parole più frequenti nell'intero corpus



La *Figura 2.1* raffigura il wordcloud delle 150 parole più frequenti, la cui grandezza e il cui colore variano proporzionalmente con la frequenza dei vocaboli presenti all'interno di tutto il corpus di testi. Questa rappresentazione è stata ottenuta a partire dalle singole unità testuali, chiamate *tokens*, ripulite dalla punteggiatura, numeri e simboli. Inoltre, sono state eliminate anche le *stopwords*, ovvero quelle parole che non hanno un vero e proprio significato ma sono funzionali per la costruzione di un discorso quali pronomi, articoli, ecc. É stato utilizzato l'elenco della libreria *stopwords* a cui sono state aggiunte anche: *constitution*, *law*, *act*, *article*, *section*, *constitutional*. Queste parole non hanno un vero e proprio significato di contenuto rispetto al corpus che si sta esaminando in quanto tutte le costituzioni sono suddivise in atti, articoli e sezioni quindi risulterebbero parole con alta frequenza ma che descrivono soltanto la struttura. Si evince che le parole maggiormente frequenti sono *president*, *court* e *person*. Da tali vocaboli emerge che: la costituzione descrive il ruolo e la figura di chi rappresenta l'intero Stato, chiamato presidente, delinea la presenza di magistrature, dette corti, e che è un documento per le persone, in quanto membri di uno Stato. Altri termini altamente presenti e che delineano il contenuto del corpus sono ad esempio: parlamento, assemblea, governo, voto, giustizia, potere, diritti.

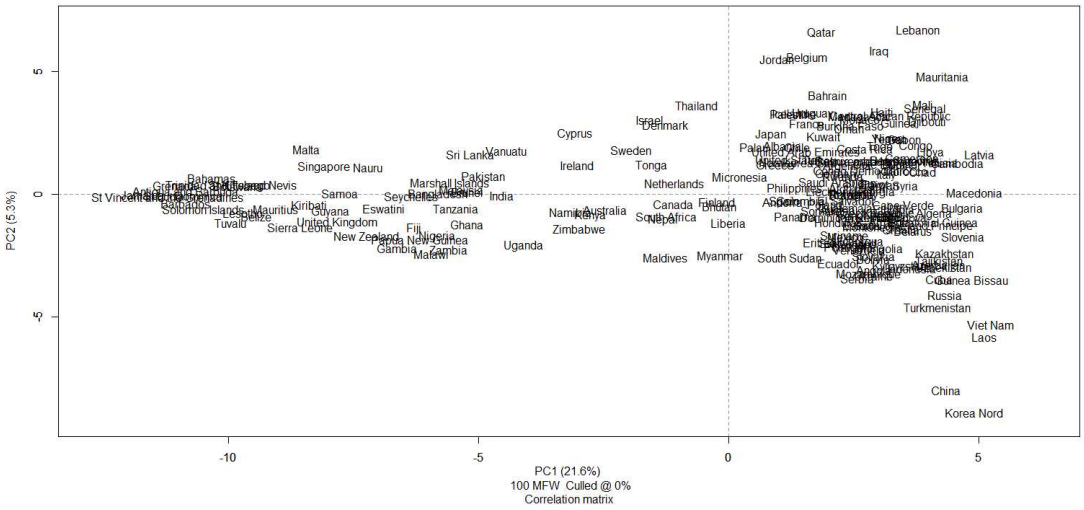
2.3 Analisi delle componenti principali (PCA)

A questo punto ci si interroga sulla presenza di similitudini e di divergenze tra i testi del corpus. In particolare, ci si chiede quali siano i testi costituzionali più simili tra loro in termini di contenuto e, a tal proposito, si effettua un'analisi delle componenti principali attraverso la libreria *stylo*. Ciascun documento del corpus è stato segmentato nelle unità di testo (tokens) ed è stato ripulito dalle *stopwords*; quindi, sono state individuate le parole più frequenti all'interno dell'intero corpus. In questo caso, si sono considerate le 3000 parole più frequenti, come suggerisce l'analisi proposta da Eder et al. (2016) [10] e sono state calcolate le rispettive frequenze dei 3000 termini per ciascuno dei testi costituzionali. Utilizzando tale matrice di dimensione 195x3000 è stata effettuata l'analisi delle componenti principali e in *Figura 2.2* viene riportato il diagramma di dispersione delle prime due componenti. Il grafico mostra, ad esempio, che le costituzioni di Cina e Corea del Nord hanno valori molto simili, presentando un evidente distaccamento rispetto agli altri paesi. I testi a loro più simili sono quelli di Laos, Vietnam, Russia e Turkmenistan. Un altro gruppo di Stati che presentano valori vicini nelle prime due componenti principali sono: Qatar, Libano, Iraq, Giordania, Mauritania e Belgio, di cui i primi 4 sono paesi del Medio Oriente geograficamente vicini. Nella parte a sinistra del grafico, invece, si può osservare un raggruppamento di molti dei paesi caraibici quali St. Vincens e Grenadine, Bahamas, Isole Salomone, St. Kitts e Nevis, Antigua e Barbuda. Nella parte centrale del grafico si concentrano alcuni paesi del nord Europa come Danimarca, Svezia, Olanda, Finlandia a cui si aggiungono anche Canada, Australia, Micronesia, Sud Africa. In *Appendice A* sono riportato alcuni zoom del grafico a dispersione per una leggibilità migliore (*Figure A.2 - A.6*). La quantità di varianza spiegata dalle prime due componenti ammonta al 21.6% per la prima e al 5.3% per la seconda; si riporta nella *Figura A.1* dell'*Appendice A* la distribuzione della varianza spiegata delle prime 30 componenti principali.

2.4 Distanze testuali e cluster analysis

Un altro modo per individuare gruppi di costituzioni simili, qualora ci fossero, è tramite l'utilizzo di distanze intertestuali. Una misura di distanza intertestuale consente di scovare somiglianze e differenze tra i modelli di frequenza dei singoli testi del corpus. In tale sezione si è scelto di utilizzare il distanza Delta di Burrows, come presentata nell'articolo di Argamon (2007) [11]. Si considerano k parole di interes-

Figura 2.2: Analisi delle componenti principali - 3000 parole più frequenti



se, in questo caso le k più frequenti con k pari a 3000 in analogia con le analisi del paragrafo 2.3. Si definisce la distanza Delta tra due testi A e B come:

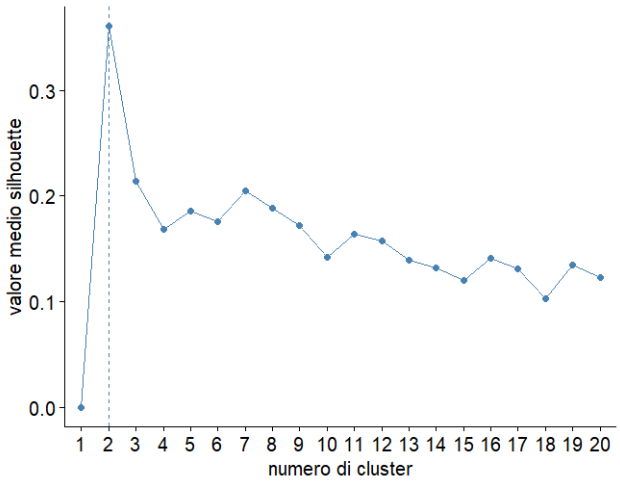
$$\Delta(AB) = \frac{1}{k}\sum_{i=1}^{k}\left|\frac{f_i(A)-\mu_i}{\sigma_i} - \frac{f_i(B)-\mu_i}{\sigma_i}\right|$$

dove: $f_i(A)$ e $f_i(B)$ le frequenze della i-esima parola rispettivamente nei documenti A e B, μ_i e σ_i la frequenza media e la deviazione standard della i-esima parola nel corpus dei testi. A partire dalla matrice delle frequenze di dimensione 195x3000 si utilizza nuovamente la libreria *stylo* per tali operazioni (Eder et al. (2016) [10]). Tale matrice delle distanze ha una dimensione 195x195 e descrive il grado di similitudine di ciascuna costituzione rispetto a ciascun'altra.

Si considera l'indice di Silhouette (Rousseeuw 1987 [12]) per determinare il numero ottimale di cluster in cui suddividere il corpus di testi in gruppi di costituzioni simili. Per effettuare tale operazione si utilizza la libreria *factoextra*; in *Figura 2.3* si evidenzia che il numero ottimale di cluster risulta essere pari a 2. Il valore massimo medio è pari a 0.36 che indica una scarsa coesione interna di ciascun gruppo e una separazione tra cluster non molto netta.

A questo punto, si sono identificati i due gruppi attraverso un clustering gerarchico agglomerativo utilizzando il criterio della minima varianza di Ward per suddividere le osservazioni. Nella *Tabella A.1* dell'*Appendice A* sono stati riportati i rispettivi cluster così ottenuti. Si può osservare che il cluster 2 è formato quasi esclusivamente dai paesi del vecchio Commonwealth delle nazioni, le cui origini arrivano dalla composizione dell'Impero britannico. Ancora oggi l'organizzazione

Figura 2.3: Indice medio di silhouette per numero di cluster - distanza delta



è attiva con l'obiettivo di far cooperare economicamente i diversi Stati membri, accomunati dalla stessa radice. Per la composizione degli Stati membri del Commonwealth delle Nazioni si fa riferimento all'elenco proposto dall'Enciclopedia Italiana Treccani - atlante geopolitico (2013) [13]. Sul totale dei 54 paesi membri odierni, soltanto Australia, Canada, Camerun, Ruanda e Mozambico non fanno parte del secondo cluster. Irlanda, Zimbabwe, Israele e Isole Marshall sono catalogati nel secondo cluster sebbene non facciano parte del Commonwealth; tuttavia si osserva che Irlanda e Zimbabwe sono due ex Stati membri dell'organizzazione. I risultati di tale clustering evidenziano quindi un'influenza britannica a livello costituzionale sui rispettivi Stati ex coloniali, con cui ha mantenuto uno stretto rapporto commerciale e istituzionale.

Alla luce dei risultati di cui sopra, per evidenziare ulteriori analogie con strutture di relazioni storiche, si è effettuato un clustering a tre gruppi analogamente a quanto descritto in precedenza. I risultati di tale clustering sono riportati *Tabella A.2* dell'*Appendice A*, utilizzando sempre il criterio di Ward. Il cluster C qui ottenuto è l'analogo del cluster 2 della *Tabella A.2* mentre i cluster A e B sono due sotto gruppi del relativo cluster 1 precedente. In questo caso, il gruppo B identifica gran parte dei paesi che facevano parte dell'ex Impero coloniale francese. Si confronta l'elenco con quello presentato su Wikipedia [14] e si osserva che ad eccezione di Andorra, Burundi e Djibouti i restanti Stati sono tutti ex colonie francesi. Anche in questo caso si può pensare che l'influenza francese nel periodo coloniale abbia impattato a livello profondo la cultura e la società di questi paesi e che tale condizionamento sia presente ancora oggi a livello costituzionale. A differenza del Commonwealth, che tuttora esercita un potere economico e strategico, la comunità francese non ha

più alcun ruolo di rilievo.

Le medesime analisi sono state effettuate anche considerando altre due misure di distanza intertestuale, quali coseno e argamon. Le distanze tra due testi A e B sono definite come:

- distanza coseno:

$$\Delta(AB) = 1 - \frac{\sum_{i=1}^{k} f_i(A) f_i(B)}{\sqrt{\sum_{i=1}^{k} f_i(A)^2} \sqrt{\sum_{i=1}^{k} f_i(B)^2}}$$

- distanza argamon:

$$\Delta(AB) = \frac{1}{k} \sum_{i=1}^{k} \sqrt{\left| \frac{f_i(A)^2 - f_i(B)^2}{\sigma_i} \right|}$$

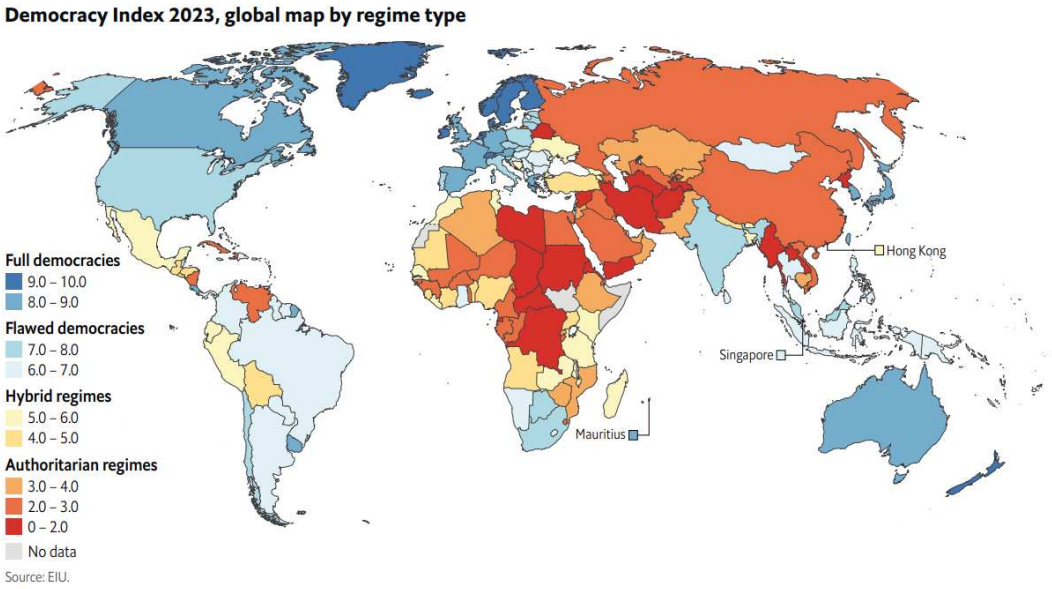
dove: $f_i(A)$ e $f_i(B)$ le frequenze della i-esima parola rispettivamente nei documenti A e B, μ_i e σ_i la frequenza media e la deviazione standard della i-esima parola nel corpus dei testi. Tuttavia, portano ai medesimi risultati presentati per la distanza Delta di Burrows che risulta essere preferibile in quanto può essere vista come una misura di differenza normalizzata tra le frequenze dei testi messi a confronto. Nelle *Figure A.7* e *A.8* dell'*Appendice A* sono riportati i valori degli indici di silhouette per la determinazione del numero ottimale di cluster in cui suddividere il corpus. Anche in questi due casi il valore massimo dell'indice si ottiene con 2 gruppi che corrispondono ai medesimi presentati sopra se si ripercorrono le stesse scelte di analisi.

2.5 Dati e distribuzione dell'indice di democrazia

Un ulteriore aspetto di interesse è quello di mettere in relazione il testo costituzionale con la rispettiva forma di governo vigente all'interno di ciascuno Stato. Come descritto nel paragrafo 1.2.1, la forma di governo di ciascuno Stato viene espressa tramite l'indice di democrazia fornito dal gruppo di ricerca e analisi Economist Intelligence Unit. In prima analisi, si vuole guardare più nel dettaglio come il grado di democrazia si distribuisce globalmente al giorno d'oggi. I dati considerati sono stati scaricati dal report *Democracy Index 2023 - Age of conflict, Tabella 2* [3]. Tale variabile si compone di 166 osservazioni che fanno riferimento rispettivamente a 166 Stati o territori indipendenti. Rispetto al dataset considerato, sono 29 i paesi man-

canti rispetto al totale delle 195 costituzioni analizzate. Tuttavia, le 166 osservazioni coprono la quasi totalità dell'intera popolazione mondiale poiché gli Stati mancanti sono tutti microstati quali ad esempio Monaco, Andorra e Maldive[1]. In *Figura 2.4* è riportata la mappa globale dell'indice di democrazia classificato per punteggio. Gran parte dell'Africa e dell'Asia sono governate da regimi autoritari, Europa, Nord America e Oceania sono caratterizzate da democrazie complete o imperfette mentre il Sud America si colloca in una posizione intermedia tra democrazie imperfette e regimi ibridi.

Figura 2.4: Mappa globale dell'indice di democrazia - anno 2023. Fonte: [3]



[1]L'elenco completo degli Stati mancanti per l'indice di democrazia è il seguente: Andorra, Antigua and Barbuda, Bahamas, Barbados, Belize, Brunei, Dominica, Grenada, Kiribati, Kosovo, Liechtenstein, Maldives, Marshall Islands, Micronesia, Monaco, Nauru, Palau, Samoa, Sao Tome and Principe, Seychelles, Solomon Islands, Somalia, South Sudan, St Kitts and Nevis, St Lucia, St Vincent and The Grenadines, Tonga, Tuvalu, Vanuatu

Capitolo 3

Topic modelling

Uno degli principali aspetti di interesse è quello di comprendere se il corpus dei testi costituzionali ha una struttura comune in termini di contenuto. Si vogliono individuare quali sono, se ci sono, le macro tematiche salienti sottostanti ai testi. A tal proposito, si considera un approccio non supervisionato di machine learning detto topic modeling che consente di individuare argomenti nascosti in un insieme di testi. Uno dei modelli più utilizzati è il Latent Dirichlet Allocation (LDA) proposto da Blei et al. (2003) [15] che considera ogni documento come una mistura di argomenti e ogni argomento come una mistura di parole. In questo capitolo si utilizza un'estensione del modello LDA che consente di introdurre dei metadati che influenzano sia gli argomenti sia le parole di ciascun topic; questo modello prende il nome di Structural Topic Model (STM).

3.1 Terminologia e notazione

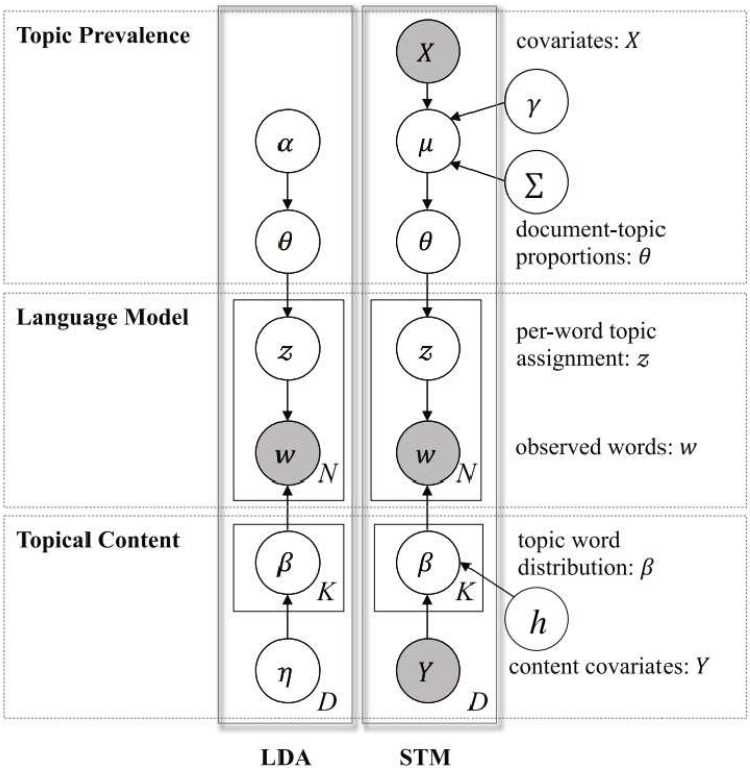
Con l'obiettivo di presentare la metodologia di topic modeling è utile introdurre i seguenti elementi di terminologia e notazione utilizzanti all'interno del capitolo. Si rappresenta con D il corpus di testi, il cui generico elemento è indicizzato da d. Il d-esimo documento è formato da un insieme di parole associate a una specifica frequenza. L'elemento w_n (detto *word token*) indica l'occorrenza della n-esima parola, dove n appartiene a $\{1,...,N_d\}$ con N_d il numero di parole del d-esimo documento. Si indica con w_v ogni singola parola (detta *word type*) indicizzata da v elemento di $\{1,...,V\}$, dove V la lunghezza del vocabolario ovvero il numero di parole distinte all'interno del corpus. Gli argomenti nascosti che si vogliono individuare sono indicizzati da k, con k che appartiene a $\{1,...,K\}$. Infine, si definisce X la matrice DxP dei metadati che influisce sulla distribuzione degli argomenti e Y la matrice DxA delle covariate che controlla l'influenza delle parole all'interno dei topic. I vettori

x_d e y_d rappresentano le covariate delle rispettive matrici del d-esimo documento con lunghezza pari a P ed A rispettivamente.

3.2 Structural Topic Modeling (STM)

Nella costruzione di uno Structural Topic Model proposta da Roberts et al. (2016) [16], si assume un approccio *bag of words* in cui i testi vengono considerati un insieme casuale di parole indipendenti tra di loro. In questo modo si perde la sequenzialità delle parole e ci si concentra sulle frequenze con cui ciascuna parola appare nel testo. Si assume, inoltre, che vi siano un certo numero di argomenti latenti comuni a tutti i testi e che ciascuno di questi sia formato da una mistura di parole.

Figura 3.1: Composizione dei modelli Latent Dirichlet Allocation e Structural Topic Model - Fonte: Bai et al. (2021)[17]



In *Figura 3.1* è riportata la struttura del modello in cui rispetto alla formulazione base LDA vengono inserite le matrici delle covariate X e Y che possono influenzare rispettivamente la distribuzione degli argomenti e quella delle parole che compongono ciascun topic. La struttura del modello può essere quindi suddivisa in tre parti:

1. *prevalenza tematica*: determina in che modo i documenti vengono assegnati a ciascun topic in funzione delle covariate X;

2. *contenuto tematico*: controlla la frequenza di ciascuna parola per ogni topic in funzione delle covariate Y;

3. *modello di osservazione centrale*: dall'unione delle due fasi precedenti si ottiene per ciascun topic la distribuzione delle parole che lo compongono.

Si analizza qui in seguito separatamente ciascuna parte del modello.

Prevalenza tematica (Topic Prevalence)

Nella fase di prevalenza tematica l'obiettivo è quello di ottenere il vettore θ_d con $d \in \{1,...,D\}$ che rappresenta la probabilità che il documento d-esimo appartenga a ciacun topic. Il vettore θ_d ha dimensione $1x(K-1)$ poiché $\sum_{i=1}^{K} \theta_{i,d} = 1$ e quindi il K-esimo valore è ottenuto come differenza. Si assume che la prevalenza tematica θ_d del d-esimo documento si distribuisca come una Normale-Logistica, la cui media è ottenuta a partire dal vettore x_d delle covariate $\mu_d = \Gamma' x_d'$ dove Γ è una matrice formata dai vettori γ_k per colonna, generati a sua volta da una Normale P-variata. Si può sintetizzare in due step come:

$$\gamma_k \sim N_P(0,\ \sigma_k^2\ I_p) \qquad k = 1,...,K-1 \tag{1}$$

$$\theta_d \sim LogN_{K-1}(\mu_d,\ \Sigma) \tag{2}$$

dove σ_k^2 viene generata a partire da una distribuzione Gamma-Inversa con parametri fissati $a = 1$ e $b = 1$ mentre Σ è la matrice $(K-1)x(K-1)$ delle covarianze condivisa da tutti i topic. É come se si stesse considerando una distribuzione *a priori* sugli argomenti in funzione dei metadati.

Contenuto tematico (Topic Contents)

Il contenuto tematico del k-esimo topic nel d-esimo documento $\beta_{.,d,k}$ rappresenta il singolo elemento della matrice di dimensione DxK. Si assume che questa matrice sia influenzata dalle modalità delle covariate Y che agiscono direttamente sulla distribuzione delle parole per ciascun topic. La probabilità della v-esima parola del d-esimo documento per il k-esimo topic è espressa come:

$$\beta_{v,d,k} = \frac{m_v + h^{(t)}_{k,v} + h^{(c)}_{y_d,v} + h^{(i)}_{y_d,k,v}}{\sum_v exp(m_v + h^{(t)}_{k,v} + h^{(c)}_{y_d,v} + h^{(i)}_{y_d,k,v})} \quad v = 1,...,V \;\; k = 1,...,K \qquad (3)$$

dove m_v esprime il tasso marginale log-trasformato della v-esima parola nel d-esimo documento e gli elementi h sono le quote di deviazione della v-esima parola in funzione dei topic, delle covariate e della loro interazione. In particolare, $h^{(t)}_{k,v}$ è la quota di deviazione della frequenza della v-esima parola del k-esimo topic, $h^{(c)}_{y_d,v}$ la deviazione del v-esimo token rispetto alle covariate del d-esimo documento e $h^{(i)}_{y_d,k,v}$ la deviazione rispetto all'interazione tra il k-esimo topic e le covariate.

Modello di osservazione centrale (Language Model)

La terza componente del modello è espressa dalla combinazione tra la prevalenza tematica e il contenuto tematico. Dal vettore θ_d si ottiene un'assegnazione tematica $z_{d,n}$ per la n-esima parola del d-esimo documento attraverso una regressione Multinomiale. Condizionatamente al contenuto tematico $\beta_{d,\, k = z_{d,n}}$ si ottiene la distribuzione del n-esimo token $w_{d,n}$ risultante sempre dalla rispettiva regressione Multinomiale. Le formule (4) e (5) esprimono in sintesi la formulazione del modello di osservazione centrale:

$$z_{d,n} \sim Multinomial_K(\theta_d) \qquad n = 1,...,N_d \qquad (4)$$

$$w_{d,n} \sim Multinomial_V(\beta_{d,\, k = z_{d,n}}) \qquad n = 1,...,N_d \qquad (5)$$

Gli autori del modello STM propongono l'algoritmo della Variational Expectation-Maximization per la stima dei parametri della distribuzione. Si fa riferimento all'articolo *Stm: An R Package for Structural Topic Models* di Roberts, Margaret E., et al. 2019 [18] per una documentazione più estesa.

3.3 Applicazione del modello ai dati

Si vuole applicare il modello presentato al corpus di testi formato dalle 195 costituzioni rappresentate dalla variabile *testo*, come descritto nel Capitolo 2. Assumendo che vi sia un certo numero K di argomenti latenti all'interno dell'insieme dei testi, si vuole individuare quali e quanti siano. Inoltre, si tiene conto che ciascuna costituzione è stata scritta ed eventualmente revisionata in anni differenti e che quindi

l'espressione del contenuto può variare rispetto all'ordine temporale. La variabile *anno* è stata ottenuta come valore massimo tra la data di emanazione e quella di revisione, esprime quindi l'anno in cui è stata revisionata l'ultima volta o direttamente l'anno di emanazione per le costituzioni più recenti. La matrice X descritta nel paragrafo soprastante coincide con la sola covariata *anno*, mentre la variabile Y non viene considerata. La variabile temporale inserita varia da un valore minimo pari a 1946, relativo all'anno di emanazione della costituzione giapponese mai revisionata, fino a un valore massimo corrispondente all'anno 2021 relativo all'ultima revisione dei paesi Cile, Ecuador e Perù. In particolare, la variabile *anno* è stata inserita come spline di regressione con 10 gradi di libertà.

Per ciascuno dei D documenti che compongono il corpus si applica l'estrazione dei tokens quali le parole di cui è formato. Si effettuano poi operazioni di pulizia quali rimozione della punteggiatura, di simboli e di numeri presenti nel testo e la rimozione delle parole prive di significato (*stopwords*). Vengono eliminate anche le parole che hanno un lunghezza inferiore a 3 caratteri e tutti i tokens vengono riportati in formato minuscolo. Si applica poi una procedura di *stemming* in cui ciascuna parola viene riportata alla sua radice (ad esempio *nation*, *nationality*, *national*, *nationalism*, *nationality* hanno tutti la medesima radice comune *nation*). Si ottiene in questo modo una Document-Term-Matrix (DTM) di dimensione 195x19135 che contiene per riga i diversi paesi e per colonna tutti i termini presenti all'interno del vocabolario formato dall'intero corpus. Ciascuna cella indica la frequenza con cui ciascun token compare all'interno del documento. Per ridurre la dimensionalità della DTM vengono escluse dal vocabolario tutte le parole che compaiono in meno di 5 costituzioni distinte e tutti i tokens che hanno un frequenza assoluta inferiore a 30. La matrice finale ha una dimensione di 195x2152.

Per l'applicazione del modello STM si fa riferimento alla libreria *stm*.

3.4 Scelta del numero di topic: CD score

L'unico aspetto del modello non ancora discusso è la scelta del numero K di argomenti che si vogliono ricercare all'interno del corpus di testi. A tal proposito si effettua la modellazione su una griglia di valori possibili per K e si sceglie il modello che si adatta meglio al corpus di testi. Per misurare la qualità dell'adattamento del modello ai dati si utilizza la metrica Consistency and Differentiation (CD) scores proposta da Sciandra et al. (2023) nell'articolo *Diagnostics for topic modelling. The dubious joys of making quantitative decisions in a qualitative environment.* [19]. Il

CD score è ottenuto come norma euclidea L_2 tra gli indici di coerenza semantica e di esclusività opportunamente normalizzati. Più nel dettaglio:

- coerenza semantica: esprime quanto le parole più frequenti compaiono insieme all'interno dei documenti del corpus. La coerenza semantica del k-esimo topic può essere espressa come:

$$C_k = \sum_{i=2}^{M} \sum_{j=1}^{i-1} log\Big(\frac{D(v_i, v_j)+1}{D(v_j)}\Big)$$

  dove $D(v_i, v_j)$ rappresenta il numero di volte in cui le parole v_i e v_j appaiono assieme all'interno del documento ed M è il numero di parole più frequenti utilizzate (in questo caso M è pari a 2152). Un alto valore di coerenza semantica indica che i topic corrispondenti sono caratterizzati da parole molto frequenti;

- esclusività: esprime quanto ciascuna parola presente all'interno dei topic è esclusiva per il k-esimo topic. Si può formulare come media armonica pesata tra il rango della v-esima parola in termini di esclusività e di frequenza, espresso come:

$$FREX_{k,v} = \Big(\frac{w}{ECDF(\beta_{k,v}/\sum_{j=1}^{K}\beta_{j,v})} + \frac{1-w}{ECDF(\beta_{k,v})}\Big)^{-1}$$

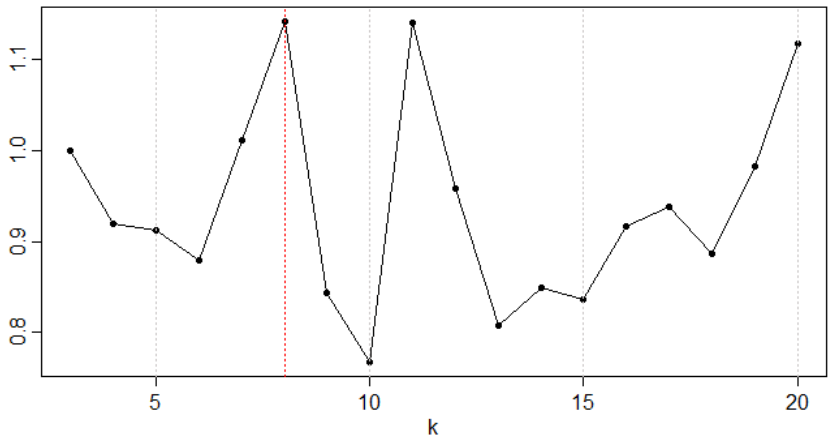
  dove w in questo caso è un parametro di pesi qui fissato a 0.7 per favorire l'esclusività e ECDF rappresenta la distribuzione empirica cumulata delle probabilità β. Un alto valore dell'indice FREX per la v-esima parola del k-esimo topic indica che, tra i termini più frequenti, tale parola caratterizza quel topic rispetto agli altri.

Il numero di topic K è stato valutato da un minimo di 3 a un massimo 20 e in *Figura 3.2* viene riportata la distribuzione del CD score. Il valore massimo è assunto per $K = 8$ con un punteggio di 1.142, leggermente superiore rispetto al CD score di $K = 11$; in ogni caso si predilige il modello più parsimonioso.

3.5 Risultati e interpretazioni

Il modello finale si compone quindi di 8 argomenti latenti trovati all'interno del corpus dei testi costituzionali, il cui contenuto è rappresentato dalla distribuzione delle parole condizionatamente a ciascun topic. Viene riportato in *Figura 3.3* l'output

Figura 3.2: Consistency and Differentiation (CD) score per numero di topic *K*



del modello adattato in cui per ciascuno degli 8 argomenti vengono affiancate le 15 parole più frequenti e le 15 parole con un indice FREX più alto.

Emerge chiaramente che molte parole sono contenute in più topic come ad esempio *shall* e *may* che sono tra le parole più frequenti di ciascun topic. Il token *shall* richiama il concetto di dovere e il primo topic è l'unico a includerlo anche tra le parole con FREX più alto; il quarto topic invece è l'unico a non includere tra le parole più frequenti *shall* ma soltanto *may* che invece esprime il concetto di possibilità di un'azione. Altre parole altamente frequenti in molti argomenti sono ad esempio *state*, *nation*, *public* e *person* che sostanzialmente rappresentano il contenuto principale della costituzione, come emerge anche dal wordcloud di *Figura 2.1*. Guardando più nel dettaglio la combinazione delle parole e dando maggiore rilievo ai tokens più esclusivi, si propone la seguente interpretazione:

topic 1 *capo di Stato*: la presenza ripetuta del token *presid* sia tra le parole più frequenti sia tra i valori di FREX può indicare la descrizione dettagliata della figura più importante dello Stato: come viene eletto, la figura di vicepresidente, gli obblighi, ecc;

topic 2 *diritti e doveri dei cittadini*: soltanto in questo topic *shall* e *right* sono le prime due parole più frequenti. I token più esclusivi in questo caso sono *peopl* e *everyon* accompagnate da *freedom* e *protect*: i diritti e i doveri sono per tutti i cittadini, al fine di garantire libertà e protezione;

topic 3 *lavoro*: il terzo token è l'unico ad essere caratterizzato da parole come *worker*, *labor* e *product*, indicando quindi un topic incentrato sul sistema lavorativo;

Figura 3.3: Output del modello STM a 8 topic - le 15 parole più frequenti e le 15 parole con indice FREX più alto per ciascun topic

TOPIC	PAROLE PIU' FREQUENTI	FREX
1	shall presid member elect nation court may assembl govern state hous minist council repres person	shall repres upon hous pursuant vice-presid membership presid islam three agenc discharg presidenti submit major
2	shall right nation state republ assembl presid govern court peopl may public elect citizen organ	peopl everyon prosecutor ensur republ stipul protect freedom self-govern citizen democrat activ adopt intern implement
3	shall feder state public may establish nation accord author provid administr servic congress offic case	feder congress art entiti municip branch worker percent product correspond indigen plan system labor resourc
4	republ presid nation may right assembl public council court organ member state chamber function elect	chamber republ compos absolut mandat senat magistr organ text assur superior can defens promulg condit
5	may court provis shall minist order person relat assembl member made king matter appli subsect	king majesti laid kingdom crown european royal appli instrument paragraph committe order record connect far
6	parliament state claus may shall govern court council servic member provinci legisl region provinc provis	claus provinci provinc parliament legislatur union governor region commonwealth schedul proclam anyth chief south notwithstand
7	shall person offic may member court appoint public provis commiss function parliament servic minist presid	advic remov offic qualifi polic appeal holder prime question person act reason vacat appoint vacant
8	person court elector must may commiss shall parti subsect offic proceed judg elect candid made	elector must return candid parti notic registrar regist poll name roll district paper subsect new

topic 4 *suddivisione dei poteri*: le parole che maggiormente emergono in questo topic sono *chamber* e *senat* che rappresentano il potere legislativo e la struttura della sua gestione. Emergono anche parole come *promulg*, *mandat* e *magistr* che rafforzano il concetto di potere legislativo.

topic 5 il quinto topic è l'unico a non avere una chiara caratteristica rispetto agli altri topic risultando non ben definito. Emergono parole come *minist* e *king* che richiamano al struttura del ministero avvicinandosi come idea al topic 7;

topic 6 *organizzazione territoriale*: il sesto argomento è l'unico a essere caratterizzato da parole come *provinc* e *region* richiamando chiaramente il concetto di organizzazione e struttura territoriale e la suddivisione del potere nei sotto organi quali province e regioni;

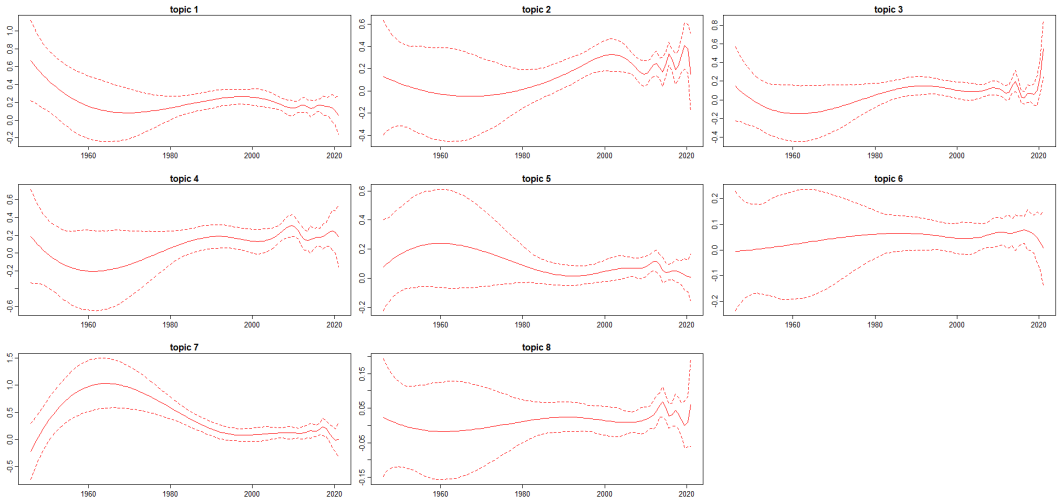
topic 7 *parlamento e ministero*: il topic 7 si avvicina come contenuto al topic 5 ma rimarcando in maniera più evidente la struttura e il funzionamento del

parlamento e del ministero attraverso i token *minist*, *parliament*, *presid* e *function*;

topic 8 *sistema elettorale*: l'ultimo topic descrive il sistema elettorale vigente all'interno del paese, chi si può candidare e come funzionano le elezioni. I token maggiormente di rilievo risultano essere *elector*, *candid* e *pool*.

Dal modello STM proposto emerge che ciascuno degli 8 topic sembra avere un proprio contenuto specifico, ad eccezione del quinto topic. Confrontando gli argomenti emersi dal topic modeling con la struttura di costituzione riportata al paragrafo 1.1 si possono notare alcune similitudini. I punti in comune con i topic risultano essere: diritti sociali ed economici (4), parlamento o legislatura (5), capo di Stato (6), governo (7) e governo sub-nazionale (9). I topic che si distanziano sono il topic 3 che però può essere inglobato all'interno dei diritti sociali ed economici (4) e il topic 8 che può essere in parte rappresentato dal punto (10) che tratta del referedum.

Figura 3.4: Effetto della covariata *anno* sui topic del modello STM



Infine, si vuole indagare l'effetto della covariata *anno* nella distribuzione di ciascun topic. In *Figura 3.4* sono stati riportati gli effetti della covariata sulla distribuzione di ciascun topic ottenuta come regressione della variabile *anno* sulla proporzione di ciascun topic utilizzata come risposta. La distribuzione relativa al topic 7 sembra mostrare un andamento oscillante nel tempo: l'anno di ultima revisione sembra avere un impatto visibile sulla distribuzione delle parole riguardante la struttura del parlamento. Questo risultato è supportato dalla significatività ($p - value < 0.01$) del primo nodo della spline di regressione per il topic 7. Effetti meno evidenti si osservano anche per i topic 1, 2 e 3: i topic 2 e 3 presentano una

leggera significatività ($p - value < 0.10$) per un solo nodo della spline (rispettivamente il nono per il topic 2 e il decimo per il topic 3) mentre il primo topic contiene una discreta significatività ($p - value < 0.05$) per il quarto, il sesto e il decimo nodo e una leggera significatività ($p - value < 0.10$) nel quinto, settimo e ottavo nodo della spline di regressione.

Capitolo 4

Modelli di regressione

L'obiettivo chiave di questa ricerca è quello di comprendere meglio se il testo costituzionale di ogni singolo paese rispecchia davvero il grado di democrazia che vige all'interno del medesimo Stato. Al netto di come una singola costituzione viene emanata, sia che rappresenti la libera espressione della maggioranza sia che venga imposta da un gruppo ristretto che ne detiene il potere, si vuole capire se vi è una relazione con la reale forma di governo.

4.1 Preprocessing: costruzione della matrice del disegno

In relazione alla domanda di ricerca, si imposta un problema di regressione il cui obiettivo è quello di riuscire a spiegare e prevedere l'indice di democrazia utilizzando come predittore il contenuto del testo costituzionale di ciascun paese. In questo capitolo viene utilizzato un approccio *bag of words*: ci si concentra sulle parole più frequenti perdendo la sequenzialità del testo e quindi anche significato specifico che ciascuna parola può assumere all'interno di un discorso. Il problema viene quindi semplificato nel chiedersi se la presenza e l'eventuale frequenza di un sottoinsieme ristretto di vocaboli che ciascuno Stato sceglie per descrivere sé stesso sia determinante per la previsione della forma di governo vigente.

La matrice del disegno utilizzata per il problema di regressione è composta dalle parole più frequenti presenti nella raccolta dei testi costituzionali. Il corpus che si considera per la fase di analisi è composto dalle 166 costituzioni di cui si possiede anche l'indice di democrazia, come illustrato al paragrafo 2.5. La costruzione della matrice delle covariate è stata effettuata con la libreria *quanteda* attraverso i seguenti passaggi:

- estrazione e pulizia dei tokens: dal corpus vengono estratti tutti i tokens ripuliti di punteggiatura, simboli e numeri;
- formattazione del carattere: tutti i caratteri dei tokens vengono messi in formato minuscolo;
- operazione di *stemming*: da ciascun token viene estratta la radice;
- creazione della Document-Term-Matrix (DTM): matrice in cui per colonna vi sono i token, per riga le osservazioni (le costituzioni dei 166 paesi) e in ciascuna cella la frequenza con cui il rispettivo token si presenta all'interno alla relativa costituzione;
- rimozione delle *stopwords*: vengono eliminate le parole vuote di significato dal testo quali pronomi, preposizioni e anche parole che ne descrivono la natura del corpus che si sta analizzando come *constitut*, *articol* e *section*[1];
- eliminazione dei tokens con lunghezza inferiore a tre caratteri;
- selezione dei tokens più frequenti: si imposta pari a 3000 la soglia di frequenza minima, tale scelta viene effettuata per ridurre la dimensionalità della matrice del disegno utilizzata in fase di modellazione.

Si ottiene in questo modo la Document-Term-Matrix relativa ai 128 tokens più frequenti all'interno del corpus. Il dataset finale considerato per le seguenti analisi si compone di 166 osservazioni e 129 variabili: 128 covariate e l'indice di democrazia come variabile risposta. Ciascuna covariata indica la frequenza assoluta con cui un determinato token compare all'interno di ciascuna costituzione e tutte le esplicative hanno come supporto l'insieme dei numeri naturali $\mathbf{N}$.

4.2 Scelte metodologiche

Tale problema di regressione viene impostato utilizzando un approccio di data mining: in questo caso l'obiettivo primario è quello di estrarre conoscenza dai dati per comprendere meglio la relazione tra le variabili considerate e allo stesso tempo quello di quantificare la capacità predittiva dei modelli utilizzati. Data l'esigua numerosità campionaria di 166 osservazioni, anche in relazione alle 129 covariate prese in considerazione, si decide di utilizzare un approccio di convalida incrociata per gestire il compromesso varianza-distorsione del modello [20]. Si considera

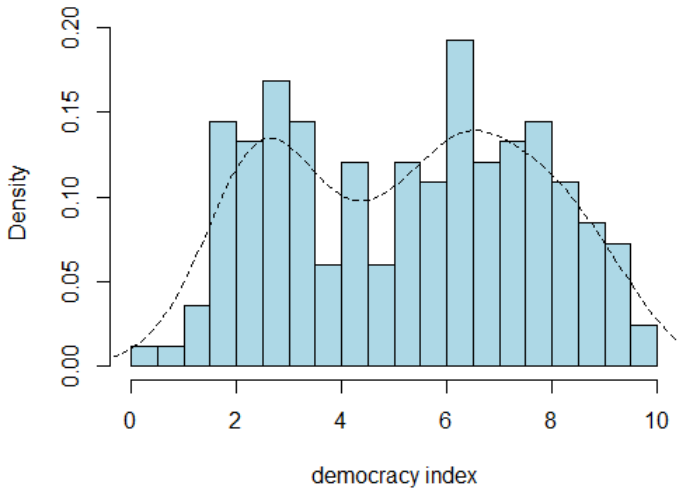
[1] per l'elenco completo delle stopwords si fa riferimento all'omonima libreria a cui sono state aggiunte: we'r, you'v, they'r, you'r, constitut, act, articl, constitutional, law, section, paragraph

quindi la suddivisione del dataset in 5 porzioni casuali: tutti i dati verranno utilizzati sia nel ruolo di stima (considerando 4 porzioni su 5) sia nel ruolo di verifica (nella restante porzione), ruotando di volta in volta il ruolo delle porzioni di dati. Le 5 differenti stime così ottenute verranno combinate assieme attraverso la loro media. In questo modo si evitano problemi di sovradattamento del modello ai dati e si può garantire comunque una generalizzazione dei risultati.

4.2.1 Distribuzione della variabile risposta

Si considera ora la variabile risposta quindi l'indice di democrazia descritto al paragrafo 1.2.1. La distribuzione marginale della variabile presenta un andamento bimodale, i cui valori massimi corrispondono ai punteggi di 3 e 6, come si osserva in *Figura 4.1*. Il punteggio medio dell'indice di democrazia dell'interno del corpus è pari a 5.23 e il valore mediano è di 5.45 che in entrambi i casi sono valori centrali rispetto al supporto della variabile [0, 10]. Tale distribuzione non rispecchia in maniera ottimale un andamento normale a causa della bi-modalità, sebbene sia piuttosto simmetrica con una deflessione nei valori centrali.

Figura 4.1: Distribuzione marginale della variabile risposta



4.2.2 Metriche di accuratezza del modello

Per valutare la scelta del modello di regressione migliore si utilizzano alcune metriche che valutano la capacità predittiva del modello rispetto alla variabile risposta. Le metriche di errore utilizzate per la valutazione dei modelli sono le seguenti:

- MSE: Mean Square Error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

quantifica l'errore medio al quadrato tra i valori predetti $\tilde{y}_i$ dal modello e i valori osservati y_i. Tale misura non è direttamente interpretabile rispetto all'ordine di grandezza della variabile risposta ed è particolarmente sensibile agli outliers;

- RMSE: Root Mean Square Error

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}$$

direttamente ottenuto dal MSE per riportare alla medesima scala della variabile risposta;

- MAE: Mean Absolute Error

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \tilde{y}_i|$$

misura l'errore medio tra le previsioni $\tilde{y}_i$ e i valori osservati y_i indipendentemente dal loro segno. Tale metrica non è influenzata dall'eventuale presenza di valori anomali ed è direttamente interpretabile nella stessa scala della variabile y.

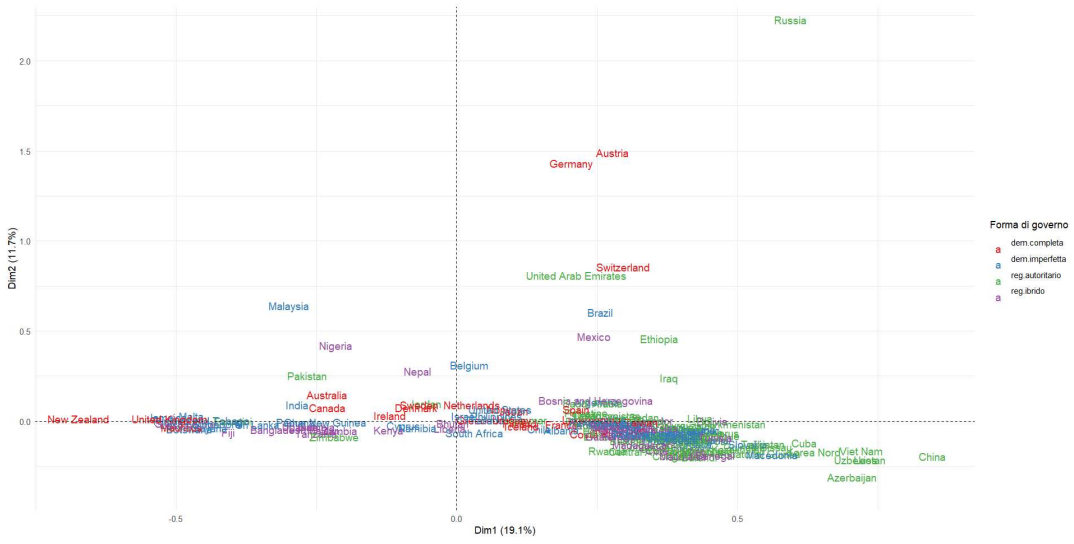
I modelli in seguito adattati vengono quindi confrontanti rispetto a tali metriche, scegliendo il modello che minimizza gli errori. Il calcolo degli errori di previsione è da considerare all'interno di un approccio di convalida incrociata: le metriche finali saranno quindi la media dei 5 errori ottenuti facendo ruotare nei 5 modi possibili il ruolo di stima e di verifica di ciascuna porzione di dati. Sebbene la scelta dei modelli migliori venga effettuata rispetto alla capacità predittiva dei modelli, si predilige l'aspetto interpretativo per rispondere alla domanda di ricerca.

4.3 Analisi esplorative bivariate

Un metodo esplorativo che non è stato presentato nel Capitolo 2 per confrontare i testi costituzionali è l'analisi delle corrispondenze. L'idea è quella di ridurre la

dimensione delle variabili con una perdita contenuta di informazione. Si utilizza quindi l'intera matrice del disegno con cui ottenere le coordinate dell'analisi delle corrispondenze che fungono da nuove variabili. Tali variabili artificiali cercano di riassumere l'intreccio di relazioni di interdipendenza tra le variabili originali quali la frequenza delle parole principali del corpus di testi. In particolare, a partire dalla matrice X vengono calcolate le frequenze relative rispetto a ciascuna covariata e si calcolano i marginali di riga e di colonna. Si calcola quindi la matrice delle frequenze attese assumendo indipendenza tra osservazioni e variabili; si ottiene poi la matrice dei residui $E = A - F$ ottenuta come differenza tra la matrice delle frequenze attese A e di quelle osservate F. A questo punto la matrice dei residui E viene pesata rispetto ai marginali di riga e di colonna e su questa nuova matrice S ottenuta si effettua la decomposizione a valori singolari (SVD). La matrice S viene decomposta come $S = U\Sigma V^T$, dove U_{nxn} contiene le coordinate delle osservazioni, V_{pxp} le coordinate delle variabili e Σ_{nxp} è la matrice dei valori singolari che catturano la quantità di varianza spiegata dalla rispettiva dimensione.

Figura 4.2: Analisi delle corrispondenze delle parole più frequenti - prime due dimensioni rispetto alla tipologia di forma di governo



Si ottengono quindi le coordinate principali F_n per ciascuna delle n osservazioni come il prodotto dato da $F_n = W_n\ U\ \Sigma$, dove W_n è la matrice diagonale dei pesi di ciascuna riga. Si visualizzano graficamente le prime due coordinate che catturano rispettivamente il 19.1% e l'11.7% della varianza dei dati e le relative coordinate di ciascuna osservazione. Più precisamente, si vogliono mettere in relazione le coordinate dei primi due assi fattoriali con la tipologia della forma di governo. Si

colorano quindi diversamente i paesi rispetto alla forma di governo vigente utilizzando la suddivisione proposta nel paragrafo 1.2.1 nelle quattro categorie. Si può osservare in *Figura 4.2* che gli Stati caratterizzati da un regime autoritario sono raggruppati nella parte in basso a destra anche se con alcune eccezioni. Per le altre forme di governo non vi è invece una chiara distinzione grafica, distribuendosi in maniera piuttosto eterogenea all'interno dello spazio. Tuttavia, sembra che la combinazione delle parole più frequenti a livello testuale possa avere una relazione con la forma di governo vigente. A partire da queste analisi esplorative si cerca quindi un modello di regressione che possa cogliere più dettagliatamente questa relazione.

In *Appendice A* nelle *Figure A.9* e *A.10* sono riportati due zoom della *Figura 4.2* per una lettura migliore. Inoltre, è presente anche il barplot della varianza cumulata in forma percentuale per le prime 50 dimensioni: si osserva che con la 50-esima dimensione si arriva a coprire quasi il 94% della varianza totale (*Figura A.11* dell'*Appendice A*).

4.4 Modelli applicati

In questo capitolo si vogliono presentare le principali tecniche di regressione utilizzate. Sono modelli che cercano di cogliere relazioni complesse tra le covariate e la variabile risposta al fine di riuscire a sapere se il contenuto del testo costituzione funge da buon predittore sulla forma di governo vigente all'interno del paese.

Nelle analisi qui descritte si utilizza come matrice del disegno X la matrice ottenuta a seguito del pre-processing, in cui le covariate rappresentano i tokens principali del corpus di testi ed esprimono la frequenza assoluta con cui tali tokens compaiono all'interno di ciascun testo costituzionale.

4.4.1 Modello MARS

Le spline di regressione multidimensionali adattive (MARS) sono il primo metodo di regressione adattato, proposto da Friedman nel 1991 [21]. Nello specifico, tale modello può essere considerato come un caso particolare delle spline di regressione e permette di utilizzare molteplici variabili esplicative. Per la costruzione del modello, si considerano le coppie di funzioni lineari a tratti del tipo $(x-\varepsilon)_+$ e $(\varepsilon-x)_+$, con un solo nodo nel punto ε, dette basi. Si definisce così l'insieme di basi di funzioni $\mathbf{C} = \{(x_j-\varepsilon)_+, (\varepsilon-x_j)_+\}$, con $\varepsilon \in \{x_{1j}, x_{2j}, ..., x_{pj}\}$ e $j = 1, 2, ..., p$.

Il modello MARS può essere espresso come:

$$f(x) = \beta_0 + \sum_{k=1}^{M} \beta_m h_m(x)$$

dove $h_m(x)$ sono le funzioni appartenenti a $\mathbf{C}$ o prodotti di almeno due di tali funzioni. Fissate le funzioni $h_m(x)$, i parametri β_m vengono stimati attraverso il criterio dei minimi quadrati. La stima del modello si concretizza quindi nella scelta di quali e di quante funzioni di base $h_m(x)$ considerare. Tale procedura è composta da due fasi: fase di crescita e fase di potatura, analogamente a quanto accade per la stima di un modello ad albero.

Fase di crescita

La fase di crescita nella stima di un modello MARS è una procedura iterativa che parte dal modello nullo con $M = 0$ e quindi con la sola intercetta β_0 inclusa. Sia M l'insieme delle funzioni di base incluse nel modello; ad ogni step $M+1$ si sceglie una nuova coppia di funzioni del tipo:

$$\widehat{\beta_{M+1}}\ h_l(x)\ (x_j - t)_+ + \widehat{\beta_{M+2}}\ h_l(x)\ (t - x_j)_+$$

dove $h_l(x) \in \mathbf{C}$, in modo che venga minimizzata la devianza residua. Si procede iterativamente fino ad includere un numero massimo prefissato di termini, ottenendo quindi un modello volutamente sovradattato.

Fase di potatura

Nella fase di potatura si effettua la regolazione del numero di prodotti tensoriali da includere nel modello, andando a sottrarre termini dal modello sovradattato in fase di crescita. Tra i diversi metodi che vi sono per fare regolazione in fase di potatura, si sceglie di utilizzare il metodo della convalida incrociata generalizzata (GCV), un'approssimazione computazionalmente efficiente della tecnica di *leave one out cross validation* per stimatori non lineari. Tale criterio si definisce come:

$$GCV(\lambda) = \frac{\sum_{i=1}^{n} \{y_i - f_\lambda(x_i)\}^2}{\{1 - d(\lambda)/n\}^2}$$

dove $d(\lambda)$ rappresenta una misura di complessità del modello ($d(\lambda)/n$ una misura di complessità *media* di ciascuna osservazione nel modello) e $f_\lambda(x_i)$ la previsione del modello adattato per la i-esima osservazione. Vengono quindi eliminati in maniera sequenziale termini dal modello sovradattato in modo da minimizzare l'errore complessivo.

Applicazione del modello ai dati

Questo modello di regressione è stato implementato utilizzando la libreria *polspline*. La regolazione del modello è stata effettuata su una griglia di valori del parametro *gcv* che consente di determinare la complessità ottimale del modello[2]. Valori alti di *gcv* producono un modello semplice che potrebbe portare a un *underfitting* mentre valori piccoli di *gcv* portano a un modello più complesso che potrebbe significare un *overfitting* del modello rispetto ai dati. In particolare, è stato effettuato un approccio di convalida incrociata con cui viene fatta sia la regolazione del modello su una griglia di valori di *gcv* sia è stato calcolato l'errore di previsione medio tra i 5 gruppi. É stato poi selezionato il modello associato all'errore medio di previsione minimo che corrisponde a un valore di *gcv* pari a 2.

4.4.2 Modello Bagging

Il modello Bagging, o boostrap aggregation, è un metodo di combinazione di modelli di previsione più semplici. In questo contesto, il modello base scelto è l'albero di regressione; per una documentazione più estesa si fa riferimento al libro *Classification And Regression Tree* di Breiman et al. (2017) [22]. Il processo di aggregazione di tanti alberi di regressione da un lato aumenta il costo computazionale ma allo stesso tempo consente di stabilizzare le stime fornendo previsioni più accurate e robuste diminuendo la varianza del modello finale.

Per la costruzione del modello, si effettuano B campioni casuali di numerosità n con reinserimento dal dataset originale. Su ciascuno dei campioni bootstrap così ottenuti si adatta un albero di regressione $f_b^A(x)$ con $b = 1, ..., B$. Il modello Bagging finale è ottenuto come la media dei predittori semplici, espresso come:

$$f_{bagging}(x) = \frac{1}{B} \sum_{b=1}^{B} f_b^A(x)$$

Per ciascuno dei B alberi adattati, si ottiene per costruzione una porzione di dati che vengono esclusi dal processo di stima; tale insieme prende il nome di *out-of-bag*. In questo modo non occorre utilizzare un approccio di convalida incrociata per calcolare l'errore di previsione commesso dal modello bensì si utilizza tale porzione di dati come se fosse un vero e proprio insieme di verifica. Il parametro di regolazione B, il numero di alberi utilizzati, viene quindi scelto in modo che si stabilizzi l'errore nell'insieme *out-of-bag*.

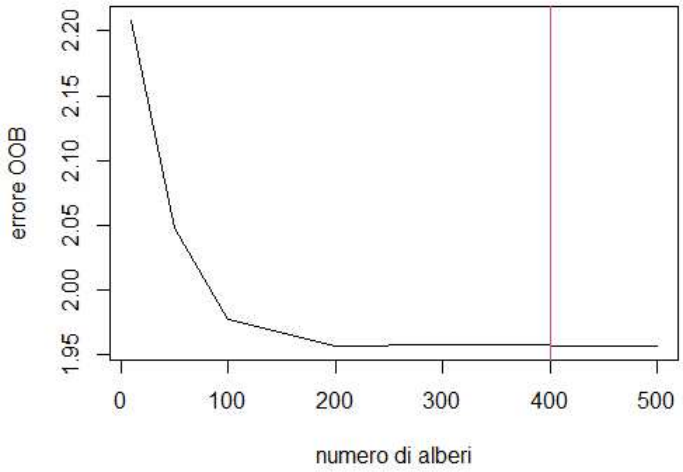
[2] il vettore di valori completo del parametro *gcv* utilizzato è: 2,3,4,5,6.

Il modello assume che gli alberi di cui si compone non siano correlati tra loro e generalmente si utilizzano alberi piuttosto grandi. É l'operazione di media che riduce la varianza del modello (per la legge dei Grandi Numeri) senza aumentarne la distorsione. Questa procedura aumenta notevolmente il costo computazionale rispetto al singolo albero di regressione ma è possibile procedere in parallelo per stimare i B modelli semplici di cui è composto, poiché sono indipendenti gli uni dagli altri. Tuttavia, si perde ogni interpretabilità diretta del modello.

Applicazione del modello ai dati

Il seguente modello è stato ottenuto con l'ausilio della libreria *ipred*. É stata effettuata la regolazione del modello utilizzando l'insieme di *out-of-bag* per calcolare l'errore di previsione commesso. In *Figura 4.3* si riporta l'andamento dell'errore quadratico medio calcolato sull'insieme *out-of-bag* rispetto al numero B di alberi adattati. Si fissa il parametro di regolazione pari a 400 poiché l'errore di stabilizza ad un valore pari a 1.96 circa.

Figura 4.3: Errore di previsione sull'insieme *out-of-bag* per numero di alberi che compongono il modello



Si calcolano poi, per ciascuna metrica, gli errori medi di previsione utilizzando un approccio di convalida incrociata a 5 gruppi.

4.4.3 Modello Random Forest

Un'altra maniera per poter combinare modelli semplici è data dal modello Random Forest; anche in questo caso viene scelto come modello base l'albero di regressione.

In questa procedura, la costruzione di ciascun albero che andrà a comporre il modello finale avverrà su un sottoinsieme F di variabili esplicative. In particolare, nella formulazione qui utilizzata, si considera l'approccio del modello Bagging descritto sopra a cui si aggiunge anche il campionamento dei predittori. La costruzione di un modello Random Forest avviene attraverso i seguenti passaggi:

1. si effettuano B campioni casuali di numerosità n con reinserimento dal dataset completo (analogamente a quanto accade per il modello Bagging);

2. per ogni b-esimo campione si costruisce un albero di regressione facendolo crescere fino alla sua massima dimensione senza poi essere potato;

3. la costruzione di ciascuno dei B alberi avviene considerando ad ogni nodo un piccolo gruppo F di variabili esplicative scelte casualmente. Ad ogni nodo il sottoinsieme di covariate viene cambiato, effettuando di volta in volta, un campionamento senza reinserimento tra le variabili disponibili;

4. si ottengono così B alberi volutamente sovradattati. Si effettua anche in questo caso un'operazione di media per ridurre la varianza del modello e migliorare l'accuratezza delle previsioni.

Il modello Random Forest finale si presenta come:

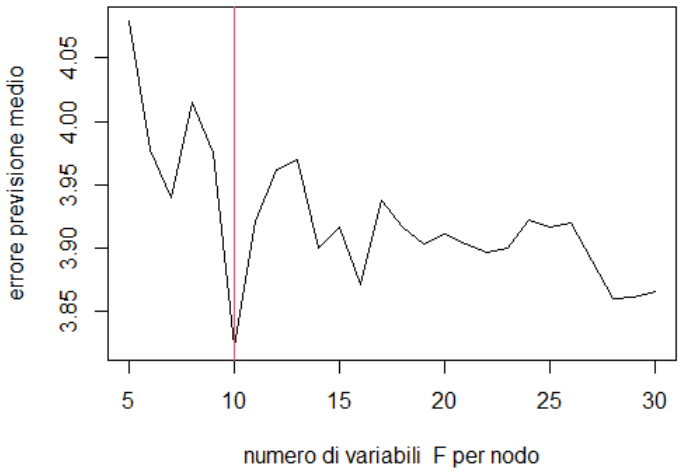
$$f_{randomforest}(x) = \frac{1}{B}\sum_{b=1}^{B} f_b^A(x)$$

analogo nella forma al modello Bagging, diverso nella sua costruzione. I parametri di regolazione risultano essere il numero di alberi B che compongono la foresta e il numero di variabili F tra cui ispezionare in fase di costruzione del modello.

Applicazione del modello ai dati

Contestualmente alla libreria *randomForest*, è stato scelto di fissare pari a 500 il numero B di alberi in modo che si possa garantire una stabilità dell'errore di previsione, anche alla luce dei risultati ottenuti col modello Bagging. Il numero F di variabili tra cui poter ispezionare in ciascun nodo durante il processo di stima del modello è stato regolato tramite convalida incrociata. In *Figura 4.4* si osserva che l'errore mimino medio di previsione corrisponde a un valore di F pari a 10 variabili.

Figura 4.4: Errore di previsione sull'insieme *out-of-bag* rispetto al numero di variabili F



4.4.4 Modello Gradient Boosting

Il modello Gradient Boosting è un terzo metodo di combinazione di previsioni che unisce l'idea del *boosting* a un algoritmo di discesa del gradiente [23]. Anche in questo caso si considera l'albero di regressione come modello base, detto predittore debole; si presenta quindi l'idea del modello Gradient Tree Boosting.

Si definisce un'opportuna funzione di perdita differenziabile $L(y,F(x))$ che misura la divergenza tra i valori osservati y e i valori predetti $F(x)$. Si considera in questo caso la funzione di perdita quadratica $L(y,F(x)) = \frac{1}{2}(y-F(x))^2$. L'algoritmo di stima del modello è iterativo, con l'obiettivo di minimizzare a ogni iterazione la funzione di perdita tramite approssimazioni consecutive del suo gradiente negativo. Ogni albero adattato sequenzialmente ha l'obiettivo di migliorare la previsione rispetto al passo precedente. Il modello finale Gradient Tree Boosting dopo M iterazioni, si presenta nella forma:

$$F_M(x) = F_0(x) + \sum_{m=1}^{M} \nu \, \gamma_m \, h_m(x)$$

dove:

- $F_0(x)$ è il modello iniziale, in questo caso dato dalla media della variabile risposta $F_0(x) = \frac{1}{n}\sum_{i=1}^{n} y_i$;

- $\nu \in (0,1)$ è un parametro di regolazione, detto tasso di apprendimento, che serve per prevenire il rischio di sovradattamento del modello;

- $h_m(x)$ è il generico albero di regressione adattato allo step m-esimo dell'algoritmo;

- γ_m è il peso associato al modello $h_m(x)$, calcolato ottimizzando la funzione di perdita $L(y, F(x))$.

Ad ogni iterazione del modello $m = 1, ..., M$ si adatta un albero di regressione $h_m(x)$ per approssimare i residui del modello corrente, in quanto coincidono con il gradiente negativo nel caso di funzione di perdita quadratica. Si prosegue poi con l'aggiornamento dei pesi γ_m e la stima del modello per la m-esima iterazione come $F_m(x) = F_{m-1}(x) + \nu\ \gamma_m\ h_m(x)$. In questo modo, il modello si propone di migliorare progressivamente la capacità predittiva arrestando l'algoritmo quando l'errore di previsione si stabilizza attorno ad un determinato valore. É quindi un modello con un costo computazionale elevato dato il discreto numero di parametri da regolare.

Applicazione del modello ai dati

La stima del modello Gradient Boosting include la regolazione di molteplici parametri di regolazione quali: il numero M di iterazioni, la profondità massima con cui viene stimato ciascun albero, il tasso di apprendimento ν e γ che controlla la riduzione della funzione di perdita. Tale procedura avviene tramite convalida incrociata con cui si stima l'errore di previsione medio delle 5 stime effettuate e si sceglie la combinazione dei parametri associata al valore minimo. La procedura porta alla selezione del modello con M pari a 500, una profondità massima di 3 nodi, un tasso di apprendimento ν pari a 0.1 e un valori di γ pari a 1[3]. La modellazione è stata effettuata utilizzando le librerie *caret* e *xgboost*.

4.4.5 Modello SVR con kernel radiale

Il modello Support Vector Regression (SVR) è un modifica dell'analogo modello di classificazione, presentato nell'articolo *Support vector networks* di Cortes & Vapnik (1995) [24]. La forma del modello che si vuole costruire è del tipo:

$$y = \langle w, \phi(x) \rangle + b$$

[3]Qui di seguito sono riportati i valori considerati in fase di regolazione: $M = \{100, 200, 300, 400, 500, 600\}$, $profondita\_massima = \{3, 6, 9\}$, $\nu = \{0.01, 0.1, 0.3\}$ e $\gamma = \{0, 1, 5\}$.

dove w è un vettore di pesi e $\phi(x)$ è una trasformazione non lineare della matrice del disegno X. Per tale problema di regressione, si utilizza una funzione di perdita del tipo:

$$L_\varepsilon(y_i, f(x_i)) = \begin{cases} 0 & se\ |y_i - f(x_i)| \leq \varepsilon \\ |y_i - f(x_i)| - \varepsilon & altrimenti \end{cases}$$

dove y_i sono i valori osservati, $f(x_i)$ i valori previsti dal modello ed ε una soglia che definisce la tolleranza entro la quale gli errori di previsione non sono penalizzati. In particolare, se la differenza in modulo tra il valore osservato y_i e il valore previsto $f(x_i)$ è minore di ε allora la funzione di perdita vale 0, se è maggiore di ε allora la funzione di perdita ammonta a tale differenza riscalata rispetto ad ε. Il problema di minimizzazione può essere formulato come:

$$\min_{w,b,\zeta} \frac{1}{2} ||w||^2 + \gamma \sum_{i=1}^{n} \zeta_i$$

soggetto ai seguenti vincoli:

$$|y_i - f(x_i)| < \varepsilon - \zeta_i$$

$$\zeta_i \geq 0$$

dove γ è un parametro di regolazione che rappresenta il costo di violazione dei margini e ζ_i è una variabile ausiliaria che indica se e quanto l'i-esima osservazione cade al di fuori del margine ε. La soluzione del problema di minimo passa attraverso la rispettiva formulazione in forma duale che consente di esprimere il modello finale come:

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) + b$$

dove α_i sono i moltiplicatori di Lagrange associati alle y_i e $K(x_i, x)$ detta *funzione nucleo* calcola il prodotto interno delle covariate o di eventuali trasformate. Nello specifico, in questo caso si utilizza un kernel radiale definito come:

$$K(x_i, x) = exp(-\sigma\ ||x_i - x||^2)$$

dove $\sigma > 0$ controlla l'ampiezza del kernel.

Applicazione del modello ai dati

La stima di quest'ultimo modello di regressione è avvenuta attraverso le librerie *caret* e *e1071* di *R*. I parametri da regolarizzare sono γ il costo di violazione dei margini e σ l'ampiezza del kernel radiale. Tale operazione viene realizzata attraverso convalida incrociata su una griglia di valori[4]. Si ottiene un errore medio minimo corrispondente alla combinazione di γ pari a 8 e σ pari a 0.016.

4.5 Risultati e confronti

Si vogliono ora confrontare i modelli adattati sulla base degli errori di previsione commessi. In *Tabella 4.1* sono riportati i valori delle metriche di accuratezza presentate nel paragrafo 4.2.2. Tali errori sono ottenuti come valori medi dei 5 errori calcolati ruotando di volta in volta il ruolo dei dati tramite convalida incrociata. Si osserva che le tre diverse metriche sono concordi e il modello migliore in termini di errore di previsione è il modello MARS con un MAE pari a 1.279. Anche la Support Vector Regression e il Gradient Boosting hanno un valore non superiore a 1.50 in termini di errore medio assoluto. L'indice di democrazia è un indice costruito con supporto $[0, 10]$ e all'interno del dataset completo assume un valore minimo pari a 0.26 e un massimo di 9.81. I modelli qui adattati riescono a prevedere il punteggio di tale indice con un errore medio assoluto di 1.279 nel migliore dei casi. La presenza e la frequenza delle parole più ricorrenti nei testi costituzionali forniscono quindi una discreta indicazione sulla forma di governo vigente all'interno dello Stato.

Tabella 4.1: Metriche di accuratezza dei modelli - valori medi ottenuti tramite convalida incrociata

modello	**MSE**	**RMSE**	**MAE**
MARS	2.466	1.568	**1.279**
Bagging	4.002	1.997	1.650
Random Forest	3.822	1.952	1.659
SVR radiale	3.344	1.819	**1.471**
Gradient Boosting	3.463	1.844	**1.504**

In fase di analisi sono stati adattati anche altre due tipologie di modelli di regressione quali il modello lineare Lasso e l'albero di regressione. Tali metodologie non sono state riportate per esteso poiché risultavano avere un valore superiore a

[4]La griglia di valori completa è data dalla combinazione dei seguenti vettori approssimati: $\gamma = \{0.03, 0.06, 0.125, 0.25, 0.5, 1, 2, 4, 8, 16\}$ e $\sigma = \{0.00049, 0.00098, 0.002, 0.004, 0.008, 0.016, 0.03, 0.0625, 0.125, 0.25, 0.5, 1\}$

2 in termini di errore medio assoluto, ottenuto tramite convalida incrociata. Per una documentazione più dettagliata, si fa riferimento agli articoli *Regression Shrinkage and Selection via the Lasso* di Tibshirani (1996)[25] per il modello Lasso e *Classification And Regression Trees* Breiman et al. (2017)[22] per l'albero di regressione.

Inoltre, per tener conto della diversa lunghezza dei testi costituzionali, sono stati adattati i medesimi modelli di regressione considerando come covariate le frequenze relative delle parole più ricorrenti. Tuttavia, i risultati così ottenuti non mostrano differenze rilevanti in termini di errore medio di previsione rispetto a quelli presentati. Infine, tutti i modelli qui proposti, sono stati replicati analogamente considerando come covariate le dimensioni dell'analisi delle corrispondenze presentate al paragrafo 4.3. I modelli così costruiti presentano un adattamento complessivo leggermente peggiore in termini di errore medio di previsione. Questi modelli non sono stati quindi presentati nel dettaglio poiché non hanno nemmeno un'interpretazione diretta delle variabili considerate in quanto costruite attraverso la decomposizione a valori singolari.

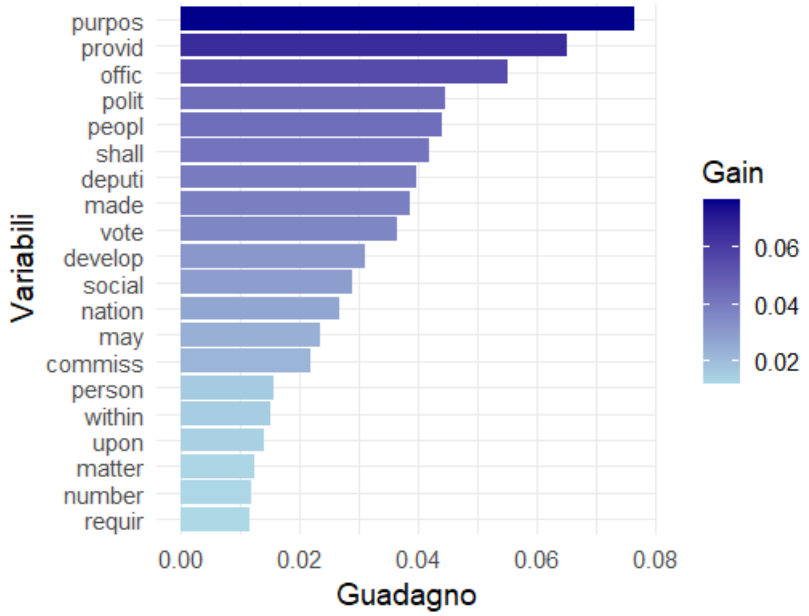
4.6 Importanza delle variabili

Si vuole a questo punto analizzare la composizione dei modelli qui proposti, andando ad osservare quali sono le variabili selezionate dai modelli migliori. Inoltre, tra le covariate incluse si vuole conoscere quali siano le più importanti con l'ottica di prevedere la forma di governo vigente nel paese.

I metodi di combinazione di modelli di previsione quali Bagging, Random Forest e Grandient Boosting possiedono per costruzione un grado di importanza delle variabili. Si considera, ad esempio, il modello Grandient Boosting e si ordinano in senso decrescente le variabili ritenute più importanti dal modello. Il grado di importanza di ciascuna esplicativa è espresso in termini di guadagno medio della funzione di perdita ottenuto ogni volta che ciascuna variabile viene utilizzata per uno split negli M alberi adattati. Questa misura esprime quanto una variabile migliora la previsione riducendo l'errore del modello. In *Figura 4.5* sono riportante le 20 variabili più importanti per il modello Gradient Boosting; *purpos* è il token più rilevante per la previsione dell'indice di democrazia, seguito da *provid*, *offic*, *polit* e *peopl*.

L'indice di democrazia utilizzato come variabile risposta è formato per costruzione da cinque macro aree quali: pluralismo e processo elettorale, funzionamento del governo, partecipazione politica, cultura politica e libertà civili (come si osserva

Figura 4.5: Importanza delle prime 20 variabili - modello Gradient Boosting



in *Figura 1.3*). Tra le radici che più richiamano il concetto di democrazia, *deputi* (l'equivalente di delegato o delega) può essere associato al concetto di separazione dei poteri (esecutivo, legislativo, giudiziario, ...) proprio e fondante una democrazia. Un altro token che può essere direttamente associato al concetto di partecipazione politica e processo elettorale è *vote*. Alla macro area riguardante le libertà civili si possono associare i tokens *peopl* e *social* in quanto sottolineano che l'aspetto sociale della cittadinanza è collegato a maggiore democrazia. Infine, la variabile *commiss* racchiude i significati di ufficio, di commissariato e di strutture specifiche quali basi militari che possono richiamare la macro area legata al funzionamento del governo.

In *Figura A.12* dell'appendice è riportato il grafico di importanza delle prime 20 variabili per il modello di regressione Random Forest. In questo caso, la misura di importanza delle variabili è in forma relativa e misura quanto l'impurità viene ridotta quando la rispettiva variabile viene utilizzata in uno split nella costruzione del modello. Essendo un problema di regressione, con impurità si intende la somma dei quadrati dei residui: maggiore è la riduzione dell'impurità che la variabile produce e maggiore sarà il valore di importanza. Si considera quindi, per ciascuna variabile, la somma totale della riduzione dell'impurità dei nodi di tutti gli alberi che compongono la foresta. Si trasformano poi i valori assoluti così ottenuti in termini relativi rispetto al valore massimo assunto. Per una documentazione più dettagliata si fa riferimento alle opzioni della funzione *randomForest* dell'omonima libreria.

In ottica di confronto di tali variabili rispetto a quelle ritenute importanti dal modello Gradient Boosting non vi sono molte differenze. Di interesse è la comparsa dei tokens *judg* e *organ* che racchiudono concetti riguardanti il funzionamento del governo e il diritto.

4.6.1 Valori SHAP

I modelli di regressione maggiormente di interesse, ovvero quelli che minimizzano l'errore di previsione, sono il modello MARS e il modello di Support Vector Regression che però non sono direttamente e facilmente interpretabili. Si richiama quindi una metodologia tratta dalla teoria dei giochi per trovare un modo di interpretare le variabili incluse nei modelli di regressione di interesse; tale metodo prende il nome di valori SHAP (SHapley Additive exPlanations). I valori Shapley sono un modo per quantificare il contributo di ciascuna esplicativa nel prevedere la variabile risposta, assumendo che le variabili siano indipendenti tra loro.

Per calcolare l'importanza dell'i-esima variabile si adatta il modello di interesse f_S per ogni possibile sottoinsieme di caratteristiche S che non includono la i-esima variabile. Si adatta l'analogo modello $f_{S\cup\{i\}}$ che include anche la variabile i-esima e si considera la differenza tra i valori previsti dai rispettivi modelli con e senza la i-esima covariata $f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)$. Tale operazione viene effettuata per tutti i possibili sotto insiemi $S \subseteq F \setminus \{i\}$ dove F è l'insieme di tutte le covariate. Si applica quindi una media ponderata di tutte le possibili differenze, tenendo conto della diversa numerosità dei sottoinsiemi S. Come proposto dall'articolo *A Unified Approach to Interpreting Model Predictions* di Lundberg e Lee (2017)[26] i valori Shapley si definiscono come:

$$\phi_i = \sum_{S\subseteq F\setminus\{i\}} \frac{|S|!\,(|F|-|S|-1)!}{|F|!}\ [f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)]$$

dove $|S|$ e $|F|$ sono rispettivamente le numerosità degli insiemi S e F. Il valore Shapley ϕ_i indica l'importanza complessiva della i-esima variabile rispetto a tutte le altre covariate. I valori ϕ_i possiedono molteplici proprietà, qui di seguito sono riportate le principali:

1. accuratezza locale: quando si approssima il modello originale f per un specifico valore x di input, l'accuratezza locale richiede che il modello di previsione $g(x')$ corrisponde al modello originale $f(x)$ quando $x = h_x(x')$, ovvero:

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i\, x'_i$$

dove ϕ_0 rappresenta il modello di previsione nullo, ovvero con nessuna covariata all'interno;

2. le variabili assenti dal modello hanno importanza nulla:

$$x_i' = 0 \quad \rightarrow \quad \phi_i = 0$$

3. linearità: se un modello di previsione è ottenuto come somma pesata di più modelli di previsione questa struttura di additività si mantiene anche per il calcolo dei valori Shapley:

$$se \quad f(x) = \alpha f_1(x) - \beta f_2(x) \quad allora \quad \phi_i = \alpha \phi_i(f_1) + \beta \phi_i(f_2)$$

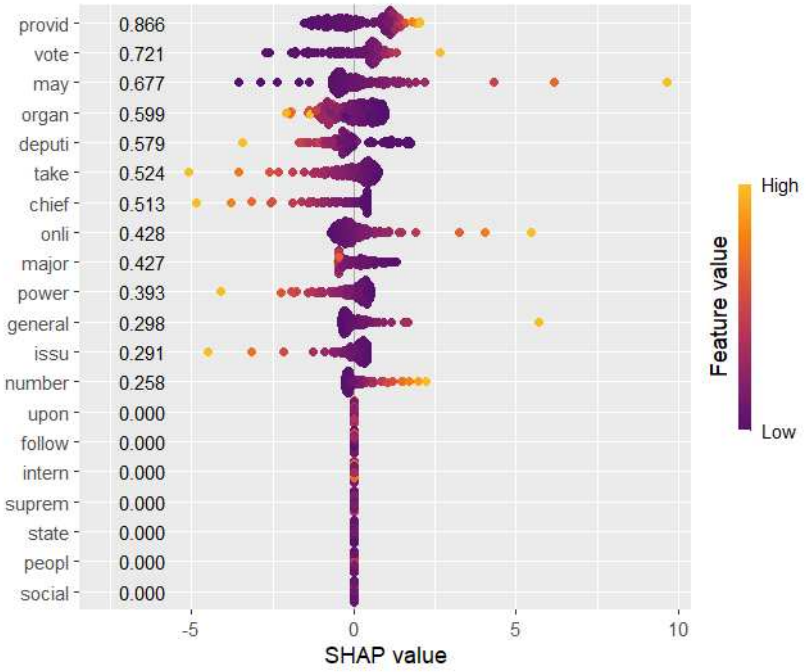
4.6.2 Kernel SHAP

Uno dei principali difetti della metodologia dei valori SHAP è l'ingente costo computazionali associato, soprattutto quando il numero di variabili coinvolte è molto elevato. Si utilizza perciò una modifica molto più efficiente rispetto alla formulazione classica che consente di non ricalcolare il modello in ogni sottoinsieme possibile delle variabili esplicative, riducendo così l'onere computazionale. Con questa metodologia, si considera un campionamento delle possibili configurazioni dato dai sottoinsiemi delle variabili coinvolte nel modello. In questo modo si esaminano soltanto le combinazioni realistiche di variabili esplicative, ottenendo una valutazione più accurata quando le covariate coinvolte sono correlate tra loro. Inoltre, questa modifica prevede l'utilizzo di una porzione di dati, chiamato *background dataset*, che viene utilizzato per rappresentare una distribuzione di riferimento per le covariate. In questo modo, quando si vuole simulare scenari in cui una o più variabili sono escluse dal modello si hanno comunque dei valori di riferimento reali per i dati di input. Si effettuano quindi le differenze tra i modelli di previsione con e senza la i-esima variabile per ciascuna delle configurazioni considerate.

Si utilizza la libreria *kernelshap* per il calcolo dei kernel SHAP nei modelli di regressione di interesse quali MARS e Support Vector Regression. Data l'esigua numerosità campionaria di 166 osservazioni, si considera come *backgroud dataset* un campionamento casuale di 40 osservazioni [27].

Sono riportati in *Figura 4.6* i valori SHAP del modello MARS, visualizzando le 20 variabili più importanti. Si osserva che 7 di queste hanno un valore nullo, sono infatti variabili che non sono incluse nel modello stimato ponendo quindi a 0 il grado della loro importanza. Le variabili incluse nel modello sono soltanto 13

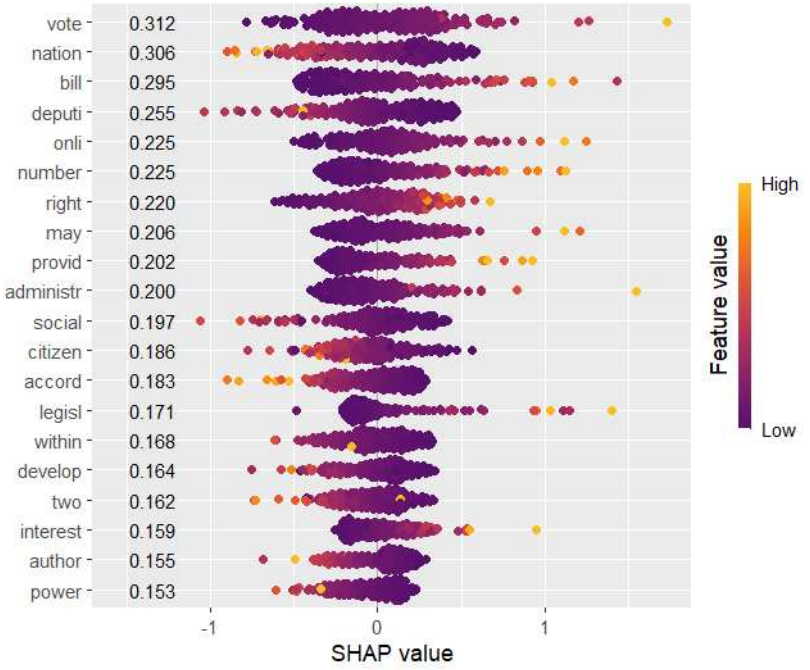
Figura 4.6: Valori SHAP - modello MARS



e il token maggiormente influente per prevedere l'indice di democrazia è *provid*, considerato altamente importante anche dal modello Gradient Boosting. Anche *vote*, *organ* e *deputi* sono tokens importanti nella previsione della forma di governo, come già emerso dai modelli di combinazione di alberi di regressione. La variabile *chief* risulta essere un predittore di discreta importanza e richiama il concetto di capo/presidente non necessariamente dello Stato ma anche in riferimento a coloro che possiedono un ruolo di vertice nella struttura del governo in tutte le sue sfaccettature.

Sono stati calcolati i valori kernel SHAP anche per il secondo modello migliore in termini di errore di previsione. In *Figura 4.7* sono riportati i valori Shapley delle 20 variabili più importanti per il modello Support Vector Regression. Si osserva che le variabili esplicative sono tutte incluse in questo modello e che assumono valori più bassi rispetto ai relativi valori del modello MARS. Anche in questo caso, le variabili esplicative più importanti nella previsione dell'indice di democrazia rispecchiano quanto visto in precedenza. I token più importanti sono *vote* e *nation* a cui si aggiunge anche *bill* che richiama il concetto di disegno di legge. Tra le altre variabili importanti e che richiamano il macro concetto di democrazia si ritrova anche *administr*, *social*, *legisl* ma soprattutto *right* che emerge soltanto in quest'ultima scala di importanza delle variabili.

Figura 4.7: Valori SHAP - modello Support Vector Regression



Capitolo 5

Large Language Models

Dal precedente capitolo emerge che la composizione dei testi costituzionali possa spiegare, seppur parzialmente, la forma di governo vigente all'interno del paese. I modelli di regressione proposti si basano su un approccio di *bag of words* (BoW), in cui i testi sono rappresentati dall'insieme casuale di parole che li compongono di cui si considerano soltanto le più frequenti. In particolare, la metodologia utilizzata è basata semplicemente sul calcolo della frequenza delle parole perdendo così la struttura e il senso del discorso. Di conseguenza, per prevedere l'indice dei democrazia, i modelli proposti si basano sulla presenza e l'eventuale frequenza di parole specifiche all'interno delle costituzioni. Sebbene tale approccio rappresenti un buon punto di partenza, i modelli basati sul BoW trascurano gran parte dell'informazione disponibile all'interno dei testi. Con l'obiettivo di superare i principali limiti delle precedenti analisi si decidere di modellare il corpus dei testi costituzionali con tecniche di elaborazione del linguaggio naturale (NLP).

In letteratura, le prime tecniche di superamento dell'approccio di *bag of words* sono quelle che trasformano il testo di input in una rappresentazione tramite vettori numerici tale che le parole con significato simile sono rappresentate da vettori numerici simili. Questo approccio prende il nome di *words embeddings* in cui ciascuna parola è rappresentata su uno spazio di coordinate multidimensionali in modo da poter cogliere anche il significato delle parole. I più comuni algoritmi di *static word embeddings* sono ad esempio *word2vec* di Mikolov et al. (2013) [28] e *GloVe* di Pennington et al. (2014) [29]. In sintesi, con questi modelli ciascuna parola viene rappresentata da un vettore numerico fisso che riflette il suo valore semantico. In questo modo, parole come *king* e *queen*, sebbene diverse morfologicamente, sono rappresentate da vettori di coordinate simili poiché condividono lo stesso significato ma con una diversa sfumatura.

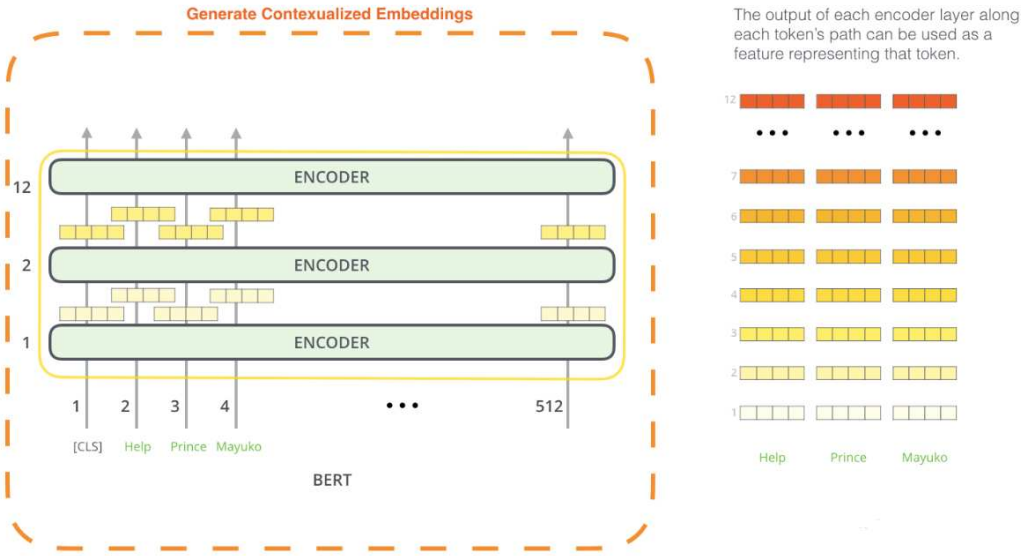
Per superare i limiti dei *word embeddings* statici, sono stati sviluppati i Lar-

ge Language Models (LLM) che permettono di rappresentare i testi tenendo conto anche del contesto specifico. Questo approccio si basa su *word embeddings* contestuali, dove ogni parola viene rappresentata in funzione del contesto in cui appare. Ciò è stato reso possibile grazie all'introduzione del modello Transformer, descritto per la prima volta da Vaswani et al. (2017) nel celebre articolo *Attention is All You Need* [30], che ha posto le basi per l'elaborazione avanzata dei dati e l'analisi generativa (di testi, immagini, audio, ecc.). Nei modelli linguistici di grandi dimensioni le parole sono rappresentate in funzione di un certo numero di termini che le precedono e che le seguono. In questo modo, frasi come "il giocatore è entrato in campo all'inizio del secondo tempo" e "qui non c'è campo, non riesco a telefonare" produrranno vettori molto diversi per la parola *campo* poiché assume significati differenti a seconda del contesto. Le rispettive rappresentazioni tramite *words embeddings* statici, invece, risulterebbero identiche in quanto non riescono a cogliere queste sfumature di significato.

5.1 Modello linguistico BERT

Il modello linguistico di grandi dimensioni utilizzato è BERT (Bidirectional Encoder Representations from Transformers) proposto da Google nel 2019 [31], nello specifico si fa riferimento alla versione di BERT-uncased-base. É un modello linguistico bidirezionale nel senso che ciascuna parola è calata nel contesto considerando la porzione di testo che la precede e la parte che la segue per una lunghezza massima di 512 tokens simultaneamente. Il modello BERT effettua la mappatura di un testo in input attraverso l'architettura di un Trasformer. Un modello Transformer è tipicamente costituito da due componenti principali: Encoder e Decoder. L'Encoder è responsabile della lettura e della comprensione del testo di input generando una rappresentazione vettoriale che riesce a cogliere il contesto e la relazione tra i token. Il Decoder utilizza la rappresentazione generata dall'Encoder per produrre un output quale, ad esempio, la generazione di un testo di risposta o la traduzione in un'altra lingua del testo originale. Il modello BERT utilizza soltanto la componente di Encoder del modello Transformer al fine di ottenere una rappresentazione numerica di un dato corpus di testi. In *Figura 5.1* si può osservare la struttura del modello BERT nella sua formulazione base: a partire dal testo di input vengono generati per ciascun token una sequenza di 12 layers ciascuno di 768 dimensioni. Ciascuno dei 12 layers è l'output di un modello di Encoder, ottenuti sequenzialmente andando a raffinare progressivamente la rappresentazione.

Figura 5.1: Struttura del modello BERT base - Fonte: [32]



Il modello BERT è ottenuto applicando sequenzialmente per 12 volte la componente di Encoder di un modello Transformer; il primo layer è ottenuto a partire dal testo di input mentre i successivi layers sono ottenuti a partire dal layer precedente. In particolare, il testo di input viene suddiviso in token attraverso un processo specifico chiamato WordPiece (Wu et al. 2016 [33]). Questa tipologia di tokenizzazione è *uncased* e *subword-based* poiché il testo viene portato tutto in carattere minuscolo e ciascuna parola può essere suddivisa anche in due token. Ad esempio la parola *playing* verrà suddivisa nei token *play* e *##ing*; in questo modo l'algoritmo riesce a trattare parole poco frequenti o morfologicamente complesse, mantenendo il vocabolario più compatto. La codifica del testo di input è sintetizzata in *Figura 5.2*; in rosa il testo di input a seguito del processo di tokenizzazione. In aggiunta vi sono anche due tokens speciali quali [CLS] e [SEP]: il primo indica l'inizio della sequenza (quindi del testo di input) e il secondo indica la fine della frase o la separazione tra due frasi. A partire dal testo così destrutturato, la codifica iniziale si scompone in tre parti:

- *token embeddings*: ogni token della sequenza viene rappresentato tramite un embeddings vettoriale di dimensione 768. Questa mappatura iniziale avviene grazie a una matrice di embeddings pre-addestrata in cui ogni token del vocabolario è associato a un vettore numerico. Il vocabolario di BERT base contiene circa 30.000 *words type* provenienti da BooksCorpus (800M parole) (Zhu et al., 2015 [34]) e *English Wikipedia* (2.500M parole);

- *segment embeddings*: ciascun token è attribuito alla frase cui appartiene per

tener conto delle relazioni strette tra le parole della stessa frase;

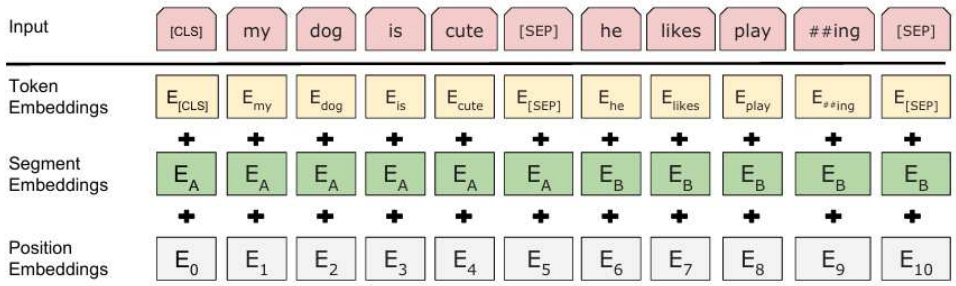
- *position embeddings*: è calcolato in base alla posizione del token, in particolare:

$$PE_{(pos,2i)} = sin\left(\frac{pos}{10000^{2i/768}}\right)$$
$$PE_{(pos,2i+1)} = cos\left(\frac{pos}{10000^{2i/768}}\right)$$

dove *pos* indica la posizione del token nell'intera sequenza, mentre *i* è indice della dimensione (quindi varia tra 0 e 768-1): *2i* è utilizzato per le dimensioni in posizione pari mentre *2i+1* per quelle dispari.

La rappresentazione iniziale di ciascun token è data dalla somma degli *token embeddings*, dei *segment* e dei *position* embeddings come si può osservare in *Figura 5.2*. La matrice finale così ottenuta viene utilizzata come input per il modello Encoder che produrrà il primo layer. A questo punto, il primo layer sarà la nuova matrice che verrà utilizzata nuovamente dall'Encoder per produrre il secondo layer, così sequenzialmente fino al 12-esimo layer.

Figura 5.2: Codifica del testo di input nel modello BERT base - Fonte: Devlin et al. (2019) [31]

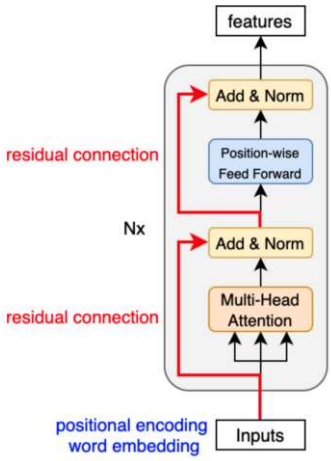


5.1.1 Modello Encoder e funzione self-attention

Nella seguente sezione si delinea brevemente il funzionamento del modello Encoder, al fine di comprendere le analisi effettuate nelle sezioni successive; per una spiegazione più dettagliata si fa riferimento agli articoli *Attention is All You Need* di Vaswani et al. (2017) [30] e *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* di Devlin et al. (2019) [31].

In *Figura 5.3* è riportata la struttura del modello Encoder, eseguito sequenzialmente $N = 12$ volte in BERT base. Il modello si compone di una prima parte

Figura 5.3: Struttura dell'Encoder di un Transformer - Fonte: sito web [35]



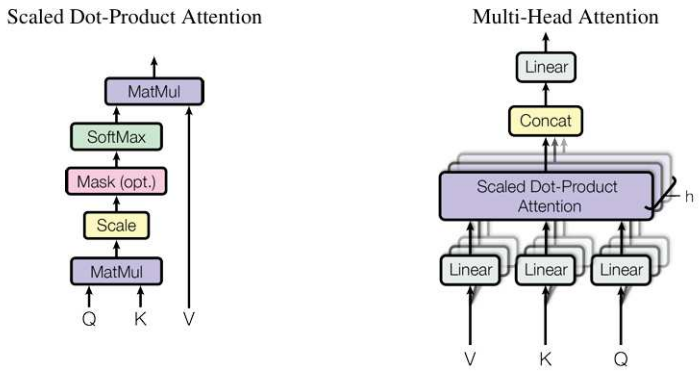
chiamata *Multi-Head Attention* formata dalla concatenazione di $h = 12$ funzioni di *self-attention*. La seconda componente del modello, chiamata *Position-wise Feed Forward Neetwork*, è una rete feed-forward a due strati che aggiunge la capacità di cogliere relazioni più sofisticate tra i token. Ciascuna di queste due componenti è seguita da una connessione con i residui e una procedura di normalizzazione. La connessione residua serve al modello per prevenire che l'informazione venga dissipata tra gli strati all'aumentare della profondità del modello, migliorando la capacità finale di rappresentazione dell'output. L'operazione di normalizzazione consente invece di stabilizzare la distribuzione dei valori e migliorare la velocità di convergenza del modello.

Multi-Head Attention

La componente di *Multi Head-Attention*, rappresentata nella parte destra della *Figura 5.3*, è formata dalla concatenazione di 12 funzioni di *self-attention*, come è illustrato nella parte sinistra della figura. In questo modo il modello consente di catturare relazioni complesse tra i token a vari livelli di contesto, estrapolando quindi un senso più generale dato dall'intreccio dei molteplici significati che compongono un testo.

Si supponga di avere un testo di input composto da n tokens, ciascuno rappresentato da un embedding di dimensione d_{model} pari a 768. La matrice iniziale del modello X è di dimensioni nxd_{model}. A questo punto, si vuole analizzare la struttura di una singola *Head*, che prende il nome di funzione *self-attention*. Per ciascuna $Head_i$, la matrice X viene pesata rispetto a tre dimensioni: Q (Query), J (Key) e V (value). Le rispettive matrici W_i^Q, W_i^K e W_i^V di dimensioni d_{model}x64 per ciascu-

Figura 5.4: Rappresentazione della Multi-Head Attention - Fonte: Vaswani et al. 2017 [30]



na *i*-esima $Head$ vengono calcolate nella fase di pre-addestramento del modello. Tramite il prodotto delle matrici dei pesi con la matrice X si ottengono:

$$Q_i = XW_i^Q \qquad K_i = XK_i^K \qquad V_i = XV_i^V$$

La funzione *self-attention* dell'*i*-esima $Head$ è espressa come:

$$Head_i = Attention\big(Q_i,\, K_i,\, V_i\big) = softmax\Big(\frac{Q_i\, K_i^T}{\sqrt{d_k}}\Big)\, V_i$$

dove d_k, nel caso di BERT, è pari a 64 ottenuto dal rapporto tra la dimensione degli embeddings e il numero di funzioni h (768/12=64) e serve a stabilizzare la distribuzione dei pesi per migliorare l'efficacia della funzione $softmax$. Il calcolo delle 12 matrici $Head$ avviene in parallelo e poi vengono concatenate tra loro:

$$Multi-Head(Q,\, K,\, V) = Concatenate(Head_1, ..., Head_{12})W^O$$

dove W^O è una matrice di pesi iniziali di dimensioni (h x d_k) x d_{model}, appresa durate la fase di pre-adattamento del modello.

Position-wise Feed Forward Network

La seconda componente del modello Encoder non è altro che una rete a due strati applicata a ciascun token in maniera indipendente. In questo modo, si garantisce una maggiore flessibilità al modello ricercando relazioni ricche e complesse mantenendo una buona efficacia poiché anche questa componente viene eseguita in parallelo. La rete è formata da due strati rappresentati da due trasformazioni lineari

connesse dalla funzione di attivazione *GELU* (Gaussian Error Linear Units). La rete feed-forward per il token x viene espressa come:

$$FFN(x) = GELU\left(xW_1 + b_1\right)W_2 + b_2$$

dove W_1 e W_2 sono le matrici dei pesi, rispettivamente per primo e del secondo strato della rete, b_1 e b_2 due vettori di errore. La funzione di attivazione è formulata come $GELU(x) = x\,\Phi(x)$ con $\Phi(x)$ che indica la funzione di distribuzione cumulativa normale standard. La dimensione del primo strato della rete consente ai token in input di aumentare la propria capacità di rappresentazione passando da 768 a 3072. L'applicazione del secondo strato riporta la dimensione di ciascun token alla propria dimensione originale di 768. Questa componente del modello introduce un elemento di non linearità che garantisce la capacità di comprendere relazioni complessi e strutture linguistiche.

5.2 Applicazione del modello ai dati

In questa sezione si vuole applicare il modello BERT al corpus dei testi costituzionali col fine di ottenere una rappresentazione completa dei testi per poi metterla in relazione con la forma di governo di ciascuno Stato. La modellazione è stata effettuata con la libreria *text* (Kjell et al. 2017 [36]) utilizzando il modello base pre-addestrato disponibile.

Per problemi computazionali e di risorse temporali limitate, è stato scelto di applicare il modello LLM soltanto a una porzione di dati ristretta. La numerosità campionaria passa quindi da 166 a 104 costituzioni, escludendo i testi più lunghi di 20.000 parole[1]. Per semplificare ulteriormente le analisi, si considera soltanto una porzione di testo equivalente alle prime 4.000 parole di ciascuna costituzione. Si assume quindi che la prima parte di ciascun documento contenga le principali informazioni e le diverse sfaccettature che definiscono la tipologia di forma di governo dichiarata a livello costituzionale. Per le 104 unità statistiche selezionate si effettua una troncatura del testo, ad eccezione della costituzione della Libia che conta soltanto 2.893 parole in totale. Si sottolinea che le seguenti analisi non pos-

[1] L'elenco esteso dei 62 paesi esclusi rispetto ai 166 considerati nei modelli di regressione del Capitolo 4 è il seguente: Albania, Angola, Austria, Bangladesh, Bolivia, Botswana, Brazil, Cape Verde, Chile, Colombia, Cyprus, Dominican Republic, Ecuador, El Salvador, Eswatini, Fiji, Gambia, Germany, Ghana, Greece, Guatemala, Guyana, Honduras, Hungary, India, Jamaica, Kenya, Lesotho, Malawi, Malaysia, Malta, Mauritius, Mexico, Mozambique, Myanmar, Namibia, Nepal, New Zealand, Nicaragua, Nigeria, Pakistan, Panama, Papua New Guinea, Paraguay, Philippines, Portugal, Sierra Leone, Singapore, South Africa, Sri Lanka, Sweden, Tanzania, Thailand, Trinidad and Tobago, Turkey, Uganda, Ukraine, United Kingdom, Uruguay, Venezuela, Zambia, Zimbabwe

sono essere generalizzabili in quanto la nuova numerosità campionaria, già esigua inizialmente, viene ulteriormente ridotta. Inoltre, la porzione di costituzione considerata per ciascun paese non è necessariamente esaustiva rispetto alla descrizione della forma di governo auto dichiarata da ciascuno Stato nel testo costituzionale completo. Si interpretano le seguenti analisi come un esperimento per valutare se un LLM adattato al corpus dei testi costituzionali possa aiutare a migliorare la stima dell'indice di democrazia.

5.2.1 Output del modello

La nuova rappresentazione matriciale del corpus dei testi costituzionali è ottenuta applicando la funzione *textEmbed* di *text*. Sono stati salvati tutti e 12 i layers creati dal modello ad esclusione del primo poiché per costruzione è decontestualizzato. Questa scelta viene effettuata per valutare se i *words embeddings* dinamici associati a diversi strati del modello hanno un ruolo e un'importanza differente, coerentemente a quando emerge dall'articolo *How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings* di Ethayarajh e Kawin 2019 [37]. In particolare, la funzione *textEmbed* viene applicata separatamente a ciascuno dei 104 testi ridotti. Per ognuno di essi, si ottiene in risposta una lista formata da due oggetti: la matrice [tokens] e il vettore [texts]. La matrice [tokens] contiene i valori delle 8448 dimensioni per ciascun token del testo, mentre il vettore [texts] contiene il valore medio di tutti i tokens del testo per ciascuna delle 8448 dimensioni. La matrice finale, di dimensione 104x8448, è ottenuta dalla concatenazione di tutti i vettori [texts] dei 104 Stati. Pertanto, il corpus di testi viene rappresentato quindi con una matrice di output di dimensione 104x8448, dove le colonne raccolgono le 768 dimensioni di ciascun layer (768x11 = 8448). In questo modo, il contenuto di ciascun testo viene sintetizzato tramite un'espressione numerica basata sulla comprensione del significato sottostante.

5.3 Modelli di regressione con word embeddings dinamici

In questa sezione vengono replicati i modelli di regressione e le scelte metodologiche rispetto a quanto visto nel Capitolo 4. Alla matrice finale descritta nel paragrafo 5.2.1 e di dimensioni 104x8448 viene applicata una riduzione della dimensionalità prima di adattare i modelli di regressione. Per mantenere e valutare la struttura dei layers proposta dal modello BERT, si decide di estrarre le componenti principali

delle dimensioni di ciascun layer. Si effettuano quindi 11 proiezioni differenti su ciascuna matrice 104x768 relative agli 11 layers salvati dall'applicazione del modello BERT. Per ciascun layer si decide di estrarre le prime *c* componenti principali che coprono il 90% della varianza totale spiegata. Come si osserva in *Tabella 5.1*, il numero *c* è diverso per ciascun layer, ottenendo una matrice finale di dimensione 104x347 che costituisce la matrice del disegno.

Tabella 5.1: Composizione della matrice del disegno - numero di componenti principali *c* per ciascun layer

Layers	2	3	4	5	6	7	8	9	10	11	12	**Totale**
c	32	33	33	33	32	32	32	29	30	30	31	**347**

Si applicano quindi i modelli di regressione MARS, Bagging, Random Forest, SVR con kernel radiale e Gradient Boosting utilizzando un approccio di convalida incrociata a 5 gruppi. Nella *Tabella 5.2* sono riportate le metriche di accuratezza dei rispettivi modelli; si può osservare che il modello migliore risulta essere ancora una volta il MARS con un MAE pari a 0.934. Rispetto al relativo modello di regressione stimato nel capitolo precedente, in cui le variabili esplicative rappresentavano la frequenza delle parole più ricorrenti, si ottiene un miglioramento di 0.345 punti sull'errore medio assoluto commesso. Le metriche mostrano un leggero miglioramento anche per i modelli SVR e Gradient Boosting mentre per i due metodi di combinazione di alberi si registra un leggero peggioramento. Il contenuto costituzionale elaborato tramite un LLM sembra migliorare la stima dell'indice di democrazia.

Tabella 5.2: Metriche di accuratezza dei modelli - valori medi ottenuti tramite convalida incrociata

modello	**MSE**	**RMSE**	**MAE**
MARS	1.337	1.145	**0.934**
Bagging	4.476	2.104	1.772
Random Forest	4.085	2.011	1.738
SVR radiale	2.995	1.729	**1.407**
Gradient Boosting	3.721	1.918	**1.592**

5.3.1 Importanza delle variabili

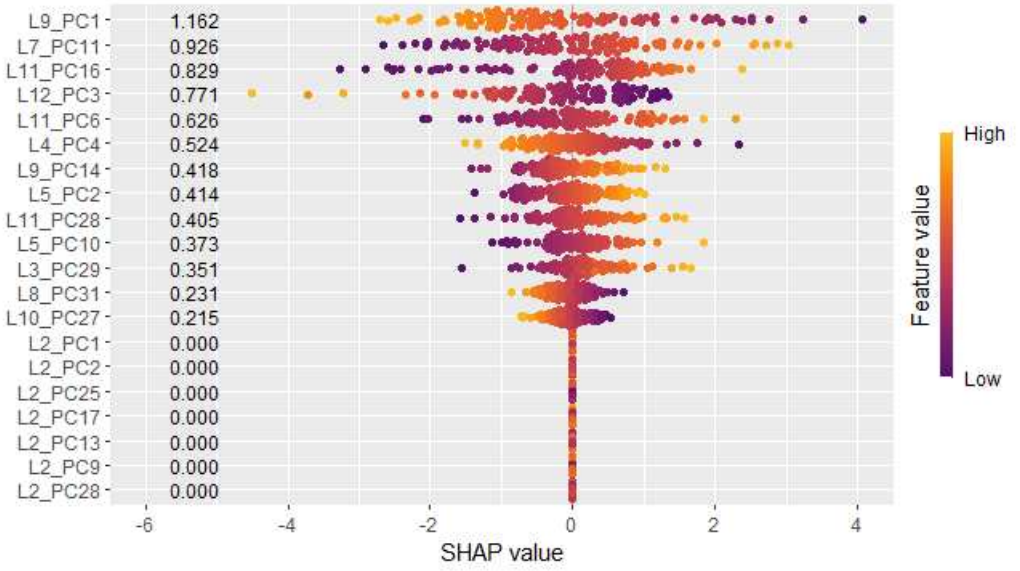
Al fine di comprendere meglio la struttura dei modelli di regressione qui adattati, si vogliono osservare quali sono le variabili esplicative più importanti rappresentate dalle componenti principali dei *word embeddings* dinamici. Analogamente a quanto fatto per i modelli del Capitolo 4, vengono calcolate delle misura di importanza delle covariate dei modelli migliori quali MARS, SVR radiale e Gradient Boosting. In *Figura 5.5* sono illustrati i valori SHAP delle 20 variabili più importanti del modello di regressione MARS. Il numero di variabili incluse nel modello risulta essere pari a 13 e la covariata più importante è *L9_PC*1 che corrisponde alla prima componente principale (*PC*) del nono layer (*L*). La composizione generale delle variabili incluse è formata da 7 componenti principali relative ai layers più profondi (9-12) del LLM, da 4 componenti di layers intermedi (5-8) e da sole 2 componenti dei primi strati superficiali (2-4). Non esiste un'interpretazione diretta dei layers mappati attraverso il modello linguistico BERT ma soltanto un'indicazione qualitativa: i primi layers del modello (1-4) catturano informazioni sintattiche e relazioni locali tra parole, quelli intermedi (5-8) estraggono relazione sematiche a lungo raggio e i layers più profondi (9-12) rappresentano la comprensione semantica e contestuale. Nell'ordine di importanza espresso dai valori SHAP del modello MARS, le esplicative più importanti sono rappresentate dalle componenti relative ai layers più profondi. Sembra quindi che siano proprio gli strati più profondi del modello linguistico di grandi dimensioni ad avere una maggiore importanza nella previsione dell'indice di democrazia, coerentemente a quanto evidenziato dallo studio di Ethayarajh e Kawin (2019) [37].

I valori SHAP calcolati per il modello Support Vector Regression con kernel radiale sono riportati in *Figura A.13* dell'appendice. Anche in questo caso sono illustrati soltanto i valori delle 20 variabili incluse più importanti e ritroviamo nelle prime 11 posizioni la prima componente principali di ciascun layer. Gli strati più profondi (9-12) si trovano anche in questo caso tra le esplicative più importanti ma si osserva che il grado di importanza per tutte e 20 le esplicative risulta essere piuttosto basso[2].

Infine, si vuole confrontare l'importanza delle variabili anche per il modello Gradient Boosting. In *Figura 5.6* è illustrato il grado di importanza delle prime 20 variabili espresso in termini di guadagno medio, analogamente a quanto fatto nel paragrafo 4.6. Le tre variabili più importanti sono in ordine: la prima componente principale del 12-esimo layer (*L12_PC1*), la terza componente principale

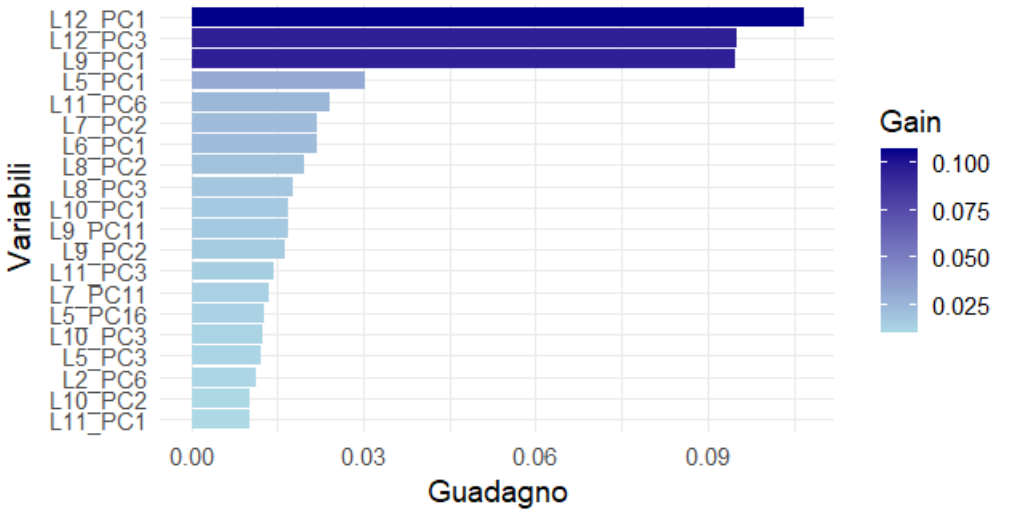
[2]la prima esplicativa *L9_PC1* corrisponde a un valore pari a 0.0901 e la 20-esima *L7_PC2* a un valore di 0.0626

Figura 5.5: Valori SHAP modello MARS - esplicative derivanti dal modello BERT



sempre del 12-esimo layer (*L12_PC3*) e la prima componente principale del nono layer (*L9_PC1*). Anche in questo caso emerge che i layers più profondi hanno una maggiore importanza nella previsione dell'indice di democrazia. La comprensione semantica del testo costituzionale risulta quindi essere un buon predittore per la stima della forma di governo vigente all'interno di un paese.

Figura 5.6: Importanza delle prime 20 variabili del modello Gradient Boosting - esplicative derivanti dal modello BERT



Conclusioni e limiti dello studio

Dalle analisi preliminari emerge che la composizione dei testi costituzionali è influenzata da fattori geografici e culturali, derivanti dalla storia di ciascun paese. Stati geograficamente vicini mostrano similitudini nelle coordinate delle prime due componenti principali. Un altro risultato significativo riguarda l'impatto del colonialismo, che è ancora presente a livello costituzionale per i paesi relativi alle ex colonie inglesi e francesi, come evidenziato dalla cluster analysis. Attraverso il topic modeling è stato possibile mappare il contenuto del corpus di testi identificando i principali temi ricorrenti. Tra questi, emergono la suddivisione dei poteri e la sua gestione, la descrizione del capo di Stato, diritti e doveri dei cittadini e il sistema elettorale. Ne consegue che il documento costituzionale racchiude gli elementi chiave per descrivere la forma di governo vigente all'interno del rispettivo paese.

Successivamente, il contenuto dei testi costituzionali è stato utilizzato per stimare il grado di democrazia reale di ciascuno Stato, espresso tramite l'indice proposto dall'Economist Intelligence Unit su una scala da 0 a 10. Attraverso diversi modelli di regressione è stato possibile prevedere l'indice di democrazia in base alla frequenza dei vocaboli più ricorrenti: il modello più accurato ha un errore medio assoluto pari a 1.28. La distribuzione della frequenza delle parole principali di ciascun testo costituzionale fornisce quindi un'indicazione approssimativa del grado di democrazia. Con il calcolo dei valori SHAP sui modelli migliori, si evidenzia che le parole maggiormente associate a un alto punteggio democratico sono *provid*, *vote*, *deputi*, *chief* e *right*. Uno dei limiti di questo studio è rappresentato dalla dimensione relativamente ridotta del campione pari a 166 costituzioni, sebbene queste coprano la quasi totalità della popolazione mondiale. Sono infatti escluse dal campione 29 microstati per i quali non è disponibile l'indice di democrazia ma che sarebbe interessante poter includere per sviluppi futuri.

L'approccio *bag of words* adottato per la stima dei modelli di regressione impone forti semplificazioni, riducendo fortemente la complessità delle informazioni presenti in ciascuna costituzione. Per migliorare la rappresentazione del corpus, è stato quindi utilizzato un modello linguistico di grandi dimensioni, col fine di estrar-

re il significato profondo dei testi costituzionali. In questo studio è stato condotto un esperimento preliminare considerando soltanto una porzione ridotta di ciascun testo per limitazioni computazionali. Questa forte semplificazione non consente la generalizzazione di questi ultimi risultati poiché è stata applicata un'ulteriore riduzione della numerosità campionaria, necessaria per escludere i testi costituzionali più lunghi. A partire dai *words embeddings* dinamici generati con il modello BERT, sono estratte le prime componenti principali e utilizzate come nuove variabili esplicative. La capacità predittiva dei modelli di regressione migliora, suggerendo che un LLM può rappresentare in modo più esaustivo il contenuto semantico dei testi. Questo risultato indica una possibile relazione tra quanto dichiarato a livello costituzionale e l'effettiva manifestazione del potere all'interno uno Stato. Un ulteriore risultato emerso riguarda l'importanza delle variabili di quest'ultimi modelli di regressione: le componenti principali relative ai layers più profondi del modello BERT risultano essere particolarmente rilevanti nella stima dell'indice di democrazia. Poiché i layers più profondi di un LLM sono associati alla comprensione semantica e contestuale di un documento, questo indica che è proprio il significato estratto dal contenuto di ciascun testo ad avere una maggiore rilevanza. Si lascia a sviluppi futuri una possibile generalizzazione dei risultati, andando ad estendere l'applicazione del modello linguistico all'intero testo costituzionale, includendo anche le costituzioni più lunghe.

Bibliografia

[1] Amnesty International. The State of World's Human Rights, April 2024. https://www.amnesty.org/en/documents/pol10/7200/2024/en/.

[2] V-Dem Institute. Democracy Report 2024, March 2024. https://www.v-dem.net/publications/democracy-reports/.

[3] Economist Intelligence Unit. Democracy Index 2023 - Age of conflict. *The Econimist Intelligence Unit Limited 2024*.

[4] Elliot W. Bulmer. What Is a Constitution? Principles and Concepts. *International Institute for Democracy and Electoral Assistance*, August 2014.

[5] James Melton Elkins Zachary, Tom Ginsburg. Constitute: The World's Constitutions to Read, Search, and Compare. https://www.constituteproject.org/constitutions?lang=en.

[6] World Population Review. Unwritten constitution countries 2024, 2024. https://worldpopulationreview.com/country-rankings/unwritten-constitution-countries.

[7] Sharon Pia Hickey. In the World of Constitution-Building in 2023. *ConstitutionNet, International IDEA*, 2024. https://constitutionnet.org/news/voices/world-constitution-building-2023.

[8] Centro Regionale di Informazione delle Nazioni Unite. Gli Stati Membri delle Nazioni Unite, 2021. https://unric.org/it/gli-stati-membri-delle-nazioni-unite/.

[9] R Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, 2024. https://www.R-project.org/.

[10] Maciej Eder, Jan Rybicki, and Mike Kestemont. Stylometry with R: A Package for Computational Text Analysis. *The R Journal*, 8(1):107, 2016.

[11] S. Argamon. Interpreting Burrows's Delta: Geometric and Probabilistic Foundations. *Literary and Linguistic Computing*, 23:131–147, October 2007.

[12] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, November 1987.

[13] Enciclopedia Treccani online. Commonwealth of Nations Commonwealth delle Nazioni, 2013. `https://www.treccani.it/enciclopedia/commonwealth-of-nations-commonwealth-delle-nazioni_(Atlante-Geopolitico)/`.

[14] Wikipedia. Impero coloniale francese, September 2024. `http://it.wikipedia.org/w/index.php?title=Impero_coloniale_francese&oldid=141122012`.

[15] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[16] Margaret E. Roberts, Brandon M. Stewart, and Edoardo M. Airoldi. A Model of Text for Experimentation in the Social Sciences. *Journal of the American Statistical Association*, 111(515):988–1003, July 2016.

[17] Xiwen Bai, Xiunian Zhang, Kevin X. Li, Yaoming Zhou, and Kum Fai Yuen. Research topics and trends in the maritime transport: A structural topic model. *Transport Policy*, 102:11–24, 2021.

[18] Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley. stm : An R Package for Structural Topic Models. *Journal of Statistical Software*, 91(2), 2019.

[19] A. Sciandra, M. Trevisani, and A. Tuzzi. Diagnostics for topic modelling. the dubious joys of making quantitative decisions in a qualitative environment. *P. Cerchiello, A. Agosto, S. Osmetti, A. Spelta (eds), Proceedings of the Statistics and Data Science Conference, Pavia University Press*, 2023.

[20] Adelchi Azzalini and Bruno Scarpa. *Analisi dei dati e data mining*. Unitext. Springer, Milano Berlin Heidelberg New York Hong Kong London Paris Tokyo, 2004.

[21] Jerome H. Friedman. Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1), March 1991.

[22] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification And Regression Trees*. Routledge, October 2017.

[23] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[24] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, September 1995.

[25] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[26] Scott Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions, November 2017. arXiv:1705.07874 [cs, stat].

[27] Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, 298:103502, September 2021.

[28] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space, September 2013. arXiv:1301.3781 [cs].

[29] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[31] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. arXiv:1810.04805 [cs].

[32] Jay Alammar. The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning), 2018. `https://jalammar.github.io/illustrated-bert/`.

[33] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. October 2016. arXiv:1609.08144 [cs].

[34] Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. June 2015. arXiv:1506.06724 [cs].

[35] kikaben.com. Transformer's Encoder-Decoder. Let's Understand The Model Architecture. `https://kikaben.com/transformers-encoder-decoder/`.

[36] Oscar Kjell, Salvatore Giorgi, and Schwartz H. Andrew. The Text-Package: An R-Package for Analyzing and Visualizing Human Language Using Natural Language Processing and Transformers. *Psychological Methods*, 28(6):1478–1498, 2023.

[37] Kawin Ethayarajh. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. September 2019. arXiv:1909.00512 [cs].

Grafici aggiuntivi

Figura A.1: Distribuzione della varianza spiegata delle componenti principali - utilizzando le 3000 parole più frequenti

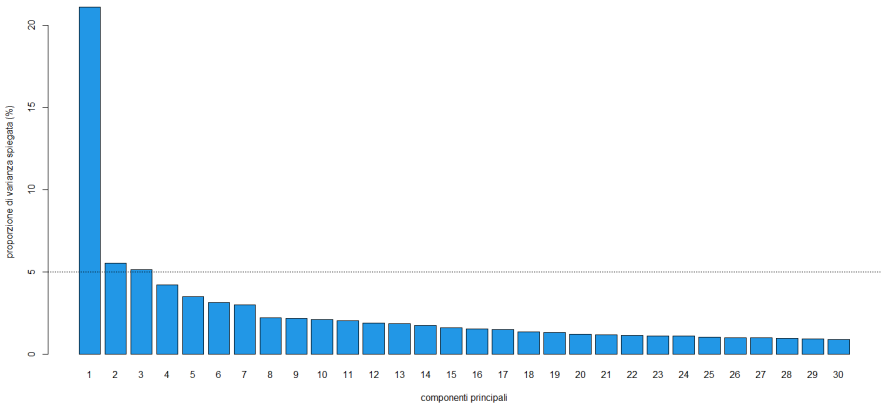


Figura A.2: Zoom1 Figura 2.2

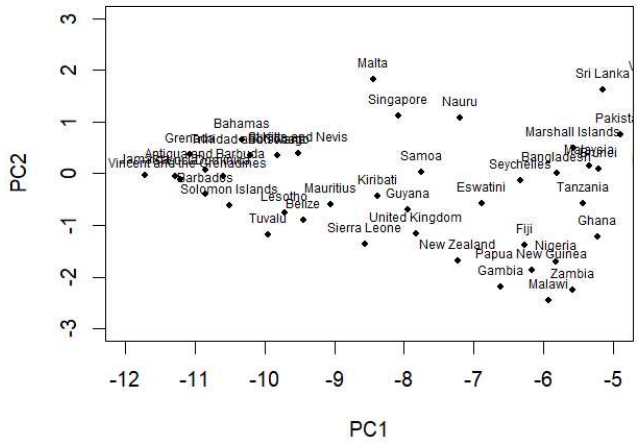


Figura A.3: Zoom2 Figura 2.2

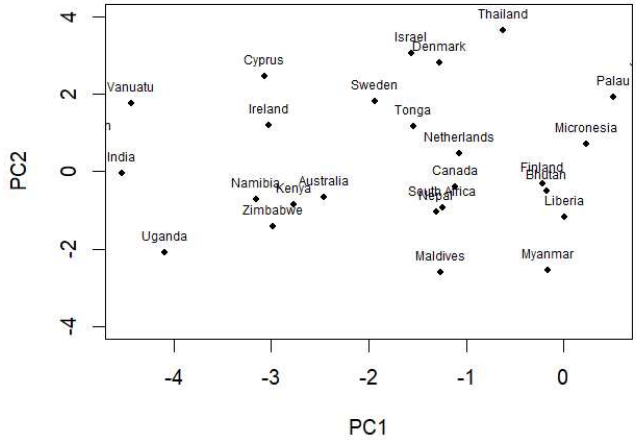


Figura A.4: Zoom3 Figura 2.2

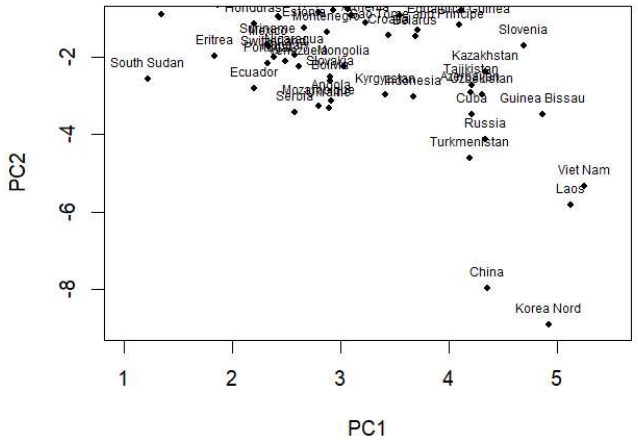


Figura A.5: Zoom4 Figura 2.2

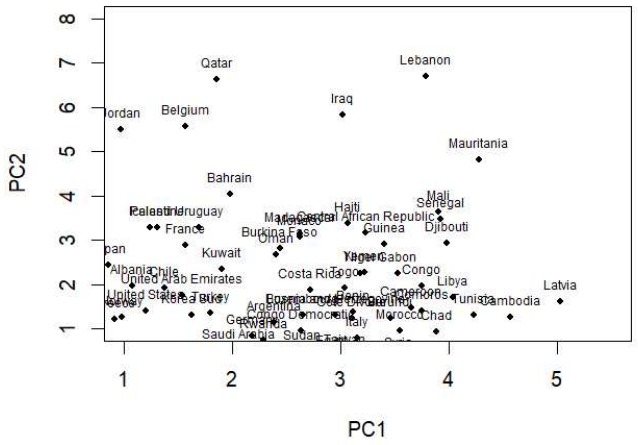


Figura A.6: Zoom5 Figura 2.2

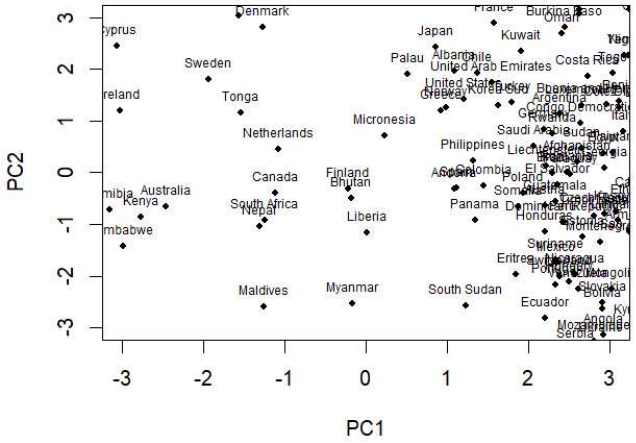


Figura A.7: Indice medio di silhouette per numero di cluster - distanza coseno

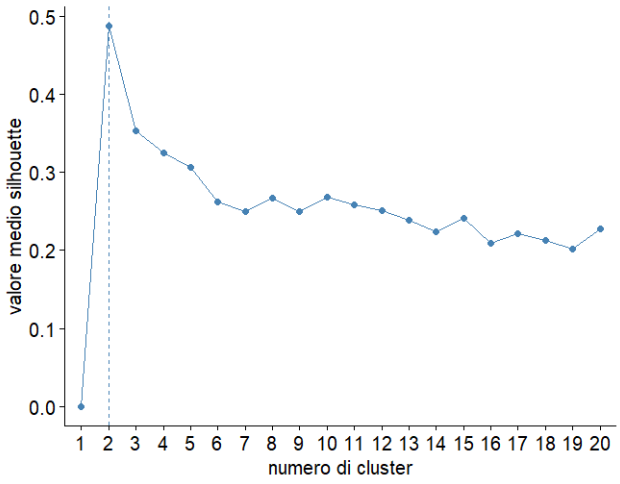


Figura A.8: Indice medio di silhouette per numero di cluster - distanza argamon

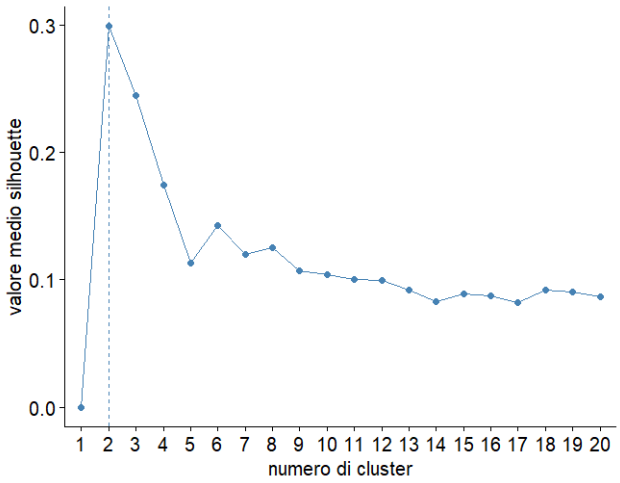


Tabella A.1: Cluster analysis - suddivisione due gruppi

CLUSTER 1:
Afghanistan, Albania, Algeria, Andorra, Angola, Argentina, Armenia, Australia, Austria, Azerbaijan, Bahrain, Belarus, Belgium, Benin, Bhutan, Bolivia, Bosnia and Herzegovina, Brazil, Bulgaria, Burkina Faso, Burundi, Cambodia, Cameroon, Canada, Cape Verde, Central African Republic, Chad, Chile, China, Colombia, Comoros, Congo, Congo Democratic, Costa Rica, Cote D'Ivoire, Croatia, Cuba, Czech Republic, Denmark, Djibouti, Dominican Republic, Ecuador, Egypt, El Salvador, Equatorial Guinea, Eritrea, Estonia, Ethiopia, Finland, France, Gabon, Georgia, Germany, Greece, Guatemala, Guinea, Guinea Bissau, Haiti, Honduras, Hungary, Iceland, Indonesia, Iran, Iraq, Italy, Japan, Jordan, Kazakhstan, Korea Nord, Korea Sud, Kosovo, Kuwait, Kyrgyzstan, Laos, Latvia, Lebanon, Liberia, Libya, Liechtenstein, Lithuania, Luxembourg, Macedonia, Madagascar, Mali, Mauritania, Mexico, Micronesia, Moldova, Monaco, Mongolia, Montenegro, Morocco, Mozambique, Myanmar, Nepal, Netherlands, Nicaragua, Niger, Norway, Oman, Palau, Palestine, Panama, Paraguay, Peru, Philippines, Poland, Portugal, Qatar, Romania, Russia, Rwanda, Sao Tome and Principe, Saudi Arabia, Senegal, Serbia, Slovakia, Slovenia, Somalia, South Sudan, Spain, Sudan, Suriname, Sweden, Switzerland, Syria, Taiwan, Tajikistan, Thailand, Timor Leste, Togo, Tunisia, Turkey, Turkmenistan, Ukraine, United Arab Emirates, United States, Uruguay, Uzbekistan, Venezuela, Viet Nam, Yemen
CLUSTER 2:
Antigua and Barbuda, Bahamas, Bangladesh, Barbados, Belize, Botswana, Brunei, Cyprus, Dominica, Eswatini, Fiji, Gambia, Ghana, Grenada, Guyana, India, Ireland, Israel, Jamaica, Kenya, Kiribati, Lesotho, Malawi, Malaysia, Maldives, Malta, Marshall Islands, Mauritius, Namibia, Nauru, New Zealand, Nigeria, Pakistan, Papua New Guinea, Samoa, Seychelles, Sierra Leone, Singapore, Solomon Islands, South Africa, Sri Lanka, St Kitts and Nevis, St Lucia, St Vincent and the Grenadines, Tanzania, Tonga, Trinidad and Tobago, Tuvalu, Uganda, United Kingdom, Vanuatu, Zambia, Zimbabwe

Tabella A.2: Cluster analysis - suddivisione tre gruppi

CLUSTER A:
Afghanistan, Albania, Algeria, Angola, Argentina, Armenia, Australia, Austria, Azerbaijan, Bahrain, Belarus, Belgium, Bhutan, Bolivia, Bosnia and Herzegovina, Brazil, Bulgaria, Cambodia, Canada, Cape Verde, Chile, China, Colombia, Costa Rica, Croatia, Cuba, Czech Republic, Denmark, Dominican Republic, Ecuador, Egypt, El Salvador, Eritrea, Estonia, Ethiopia, Finland, Georgia, Germany, Greece, Guatemala, Guinea Bissau, Haiti, Honduras, Hungary, Iceland, Indonesia, Iran, Iraq, Italy, Japan, Jordan, Kazakhstan, Korea Nord, Korea Sud, Kosovo, Kuwait, Kyrgyzstan, Laos, Latvia, Lebanon, Liberia, Libya, Liechtenstein, Lithuania, Luxembourg, Macedonia, Mexico, Micronesia, Moldova, Monaco, Mongolia, Montenegro, Mozambique, Myanmar, Nepal, Netherlands, Nicaragua, Norway, Oman, Palau, Palestine, Panama, Paraguay, Peru, Philippines, Poland, Portugal, Qatar, Romania, Russia, Rwanda, Sao Tome and Principe, Saudi Arabia, Serbia, Slovakia, Slovenia, Somalia, South Sudan, Spain, Sudan, Suriname, Sweden, Switzerland, Syria, Taiwan, Tajikistan, Thailand, Timor Leste, Tunisia, Turkey, Turkmenistan, Ukraine, United Arab Emirates, United States, Uruguay, Uzbekistan, Venezuela, Viet Nam, Yemen
CLUSTER B:
Andorra, Benin, Burkina Faso, Burundi, Cameroon, Central African Republic, Chad, Comoros, Congo, Congo Democratic, Cote DìIvoire, Djibouti, Equatorial Guinea, France, Gabon, Guinea, Madagascar, Mali, Mauritania, Morocco, Niger, Senegal, Togo
CLUSTER C:
Antigua and Barbuda, Bahamas, Bangladesh, Barbados, Belize, Botswana, Brunei, Cyprus, Dominica, Eswatini, Fiji, Gambia, Ghana, Grenada, Guyana, India, Ireland, Israel, Jamaica, Kenya, Kiribati, Lesotho, Malawi, Malaysia, Maldives, Malta, Marshall Islands, Mauritius, Namibia, Nauru, New Zealand, Nigeria, Pakistan, Papua New Guinea, Samoa, Seychelles, Sierra Leone, Singapore, Solomon Islands, South Africa, Sri Lanka, St Kitts and Nevis, St Lucia, St Vincent and the Grenadines, Tanzania, Tonga, Trinidad and Tobago, Tuvalu, Uganda, United Kingdom, Vanuatu, Zambia, Zimbabwe

Figura A.9: Analisi delle corrispondenze - zoom 1 della *Figura 4.2*

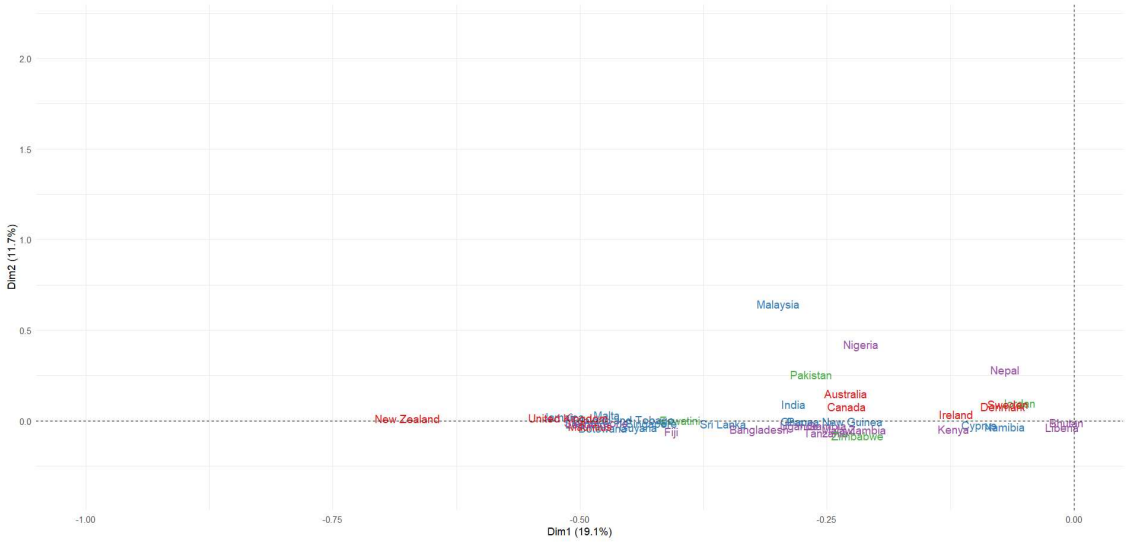


Figura A.10: Analisi delle corrispondenze - zoom 2 della *Figura 4.2*

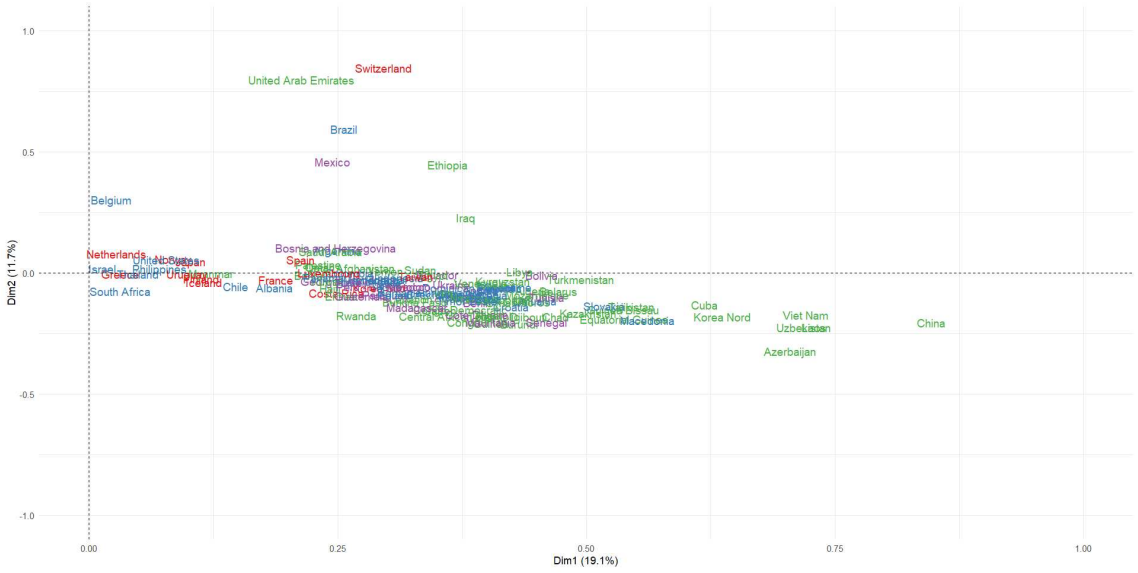


Figura A.11: Varianza cumulata per le prime 50 dimensioni dell'analisi delle corrispondenze

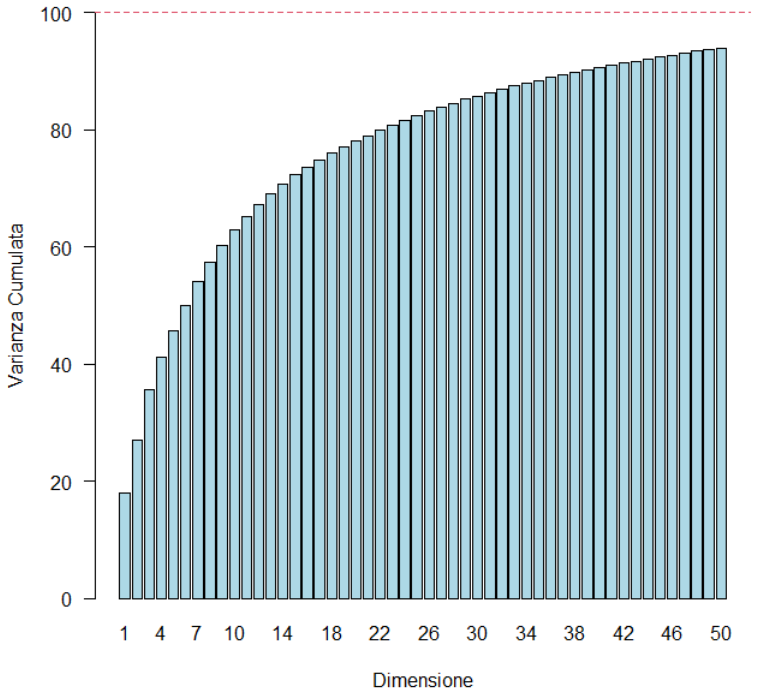


Figura A.12: Importanza delle prime 20 variabili - modello Random Forest del Capitolo 4

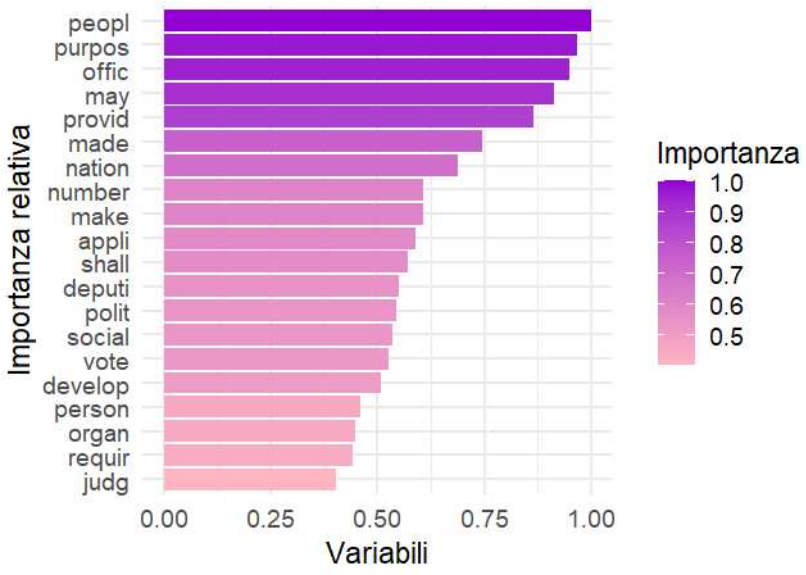


Figura A.13: Valori SHAP modello SVR con kernel radiale - esplicative derivanti dal modello BERT