Università di Pisa

Scuola di Ingegneria

Dipartimento di Ingegneria dell'Informazione

Laurea triennale in Ingegneria Informatica

Sentiment analysis e Reddit scores a confronto: trends nel contesto delle presidenziali USA 2020

Relatori:

prof. Marco Avvenuti

ing. Lorenzo Cima

Candidato:

Gabriele Frassi

Anno accademico 2023/24

Indice

Capitolo 1

Introduzione

1.1 Reddit e dibattito politico

Reddit è uno dei principali *social network*. Al 1 marzo 2024 risulta essere il 15esimo sito più visitato del mondo e il nono sito più visitato degli Stati Uniti d'America[1].

Figura 1.1: Logo di Reddit

La sua struttura richiama quella dei forum tradizionali:

- la piattaforma è caratterizzata da *subreddit*, ciascuno dedicato a specifici argomenti;
- l'utente decide a quali subreddit partecipare;
- l'utente può pubblicare *submission* e stimolare la discussione di nuovi argomenti all'interno di particolari subreddit;
- l'utente può pubblicare *comments* all'interno di particolari submission.

La differenza rispetto ai forum tradizionali sta nel passaggio da un approccio *top-down* a un approccio *bottom-up*: nei forum tradizionali sezioni e regole di moderazione sono determinate dagli amministratori della piattaforma, mentre su Reddit ciò è determinato dagli utenti iscritti. Tali caratteristiche hanno permesso a Reddit di diventare una delle piattaforme più apprezzate, ma allo stesso tempo hanno favorito l'emergere di nuove sfide e controversie in un contesto politico sempre più polarizzato. Citiamo, a tal proposito, due importanti paper che hanno analizzato le conseguenze del cosiddetto *The Great Ban*, un'operazione di *deplatforming* attuata dagli amministratori della piattaforma a Giugno 2020 per rimuovere contenuti politicamente estremi (operazione che è stata preceduta da interventi di moderazione di entità minore):

- *Make reddit great again: assessing community effects of moderation interventions on* `r/the_donald` [4].

  Studio che attesta come gli interventi di moderazione attuati sul subreddit `r/the_donald` tra Giugno 2019 e Giugno 2020 hanno sì permesso una riduzione delle attività di utenti controversi e limitato la diffusione di contenuti politicamente estremi, ma allo stesso tempo hanno comportato un aumento di tossicità e polarizzazione.

[1] Dati a cura di *similarweb*: `https://www.similarweb.com/top-websites/`

- *The Great Ban: Efficacy and Unintended Consequences of a Massive Deplatforming Operation on Reddit*[1].

  Studio che affronta gli effetti del *The Great Ban* in maniera trasversale, considerando i quindici subreddit più popolari tra quelli coinvolti nell'operazione di *deplatforming*. Si sottolinea, come nel paper precedente, che queste operazioni hanno comportato l'emergere di *resentful users*.

Tali passaggi sono avvenuti nei mesi antecedenti le elezioni presidenziali statunitensi del 2020 (3 Novembre 2020), che hanno visto Joe Biden (Democratico) sconfiggere il presidente uscente Donald Trump (Repubblicano). Gli eventi successivi al 3 novembre, in primis il rifiuto del presidente uscente di riconoscere l'esito del voto, hanno influito pesantemente sul dibattito social.

1.2 *up/down score*

Reddit offre ai propri utenti la possibilità di esprimere apprezzamento o disappunto verso *submissions* e *comments* attraverso un *up/down score*:

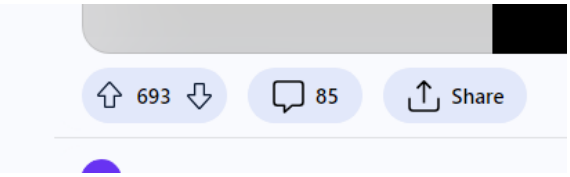

Figura 1.2: Bottoni up and down

- ogni *submission* e *comment* è caratterizzata da uno *score* un valore numerico che può avere segno positivo o negativo;
- l'utente ha a disposizione due tasti, uno per esprimere gradimento (quindi incrementare lo score) e uno per esprimere disappunto (quindi decrementare lo score);
- maggiore è lo score maggiore è l'apprezzamento degli utenti verso quel particolare commento, minore è lo score minore è l'apprezzamento degli utenti.

1.3 Obiettivo dello studio

Obiettivo del presente studio è analizzare la correlazione tra *up/down scores* di un particolare dataset di *comments* (pubblicati durante le elezioni presidenziali statunitensi del 2020) e l'esito della *sentiment analysis* sugli stessi commenti. Vogliamo verificare se esistono particolari trend nel contesto di forte polarizzazione politica precedentemente descritto.

- ***subreddit* di nostro interesse**.

  Si è deciso di coinvolgere nello studio due particolari subreddit, uno relativo all'area politica democratica e uno relativo all'area politica repubblicana:

  - `r/Republican` (`https://www.reddit.com/r/Republican/`)
    - **Numero di membri** (al 01 marzo 2024): 191K
    - **Descrizione**: */r/Republican is a partisan subreddit. This is a place for Republicans to discuss issues with other Republicans.*

  - `r/democrats` (`https://www.reddit.com/r/democrats/`)
    * **Numero di membri** (al 01 marzo 2024): 437K
    * **Descrizione**: *The Democratic Party is building a better future for everyone and you can help. Join us today and help elect more Democrats nationwide! This sub offers daily news updates, policy analysis, links, and opportunities to participate in the political process. We are here to get Democrats elected up and down the ballot.*

- **Periodo di nostro interesse**.

  Abbiamo considerato un range di giorni che permette di analizzare i comportamenti degli utenti sia nel periodo antecedente l'*election day*, sia nel periodo successivo. I giorni considerati vanno dal 07 ottobre 2020 al 1 dicembre 2020.

- ***VADER*, tool per la *sentiment analysis***.

  Per la sentiment analysis si è deciso di fare affidamento a VADER (*Valence Aware Dictionary and sEntiment Reasoner*[2]), tool pensato esplicitamente per la sentiment analysis sui social media. VADER ha acquisito grande consenso in letteratura, con studi che ne hanno confermato la validità nel contesto dei social media e in confronto ad altri tool[2].

[2] `https://github.com/cjhutto/vaderSentiment`

Capitolo 2

Costruzione del *dataset*

2.1 Dataset di partenza

Il primo passaggio è stato l'acquisizione dei commenti pubblicati su Reddit nel periodo di nostro interesse. I relativi dataset sono disponibili sulla piattaforma *academictorrents*[3], divisi per mesi e posti in formato compresso *zst*. I file scaricati sono i seguenti, con essi copriamo i commenti pubblicati su Reddit nei mesi di ottobre, novembre e dicembre 2020.

Nome	**Dimensione**
`comments/RC_2020-10.zst`	18.75GB
`comments/RC_2020-11.zst`	18.42GB
`comments/RC_2020-12.zst`	19.22GB

Abbiamo posto il focus sul periodo temporale intorno alle elezioni presidenziali USA del 2020, tuttavia dobbiamo ancora restringere il dataset ai soli commenti di nostro interesse.

2.2 *data filtering*

2.2.1 Limitazione ai *subreddit* di nostro interesse

Attuiamo il primo filtraggio con *PushshiftDumps*[1], tool appositamente pensato per il filtraggio dei dataset scaricati dalla repository di cui al passaggio precedente. Dato il dataset di partenza vogliamo ottenere esclusivamente i commenti relativi ai subreddit `r/democrats` e `r/Republican`, di cui abbiamo parlato nell'introduzione: il risultato è il seguente

Nome	**# utenti**	**# commenti**	**Dimensione**
`2020-10`	12707	59080	74.6MB
democrats	4883	23584	
Republican	8048	35496	
`2020-11`	15022	71795	90.6MB
democrats	6181	30787	
Republican	9131	41008	
`2020-12`	10257	51589	65.8MB
democrats	3535	17860	
Republican	6863	33729	

Le somme degli utenti coinvolti nei subreddit non è uguale al numero degli utenti presenti nel file in quanto esistono utenti che hanno pubblicato commenti in entrambi i subreddit:

- 223 utenti in `2020-10`

[1] `https://github.com/Watchful1/PushshiftDumps`

- 289 utenti in `2020-11`
- 140 utenti in `2020-12`

2.2.2 Costruzione di due finestre temporali

Vogliamo costruire due finestre temporali su cui lavorare, ciascuna di 28 giorni:

- la prima riguarda i giorni antecedenti le elezioni presidenziali e va dal 07 ottobre 2020 al 3 novembre 2020 (abbiamo deciso di includere in questa finestra i commenti pubblicati durante l'*election day*);
- la seconda riguarda i giorni successivi e va dal 4 novembre 2020 al 1 dicembre 2020.

I file precedenti (`2020-10`, `2020-11`, `2020-12`) sono stati aperti, filtrati e riorganizzati sfruttando le funzionalità della libreria Python *Pandas*: il risultato sono i file `first_window` e `second_window`, dove le finestre temporali sono distinte in maniera netta

Nome	**# utenti**	**# commenti**	**Dimensione**
`first_window`	12101	55751	74.6MB
democrats	4857	23163	
Republican	7467	32588	
`second_window`	13811	64823	90.6MB
democrats	5516	26840	
Republican	8536	37983	

Anche qua vale lo stesso discorso degli utenti che hanno pubblicato commenti in entrambi i subreddit:

- 222 utenti in `first_window`
- 240 utenti in `second_window`

2.2.3 Rimozione di commenti di utenti eliminati

All'interno del dataset sono mantenuti commenti di utenti eliminati da Reddit: questi risultano avere username (attributo `author`) *[deleted]*. I commenti relativi a questi utenti devono essere rimossi:

- non è possibile distinguere gli utenti eliminati utilizzando altri attributi;
- i commenti relativi ad utenti con username *[deleted]* risulteranno rimossi in ogni caso in virtù della condizione (3) con cui identifichiamo un utente come *non umano*.

Nella prima finestra sono stati rimossi 12650 commenti, nella seconda 14225.

Nome	**# Prima**	**# Dopo**	**Δ**
`first_window`	55751	43101	12650
democrats	23163	17540	5623
Republican	32588	25561	7027
`second_window`	64823	50598	14225
democrats	26840	20833	6007
Republican	37983	29765	8218

Ci chiediamo se l'eliminazione di oltre ventiseimila commenti sia impattante in senso negativo sulla qualità del dataset. Calcoliamo per ogni giorno il numero di commenti rimossi e mettiamo tale valore a confronto col numero di commenti pubblicati nel giorno stesso (si intende il numero di commenti prima della rimozione)

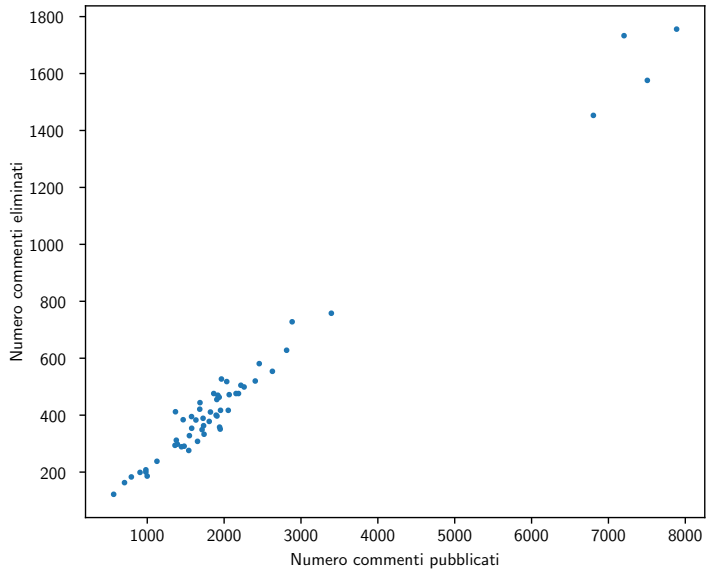

Figura 2.1: Correlazione tra num. di commenti pubblicati e num. commenti cancellati

Gli *outliers* (i quattro punti in alto a destra) rappresentano il picco di attività nei quattro giorni successivi all'*election day*.

Giorno	**# pubblicati**	**# cancellati**
04-11	7887	1756
05-11	7203	1733
06-11	6804	1453
07-11	7506	1576

Osserviamo dal grafico la correlazione lineare tra numero di commenti pubblicati e numero di commenti rimossi, il che è positivo (numero di commenti eliminati sostanzialmente proporzionale al numero di commenti pubblicati).

2.2.4 Rimozione di commenti di utenti "non umani"

Siamo interessati unicamente ai commenti di persone umane, pertanto i commenti riconducibili a *bot* devono essere rimossi. Ai fini del presente elaborato consideriamo un utente "non umano" se almeno una delle seguenti condizioni risulta essere soddisfatta:

1. l'utente ha un username (attributo `author`) che inizia con *auto* e/o finisce con *auto*
2. l'utente ha un username (attributo `author`) che inizia con *bot* e/o finisce con *bot*
3. l'utente ha pubblicato almeno due commenti con lo stesso timestamp (attributo `created_utc`, commenti pubblicati con lo stesso timestamp forniscono l'avvisaglia di comportamenti automatizzati)

Rispetto alle condizioni (1) e (2) si è deciso di verificare la presenza delle keyword solo all'inizio e alla fine della stringa per minimizzare i falsi positivi. Stesse ragioni dietro la decisione di limitarsi all'individuazione di commenti con lo stesso timestamp (condizione (3)) e di non valutare la rimozione degli utenti analizzando la vicinanza tra timestamp di commenti consecutivi (un utente attivo può essere estremamente veloce nella pubblicazione di commenti). Troviamo che

	First Window		Second Window	
Cond.	**# utenti**	**# commenti**	**# utenti**	**# commenti**
(1)	3	2534	6	1977
democrats	1	912	3	774
Republican	3	1622	4	1203
(2)	29	127	32	133
democrats	4	31	11	46
Republican	25	96	21	87
(3)	3	8	3	6
democrats	3	8	0	0
Republican	0	0	3	6

Unendo il tutto otteniamo le seguenti variazioni nel numero di commenti

	Prima		Dopo		Differenze	
Nome	**# utenti**	**# commenti**	**# utenti**	**# commenti**	ΔU	ΔC
`first_window`	12100	43101	12065	40282	35	2819
democrats	4856	17540	4848	16439	8	1101
Republican	7466	25561	7438	23843	28	1718
`second_window`	13810	50598	13769	48416	41	2182
democrats	5515	20833	5500	20011	15	822
Republican	8535	29765	8507	28405	28	1360

Anche qua si evidenzia la presenza di utenti che risultano aver pubblicato commenti sia in `r/democrats` che in `r/Republican`: 221 nella prima finestra, 238 nella seconda.

2.2.5 Rimozione di attributi non necessari ai fini della nostra analisi

Riduciamo la dimensione del dataset eliminando gli attributi che non sono di nostro interesse. Si è deciso di mantenere i seguenti attributi:

- `id`

  Identificativo alfanumerico del commento.

- `author_fullname`

  Identificativo alfanumerico dell'utente che ha pubblicato il commento.

- `body`

  Testo del commento.

- `subreddit`

  subreddit dove il commento è stato pubblicato. A seguito dei precedenti filtraggi l'attributo può assumere solo due valori: *Republican* o *democrats*.

- `parent_id`

  Identificativo alfanumerico dell'entità (*submission* o *comment*) commentata dall'utente.

- `link_id`

  Identificativo alfanumerico della *submission* relativa al commento.

- `created_utc`

  timestamp relativo al commento.

- `score`

  up/down score relativo al commento.

A seguito di questo passaggio il numero di colonne del dataset si è ridotto da 51 a 8.

2.3 Introduzione dell'attributo *vader compound*[2]

Concludiamo la costruzione del nostro dataset inserendo per ogni commento l'esito della *sentiment analysis* compiuta da VADER. Per ogni commento presente nel dataset andiamo ad eseguire le seguenti righe di codice

```
1 analyzer = SentimentIntensityAnalyzer()
2
3 for element in dataset.index :
4     vader_analysis = analyzer.polarity_scores(dataset['body'][element])
5     dataset.loc[element, 'vader_compound'] = vader_analysis['compound']
```

La funzione `polarity_scores` restituisce un dizionario costituito da quattro elementi: l'unico attributo di nostro interesse è `compound`, che riassume l'esito della *sentiment analysis* con un valore compreso tra -1 e $+1$.

Interpretazione dell'attributo L'esito dell'analisi è determinato dai seguenti *threshold*, quelli suggeriti nel README del tool.

- ***Positive sentiment***: score ≥ 0.05
- ***Neutral sentiment***: score $\in]-0.05, +0.05[$
- ***Negative sentiment***: score ≤ -0.05

Con l'introduzione di questo attributo il numero di colonne del dataset aumenta da 8 a 9.

2.4 Dataset finale

Il dataset ottenuto dai passaggi precedenti è scomposto in due *subset*: `first_window` e `second_window`, aventi le dimensioni indicate.

Nome	# utenti	# commenti	Dimensione
`first_window`	12065	40282	13.3MB
democrats	4848	16439	
Republican	7438	23843	
`second_window`	13769	48416	17MB
democrats	5500	20011	
Republican	8507	28405	

- Si definisce utente un account Reddit che ha pubblicato almeno un commento all'interno dei subreddit di nostro interesse. Il numero di commenti presenti all'interno di un particolare file rappresenta il numero di *results* che costituiscono il *subset*.

[2] `https://github.com/cjhutto/vaderSentiment`

- In `first_window` si hanno 221 utenti che hanno pubblicato commenti in entrambi i subreddit di nostro interesse (`r/democrats` e `r/Republican`). In `second_window` si hanno 238 utenti con la medesima caratteristica. Osserviamo un leggero aumento di questo numero a seguito dell'*election day*.
- Il numero di commenti e utenti è maggiore nella `second_window`: +1704 utenti e +8134 commenti. Assistiamo ad un aumento anche rispetto ai singoli subreddit:
  - `r/democrats` con +652 utenti e +3572 messaggi
  - `r/Republican` con +1069 utenti e +4562 messaggi

Capitolo 3

Analisi statistica

3.1 Numero di commenti giornalieri all'interno delle finestre

Consideriamo le due finestre temporali e all'interno di ciascuna calcoliamo il numero di commenti giornalieri. Il grafico risultante è il seguente, dove in nero si registra l'andamento generale, in rosso quello del subreddit `r/Republican` e in blu quello di `r/democrats`

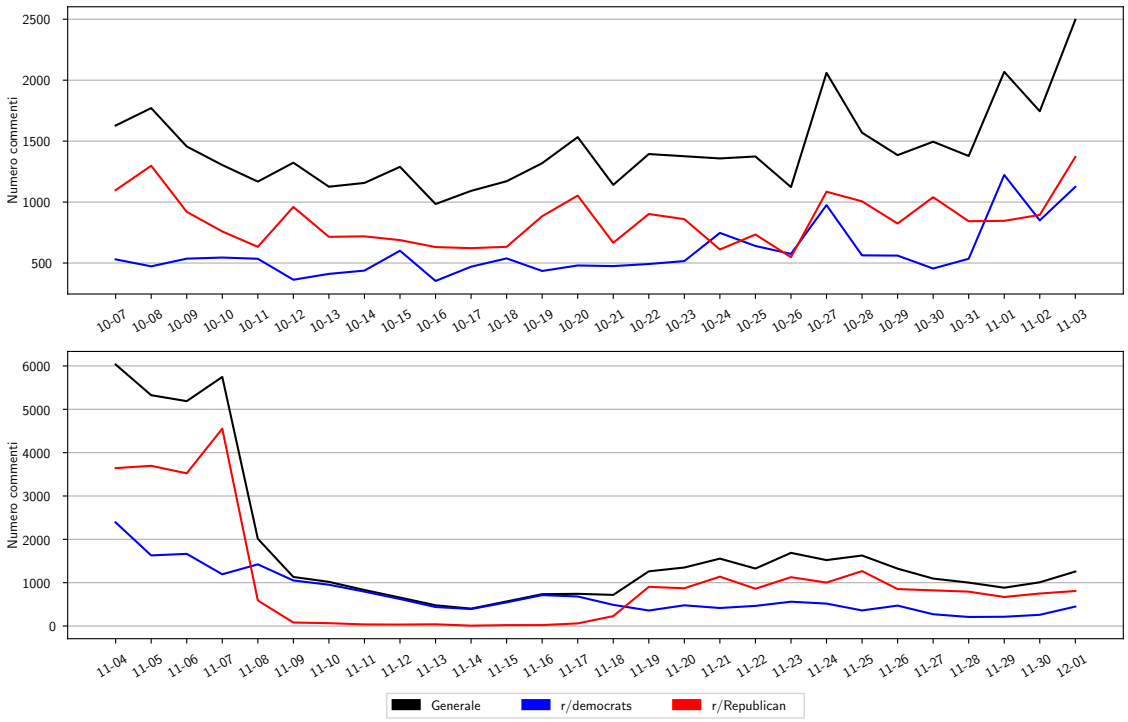

Figura 3.1: Numero di commenti giornalieri all'interno delle due finestre

Calcoliamo le medie del numero di commenti pubblicati.

- **Prima finestra**. La media globale è ≈ 1438.64. Ponendo il focus sui subreddit otteniamo ≈ 587.11 per `r/democrats` e ≈ 851.54 per `r/Republican`
- **Seconda finestra**. La media globale è ≈ 1731.71. Ponendo il focus sui subreddit otteniamo ≈ 714.75 per `r/democrats` e ≈ 1016.96 per `r/Republican`

In media `r/Republican` presenta un maggior numero di commenti giornalieri, anche se con un brusco calo del numero di commenti nel periodo dal 09 novembre al 17 novembre. Tale calo non è riconducibile alle operazioni di *data filtering* compiute per la costruzione

del dataset (in particolare non è riconducibile alla rimozione dei commenti degli utenti eliminati di cui alla sezione 2.2.4). Si assiste anche ad un aumento rilevante del numero di commenti durante l'*election day* e i quattro giorni successivi (cosa già mostrata con gli *outliers* della figura 2.1).

3.2 Frequenza dei valori dello *score* nelle finestre

Analizziamo la distribuzione dei valori assunti dallo *score* per mezzo di un istogramma. Poniamo anche uno "zoom" sulla parte bassa dell'istogramma, in modo tale da visionare gli intervalli con frequenze basse.

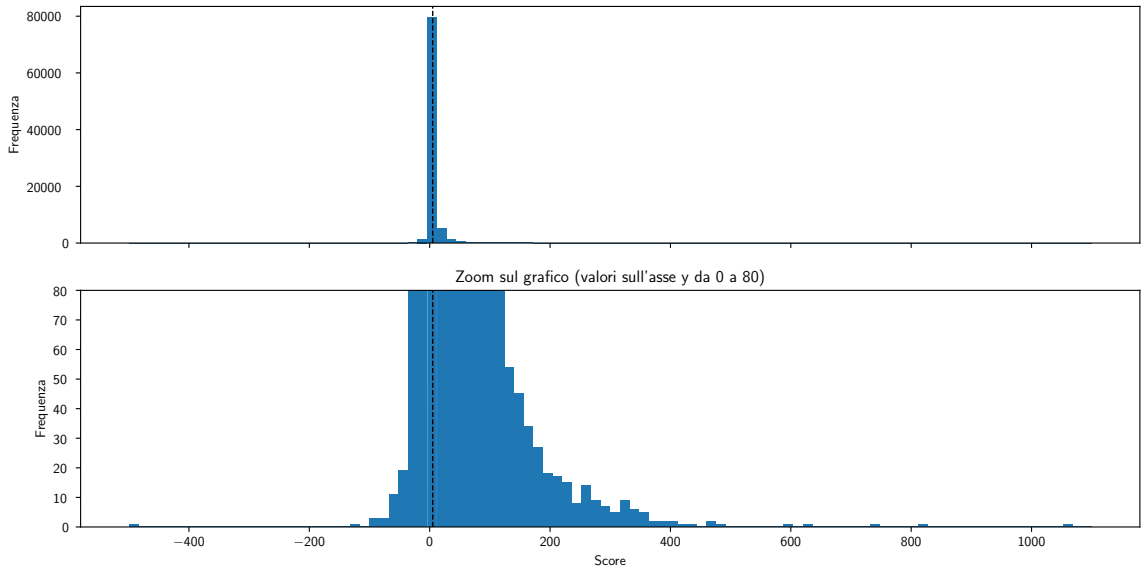

Figura 3.2: Frequenza dei valori dello *score* (globale). In tratteggio lo *score medio*

Osserviamo i seguenti valori:

- il valore minimo assunto dallo *score* è -498
- il valore massimo assunto dallo *score* è $+1060$
- il valor medio assunto dallo *score* è ≈ 5.04, positivo.

La seguente tabella indica gli intervalli di valori con una frequenza ≥ 100

Intervallo	#
$[-36, -20[$	116
$[-20, -4[$	1412
$[-4, +12[$	79469
$[+12, +28[$	5082
$[+28, +44[$	1169
$[+44, +60[$	564
$[+60, +76[$	273
$[+76, +92[$	180
$[\mathbf{-36}, \mathbf{+92}[$	**88265**

Emerge che 79469 commenti su 88698 (l'89.6%) ha uno score compreso tra -4 e $+12$! L'intervallo $[-36, +92[$ contiene gli *score* assunti dal 99.5% dei commenti!

3.2.1 Prima finestra temporale

Consideriamo la prima finestra temporale.

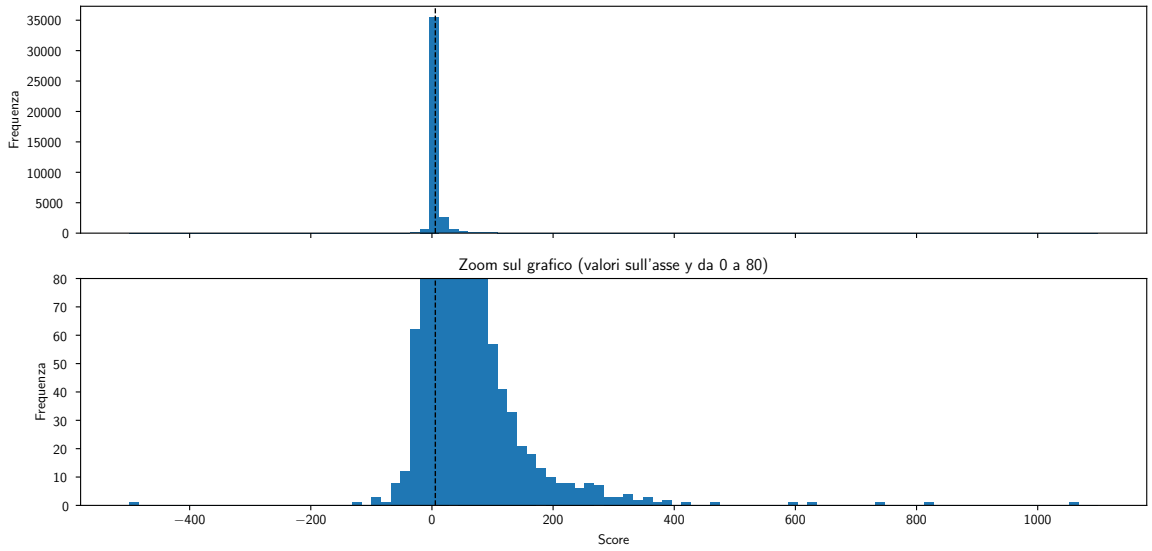

Figura 3.3: Frequenza dei valori dello *score* (prima finestra). In tratteggio lo *score medio*

Osserviamo i seguenti valori:

- il valore minimo assunto dallo *score* è -498 (stesso valore della finestra globale)
- il valore massimo assunto dallo *score* è $+1060$ (stesso valore della finestra globale)
- il valor medio assunto dallo *score* è ≈ 5.58, maggiore del valor medio globale.

La seguente tabella indica gli intervalli di valori che abbiamo già considerato nell'analisi globale

Intervallo	#
$[-36, -20[$	62
$[-20, -4[$	680
$[-4, +12[$	35516
$[+12, +28[$	2602
$[+28, +44[$	616
$[+44, +60[$	284
$[+60, +76[$	152
$[+76, +92[$	89
$[\mathbf{-36}, \mathbf{+92}[$	**40001**

35516 commenti su 40282 (l'88.17%, percentuale inferiore all'89.6% dell'analisi globale) ha uno score compreso tra -4 e $+12$! L'intervallo $[-36, +92[$ contiene gli *score* assunti dal 99.30% dei commenti (percentuale inferiore al 99.5% dell'analisi globale)!

3.2.2 Seconda finestra temporale

Consideriamo adesso la seconda finestra temporale.

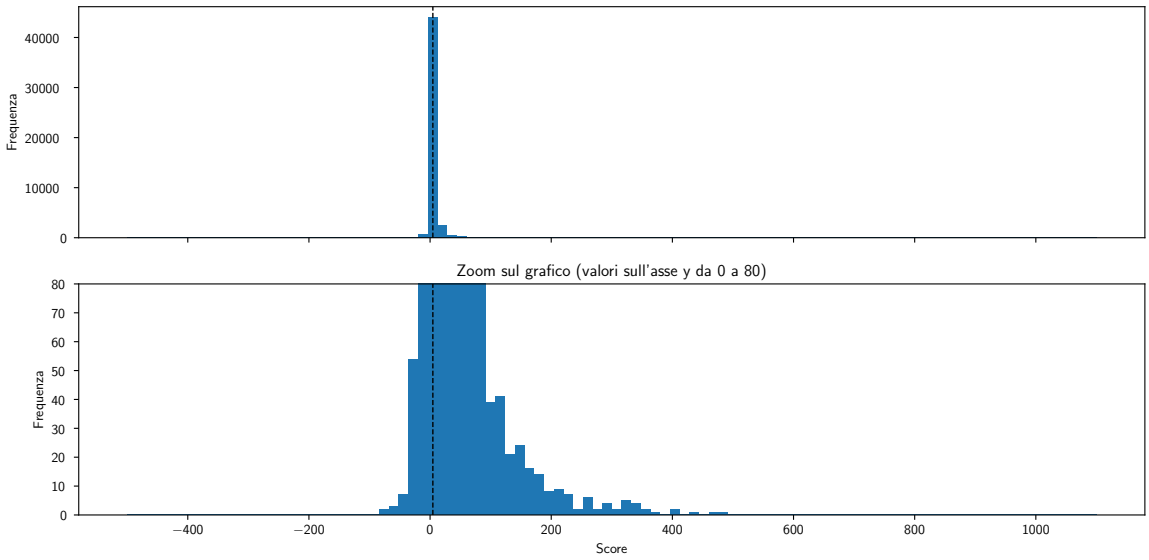

Figura 3.4: Frequenza dei valori dello *score* (seconda finestra).

Osserviamo i seguenti valori:

- il valore minimo assunto dallo *score* è -81 (valore $>>$ maggiore del minimo globale)
- il valore massimo assunto dallo *score* è $+488$ (valore $<<$ minore del massimo globale)
- il valor medio assunto dallo *score* è ≈ 4.59, minore del valor medio globale.

La seguente tabella indica gli intervalli di valori che abbiamo già considerato nell'analisi globale

Intervallo	**#**
$[-36, -20[$	54
$[-20, -4[$	732
$[-4, +12[$	43953
$[+12, +28[$	2480
$[+28, +44[$	553
$[+44, +60[$	280
$[+60, +76[$	121
$[+76, +92[$	91
$[\mathbf{-36}, \mathbf{+92}[$	**48264**

43953 commenti su 48416 (il 90.78%, percentuale superiore all'89.6% dell'analisi globale e all'88.17% della prima finestra) ha uno score compreso tra -4 e $+12$! L'intervallo $[-36, +92[$ contiene gli *score* assunti dal 99.69% dei commenti (percentuale superiore al 99.5% dell'analisi globale e al 99.3% della prima finestra)! Il range di valori assunti è decisamente più piccolo e, contrariamente alla prima finestra, non si hanno *outlier*.

3.3 Frequenza dei valori del *compound* nelle finestre

Analizziamo la distribuzione dei valori assunti dal *compound* per mezzo di un istogramma.

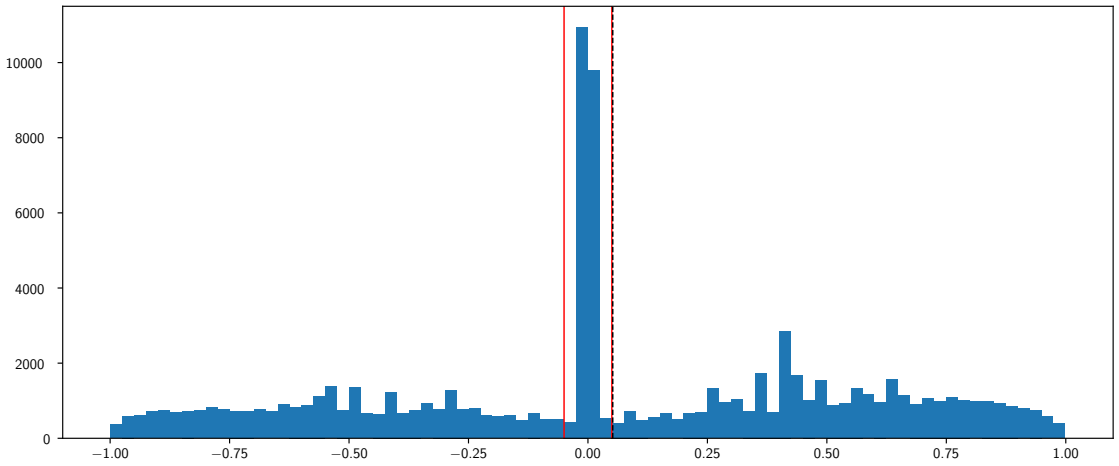

Figura 3.5: Frequenza dei valori del *compound* (globale).

Osserviamo che la maggioranza dei valori assunti si colloca indisputabilmente nell'intervallo $]-0.05, +0.05[$ (i valori che si classificano come *neutral sentiment*), di questi:

- oltre diecimila appartengono all'intervallo $[-0.025, 0[$ (rettangolo più alto)
- quasi diecimila appartengono all'intervallo $[0, +0.025[$ (secondo rettangolo più alto)

Il *compound medio* è pari a $\approx +0.052$, che si classifica come *positive sentiment* essendo leggermente superiore a $+0.05$. Procediamo analizzando la distribuzione all'interno delle singole finestre temporali.

3.3.1 Prima finestra temporale

Consideriamo la prima finestra temporale

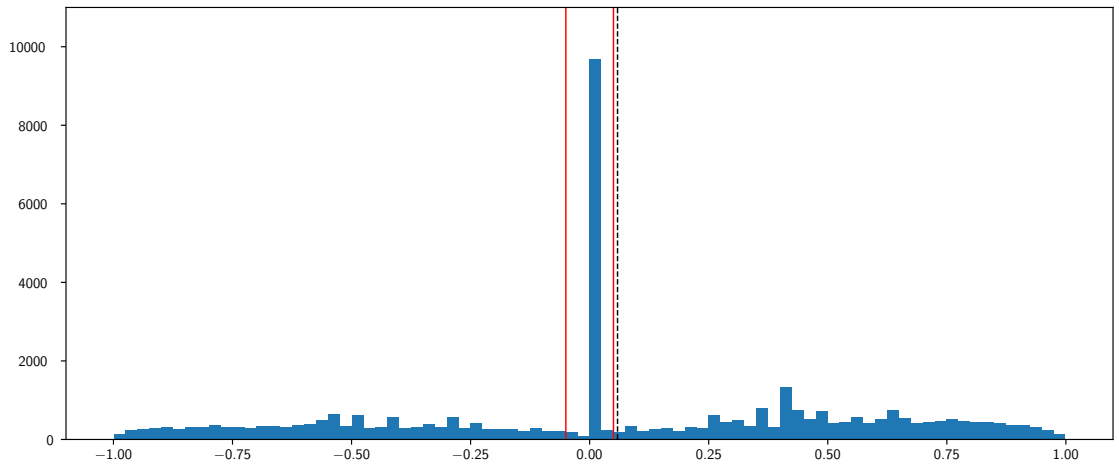

Figura 3.6: Frequenza dei valori del *compound* (prima finestra).

I valori prevalenti sono quelli appartenenti all'intervallo $[0; +0.025[$ (intervallo rappresentato dal rettangolo più alto). Il *compound medio* all'interno della finestra è pari a $\approx +0.059$ (*positive sentiment*): esso risulta essere maggiore del *compound medio* globale.

3.3.2 Seconda finestra temporale

Consideriamo adesso la seconda finestra temporale

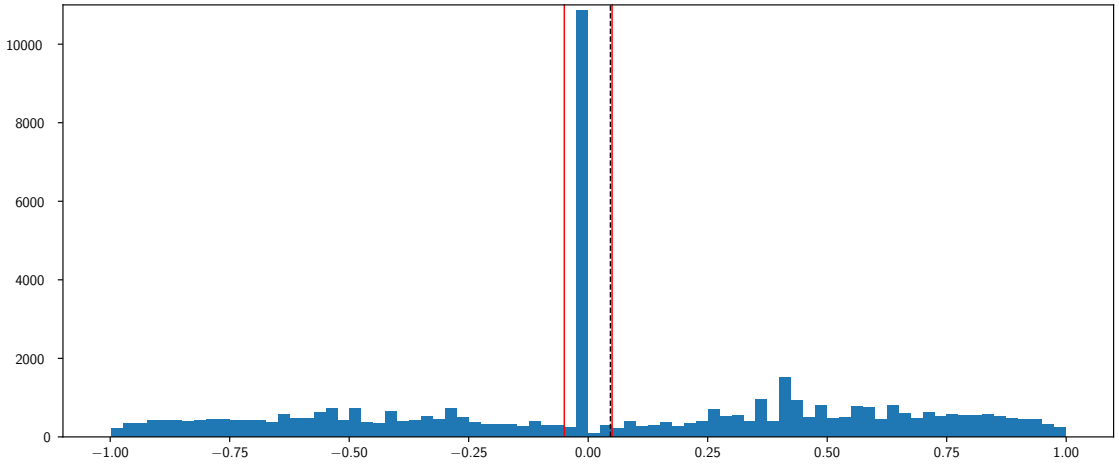

Figura 3.7: Frequenza dei valori del *compound* (seconda finestra).

I valori prevalenti sono quelli appartenenti all'intervallo $[-0.025; 0[$ (intervallo rappresentato dal rettangolo più alto). Il *compound medio* all'interno della finestra è pari a $\approx +0.046$: esso non è solo minore del *compound medio* globale, ma anche classificato diversamente (*neutral sentiment*).

3.4 Correlazione tra *score* e *compound* rispetto a commenti

Costruiamo degli scatterplot con cui cerchiamo di individuare correlazioni tra *score* e *compound* rispetto ai commenti pubblicati:

- lungo l'asse x poniamo lo *score*;
- lungo l'asse y poniamo il *compound*;
- ogni punto presente nello scatterplot rappresenta un commento.

Poniamo la scala logaritmica per entrambi gli assi, visto che lo *score* può assumere valori molto elevati (non si ha un minimo e un massimo, a differenza del *compound*).

3.4.1 Correlazione rispetto alle finestre temporali

Consideriamo prima e seconda finestra temporale: osserviamo l'assenza di correlazioni significative tra *score* e *compound*

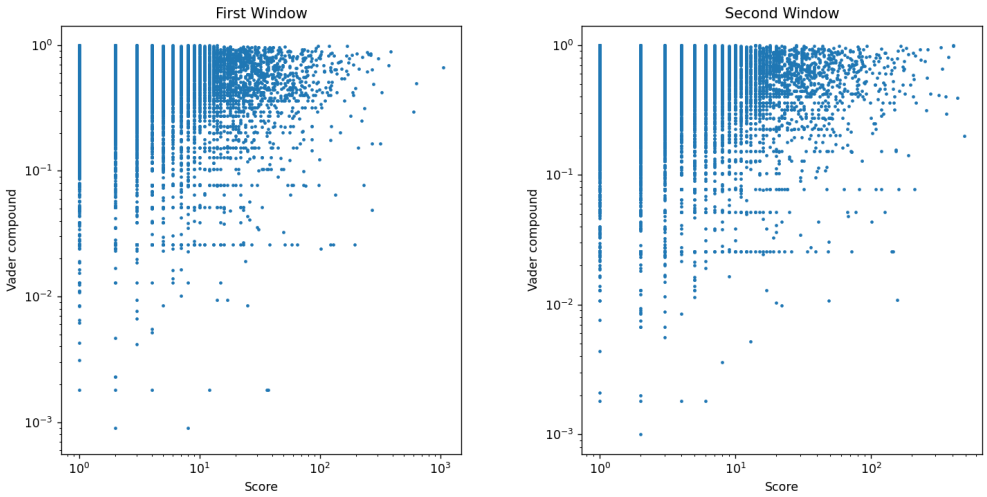

Figura 3.8: scatterplot con *score* e *compound* (Finestre temporali a confronto)

3.4.2 Correlazione rispetto ai subreddit considerati

Facciamo lo stesso confronto, ma rispetto ai singoli *subreddit*. Anche in questo confronto non emergono correlazioni significative.

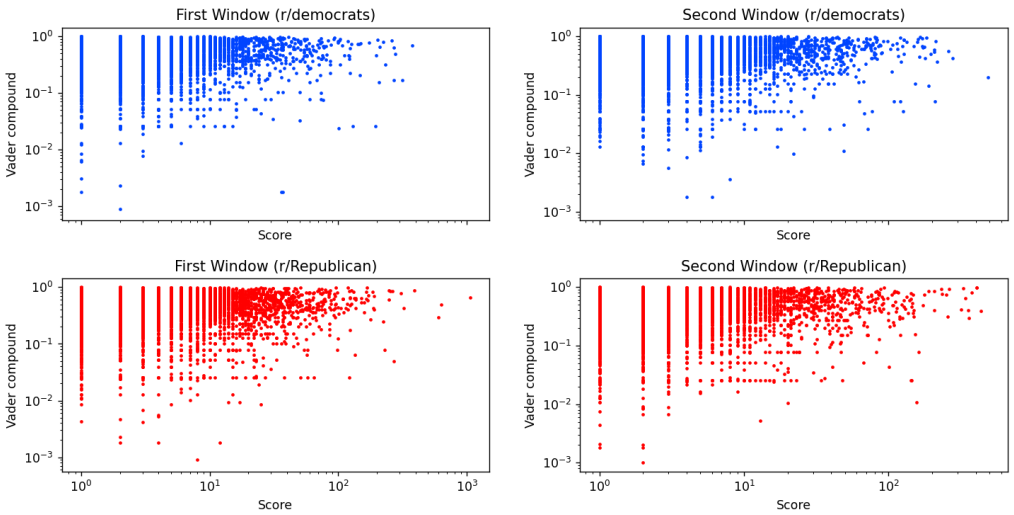

Figura 3.9: scatterplot con *score* e *compound* (*subreddit* a confronto)

3.5 Variazione di *score* e *compound* medi degli utenti

Preso atto dell'assenza di particolari trend rispetto ai commenti procediamo ricercando possibili correlazioni rispetto ai singoli utenti. Muoviamoci in due step:

1. analisi degli utenti attivi in entrambe le finestre;
2. analisi degli utenti attivi in una sola delle due finestre.

Informazione	#
Utenti attivi in entrambe le finestre temporali	3832
Utenti attivi solo in `first_window`	8233
Utenti attivi solo in `second_window`	9940
TOTALE	**22005**

3.5.1 Utenti attivi in entrambe le finestre temporali

Definiamo *utente attivo* un utente che ha pubblicato almeno un commento nel periodo della prima finestra temporale e almeno un commento nella seconda (3832 utenti).

- Consideriamo ogni singolo utente.
  - Calcoliamo *score medio* e *compound medio* dell'utente rispetto alla prima finestra. Facciamo la stessa cosa per la seconda finestra.
  - Calcoliamo la differenza tra score medi e la differenza tra compound medi.
- Costruiamo uno scatterplot dove
  - ogni punto rappresenta un utente,
  - l'asse x rappresenta le differenze tra score medi, e
  - l'asse y rappresenta le differenze tra compound medi.

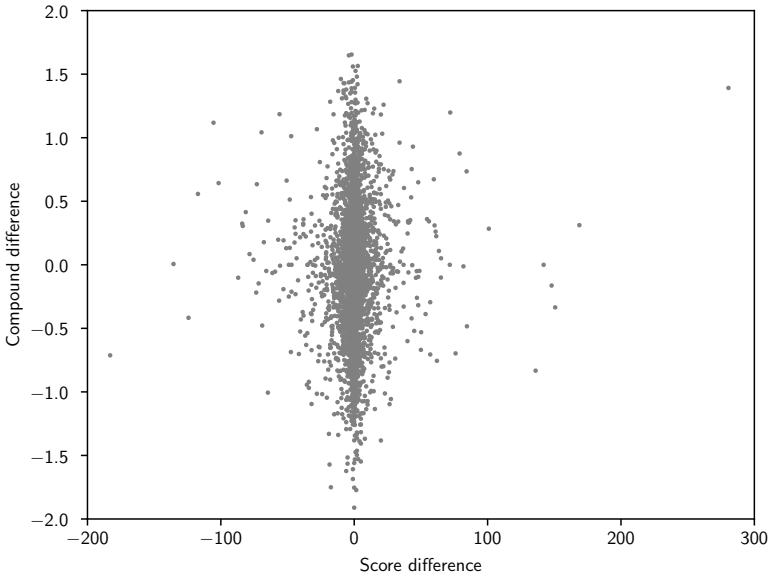

Figura 3.10: scatterplot con differenze di *score* e *compound* medi

Le variazioni di *score medio* risultano limitate nell'intorno di 0, mentre le variazioni di *compound medio* sono significative e coprono quasi interamente l'intervallo di valori $[-2; +2]$ (2 è il valore assoluto massimo che la variazione può assumere). Se generiamo scatterplot relativi ai singoli *subreddit* otteniamo un andamento simile per le variazioni di *compound medio*, mentre il *subreddit* `r/Republican` presenta maggiori differenze di score medio.

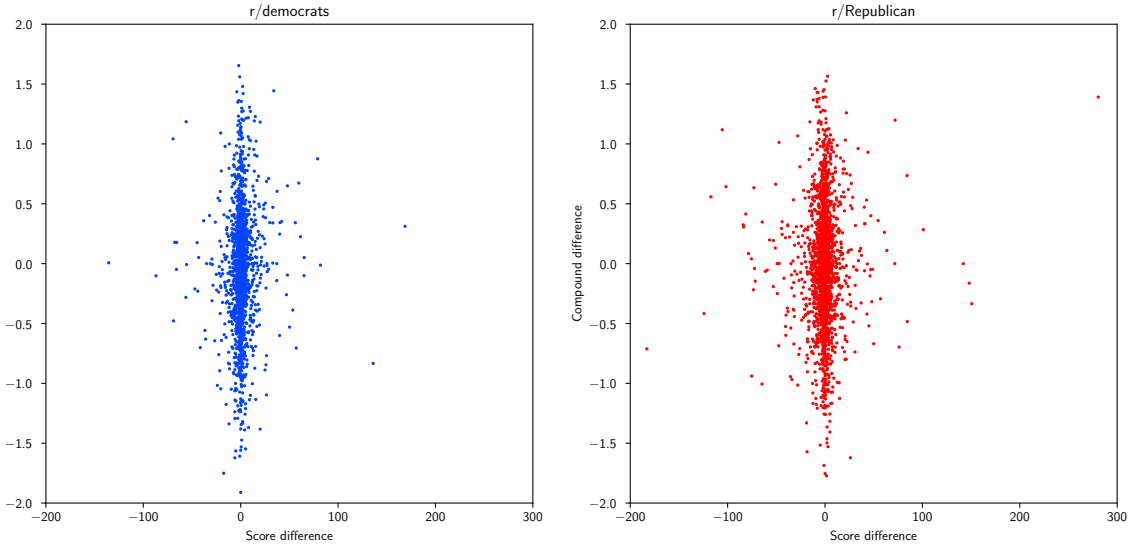

Figura 3.11: scatterplot con diff. di *score* e *compound* medi, distinti per *subreddit*

Complessivamente non sembrano emergere particolari correlazioni.

3.5.2 Utenti attivi in una sola delle due finestre temporali

In questa sezione consideriamo gli utenti che hanno pubblicato almeno un commento in una sola delle due finestre temporali (8233 utenti per la prima finestra temporale, 9940 per la seconda). Per questi utenti dobbiamo adottare un approccio diverso, dato che la presenza in una sola delle due finestre impedisce di calcolare le differenze tra *score* e *compound* medi.

- Consideriamo ogni singolo utente: calcoliamo *score medio* e *compound* medio all'interno della relativa finestra temporale.
- Costruiamo uno scatterplot dove
  - ogni punto rappresenta un utente,
  - il colore del punto dipende dalla relativa finestra temporale,
  - l'asse x rappresenta lo *score medio*, e
  - l'asse y rappresenta il *compound medio*.

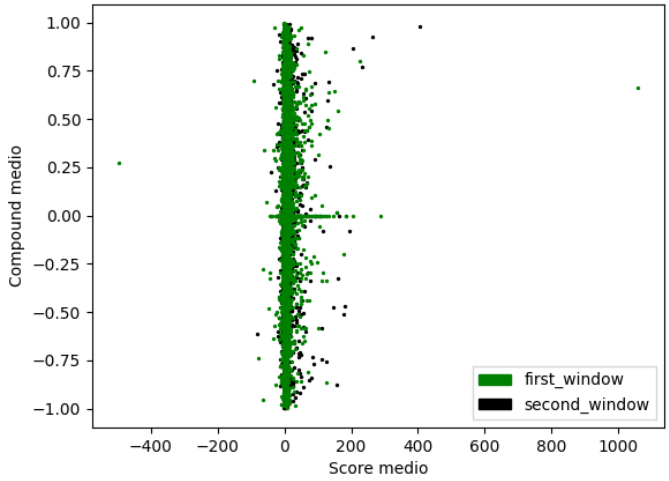

Figura 3.12: scatterplot con *score* e *compound* medi di utenti attivi solo in single finestre

Anche qua non emergono particolari correlazioni. Abbiamo la presenza di un principio di "topologia a stella" ("linea verticale", chiaramente visibile, e "linea orizzontale", costituita da punti con *score medio* diverso e *compound medio* uguale e nullo): a grandi cambiamenti di *score* non corrispondono grandi cambiamenti di *compound* e viceversa. Possiamo fare discorsi simili se generiamo scatterplot relativi ai singoli *subreddit*.

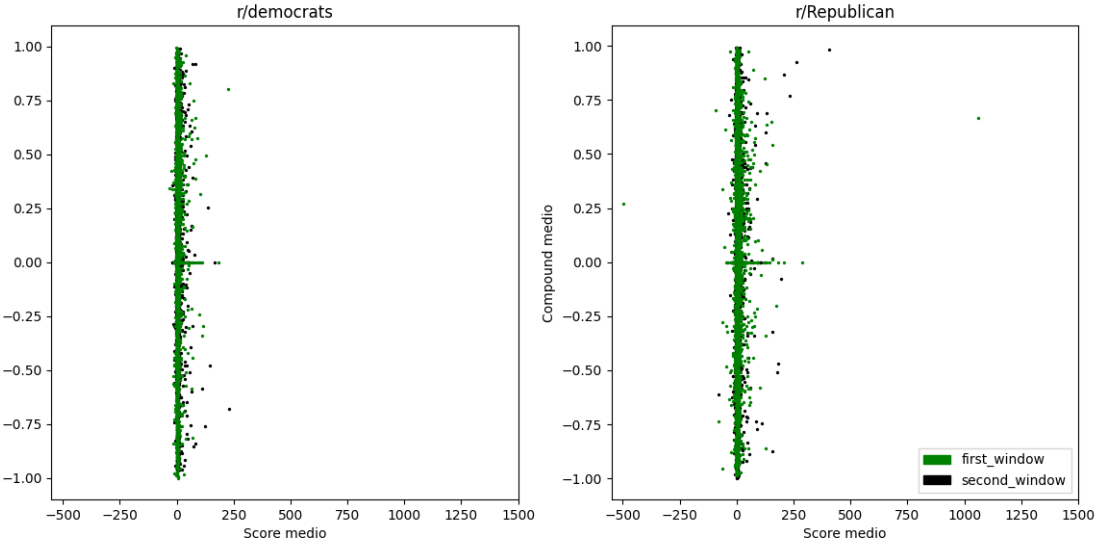

Figura 3.13: scatterplot con *score* e *compound* medi di utenti attivi in singole finestre, distinti per *subreddit*

3.6 Evoluzione giornaliera di *score* e *compound* medi

Non abbiamo ancora individuato trend significativi: cambiamo approccio. Abbiamo visto nella sezione 3.3.2 che la `second_window` è caratterizzata da un *compound medio* ridotto. Vogliamo capire meglio le cause: consideriamo l'evoluzione di *score* e *vader compound*.

1. Consideriamo tutti i commenti pubblicati nelle due finestre temporali.
2. Raggruppiamo i commenti per giorno.
3. Per ogni giorno calcoliamo la media dello *score* e la media del *vader compound*.
4. Facciamo la stessa cosa anche rispetto ai singoli subreddit (`r/democrats` e `r/Republican`)

Il risultato è il seguente grafico

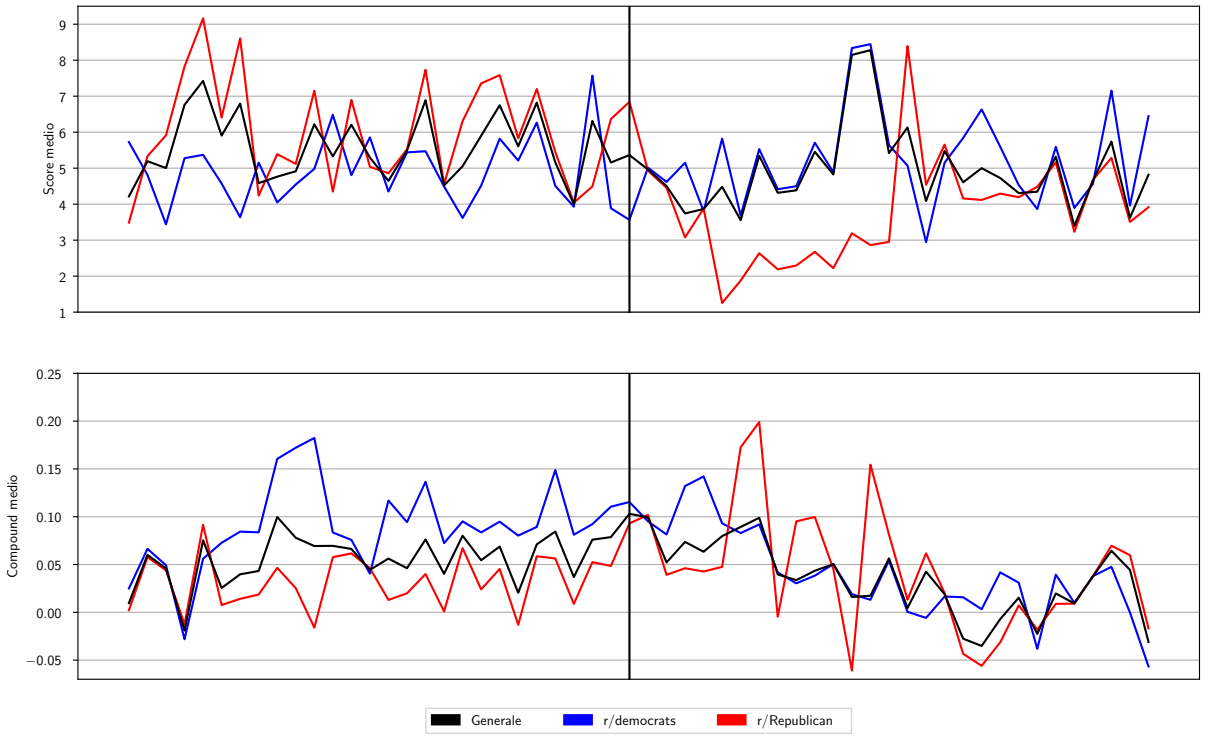

Figura 3.14: Numero di commenti giornalieri

Procediamo analizzando le due finestre temporali singolarmente.

3.6.1 Prima finestra temporale

Consideriamo la prima finestra temporale.

- Lo *score medio* giornaliero ha valori compresi tra 4.0 e 7.5
- Il *compound medio* giornaliero ha valori compresi tra -0.02 e 0.10.

Il *compound medio* non assume un valore nettamente positivo (*positive sentiment*) o nettamente negativo (*negative sentiment*), ma in un numero significativo di giorni si assume un valore leggermente superiore a $+0.05$ (classificabile come *positive sentiment*).

	Classificazione	#
Giorni con *compound medio* $\geq +0.05$	*Positive sentiment*	17
Giorni con *compound medio* > -0.05 e $< +0.05$	*Neutral sentiment*	11

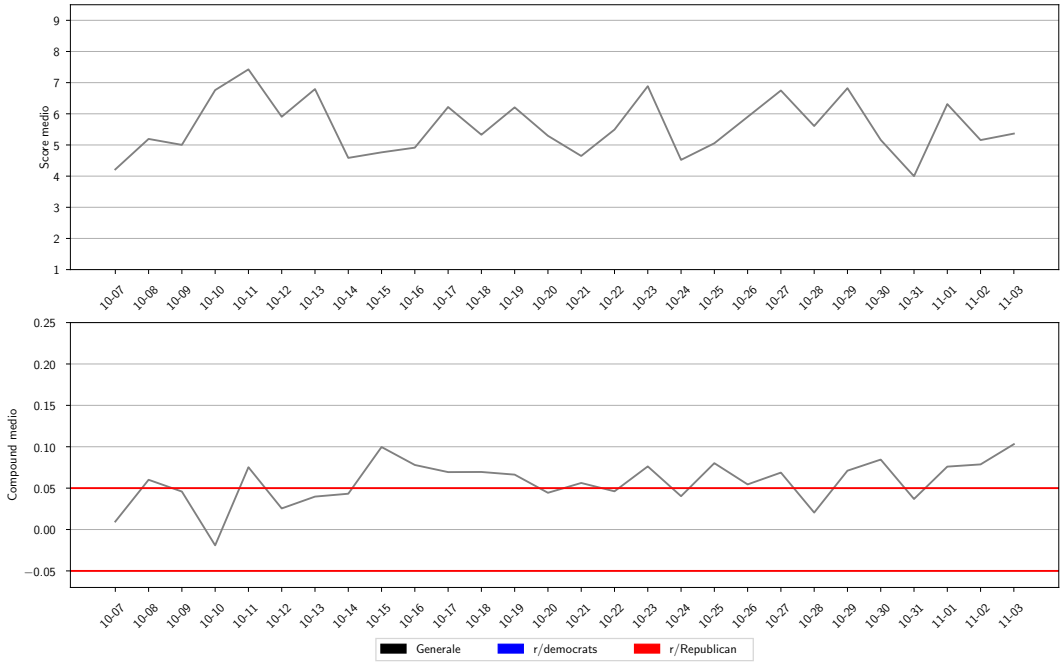

Figura 3.15: Evoluzione giornaliera di *score medio* e *compound medio*

Cosa succede rispetto ai singoli subreddit? Nel caso di `r/democrats` abbiamo

- *score medio* giornaliero che ha valori compresi tra $\approx +3.56$ e $\approx +7.57$, e
- *compound medio* giornaliero che ha valori compresi tra ≈ -0.03 e $\approx +0.18$

Il range di valori assunti dal *compound medio* è molto più grande, ma soprattutto è maggiore il numero di giorni in cui si assume un valore classificabile come *positive sentiment*

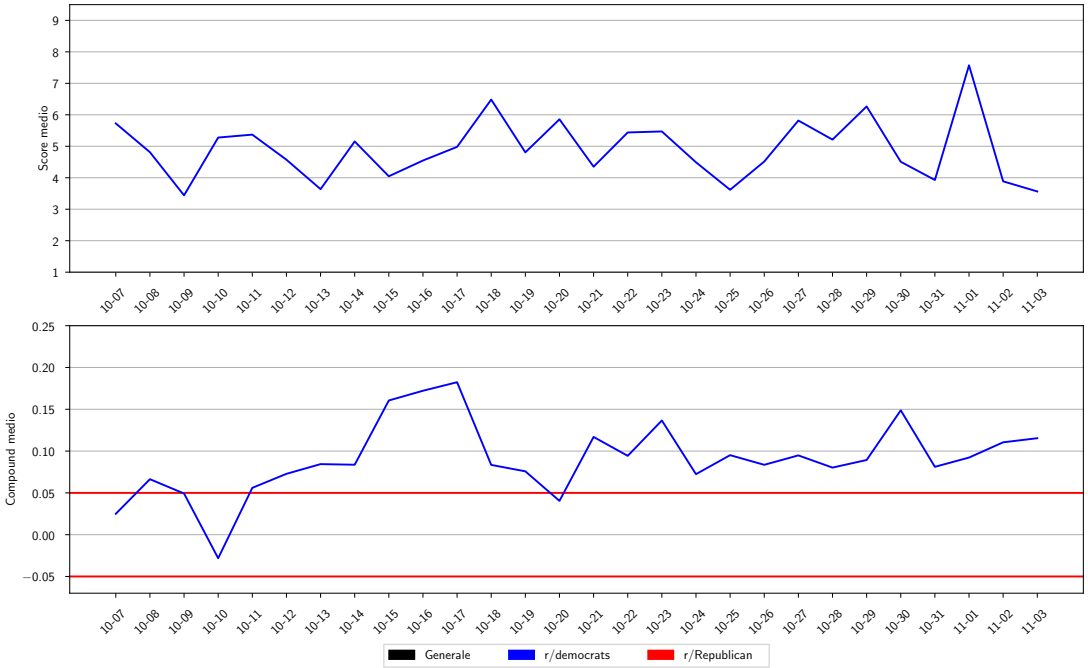

Figura 3.16: Evoluzione di *score medio* e *compound medio* su **r/democrats**

	Classificazione	#
Giorni con *compound medio* $\geq +0.05$	*Positive sentiment*	24
Giorni con *compound medio* > -0.05 e $< +0.05$	*Neutral sentiment*	4

Consideriamo adesso il subreddit `r/Republican`. Abbiamo

- *score medio* giornaliero con valori compresi tra $\approx +3.49$ e $\approx +9.16$, e
- *compound medio* giornaliero con valori compresi tra ≈ -0.016 e $\approx +0.092$

Notiamo come gli utenti di `r/Republican` siano maggiormente "generosi" in termini di *score*. Allo stesso tempo constatiamo che il *compound medio* assume valori $< +0.05$ (valori classificabili come *neutral sentiment*) nella maggior parte dei giorni

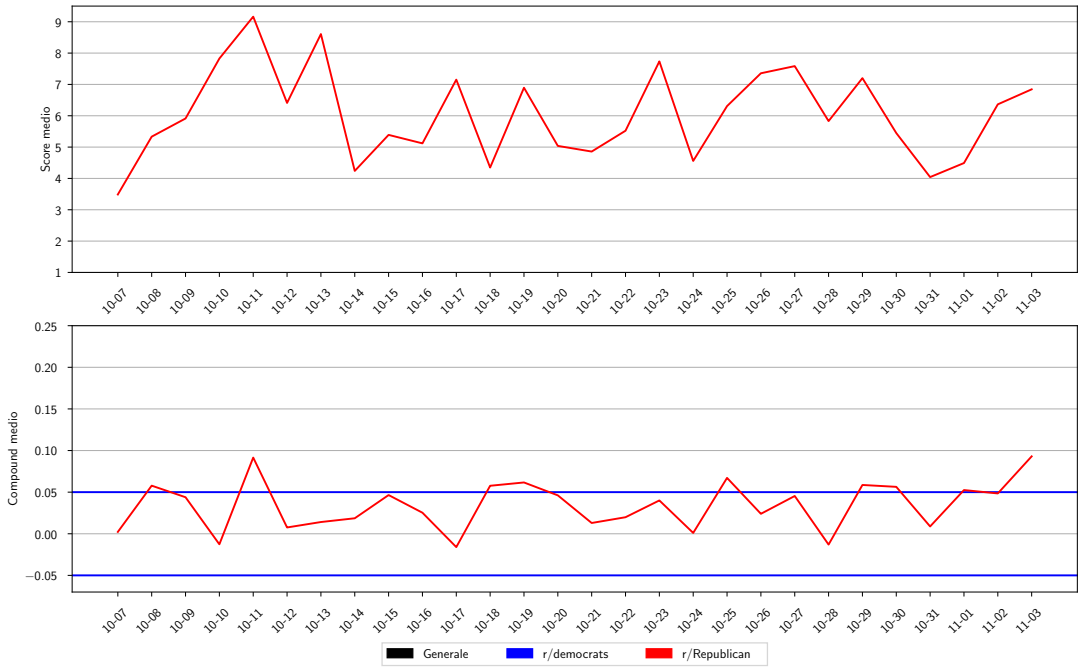

Figura 3.17: Evoluzione di *score medio* e *compound medio* su **r/Republican**

	Classificazione	#
Giorni con *compound medio* $\geq +0.05$	*Positive sentiment*	11
Giorni con *compound medio* > -0.05 e $< +0.05$	*Neutral sentiment*	17

3.6.2 Seconda finestra temporale

Consideriamo adesso la seconda finestra temporale.

- Lo *score medio* giornaliero ha valori compresi tra $\approx +3.40$ e $\approx +8.28$
- Il *compound medio* giornaliero ha valori compresi tra ≈ -0.035 e $+0.10$.

Il range di valori assunto dallo *score medio* risulta ampliato rispetto alla prima finestra, segnale di una maggiore polarizzazione. Il *compound medio* assume un valore $\in]-0.05; +0.05[$ nella maggior parte dei giorni della finestra (novità rispetto alla prima finestra temporale, dove prevalevano i giorni con *compound medio* $\geq +0.05$)

	Classificazione	#
Giorni con *compound medio* $\geq +0.05$	*Positive sentiment*	10
Giorni con *compound medio* > -0.05 e $< +0.05$	*Neutral sentiment*	18

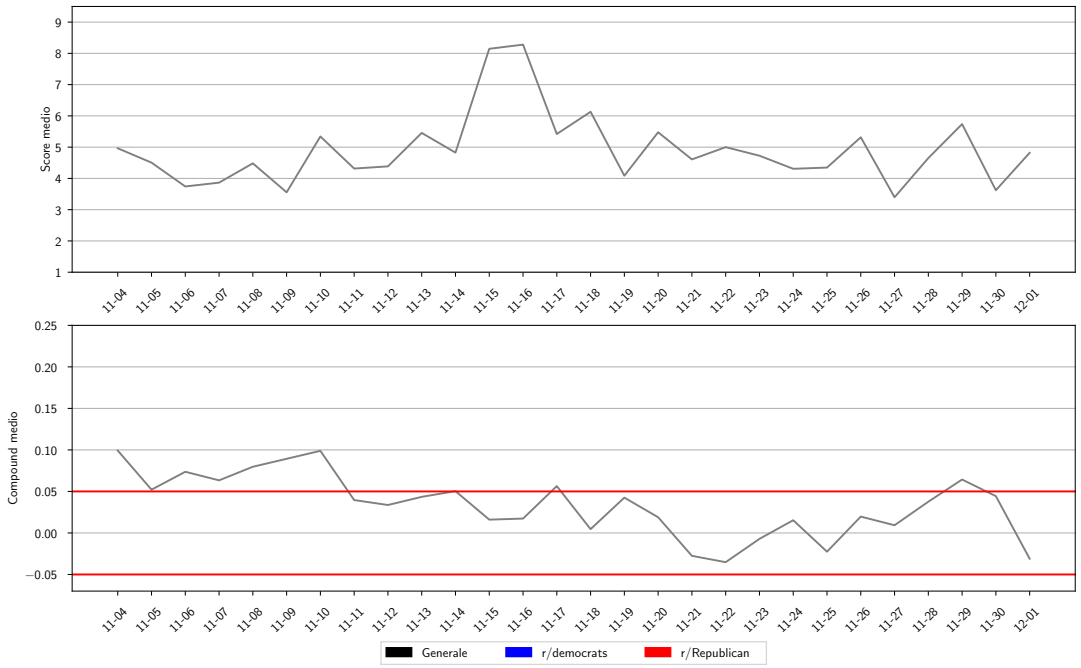

Figura 3.18: Evoluzione giornaliera di *score medio* e *compound medio*

Consideriamo i singoli subreddit. Nel caso di `r/democrats` abbiamo

- *score medio* che ha valori compresi tra $\approx +2.94$ e $\approx +8.45$, e
- *compound medio* che ha valori compresi tra ≈ -0.06 e $\approx +0.14$

Il range di valori assunti dallo *score medio* è più grande rispetto alla prima finestra, ma si rovescia quanto già visto nella stessa rispetto al *compound medio*: non solo prevalgono i giorni con valori classificabili come *neutral sentiment*, ma abbiamo addirittura un giorno con un *compound medio* ≤ -0.05 (quindi classificabile come *negative sentiment*).

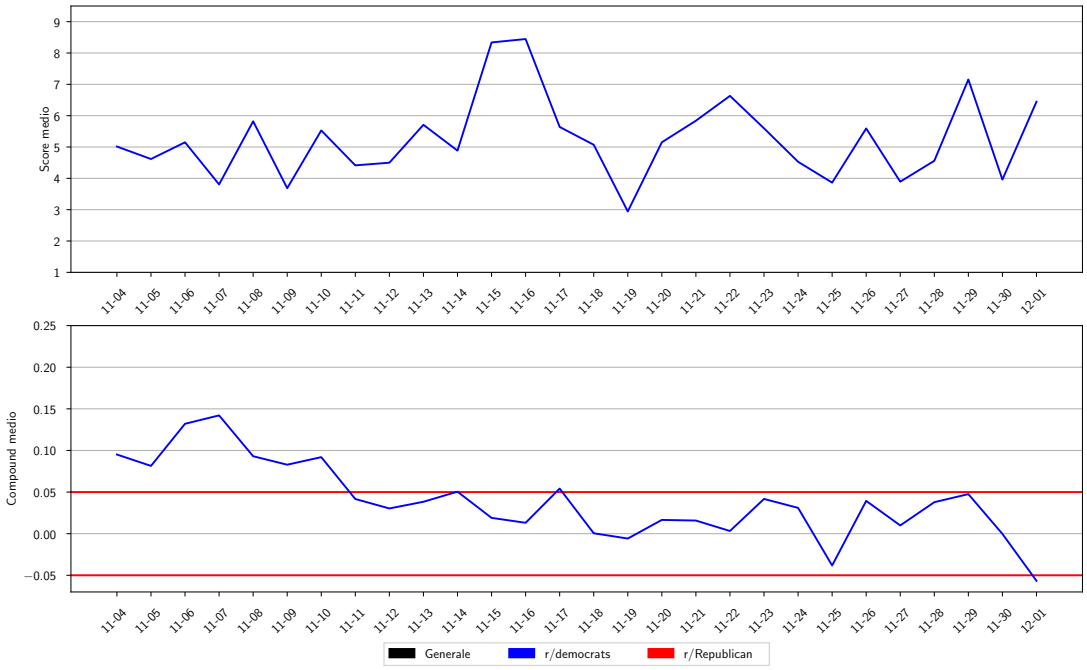

Figura 3.19: Evoluzione di *score medio* e *compound medio* su **r/democrats**

	Classificazione	#
Giorni con *compound medio* ≥ 0.05	*Positive sentiment*	9
Giorni con *compound medio* > -0.05 e $< +0.05$	*Neutral sentiment*	18
Giorni con *compound medio* ≤ -0.05	*Negative sentiment*	1

Consideriamo adesso il subreddit `r/Republican`. Abbiamo

- *score medio* giornaliero con valori compresi tra $\approx +1.25$ e $\approx +8.39$, e
- *compound medio* giornaliero con valori compresi tra ≈ -0.06 e $+0.20$

Si conferma il trend visto nella prima finestra temporale, dove i valori con sentiment "non positivo" prevalgono. Osserviamo che in questa finestra i giorni con sentiment "non positivo" aumentano di una unità, da 17 a 18: di questi 18 due hanno addirittura un *compound medio* ≤ -0.05 (mentre prima erano tutti valori classificabili come *neutral sentiment*).

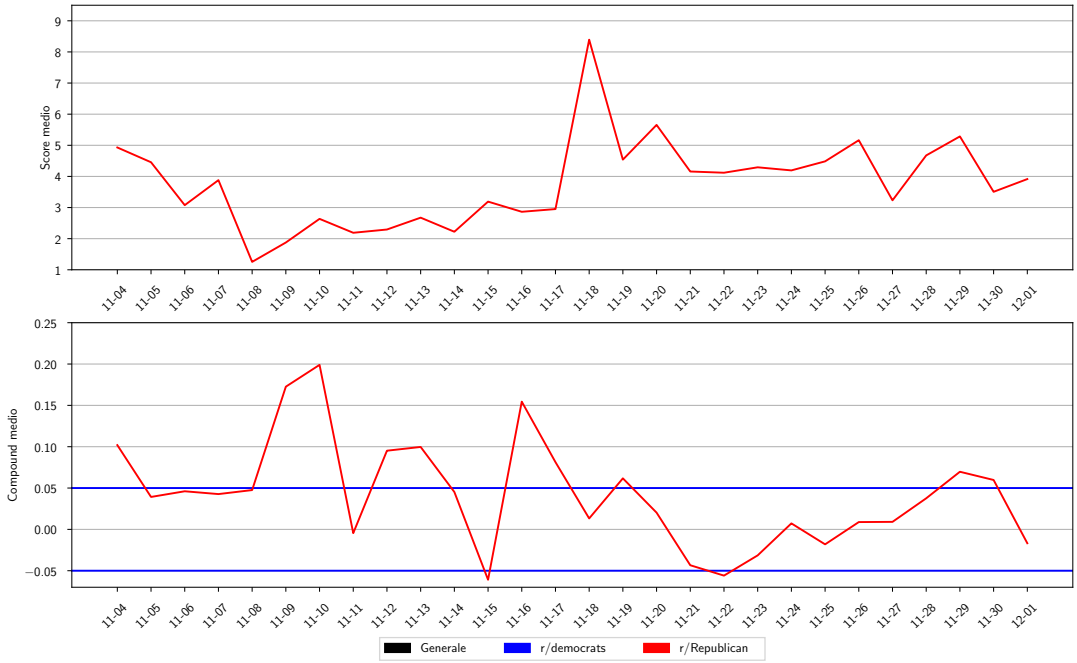

Figura 3.20: Evoluzione di *score medio* e *compound medio* su **r/Republican**

	Classificazione	#
Giorni con *compound medio* ≥ 0.05	*Positive sentiment*	10
Giorni con *compound medio* > -0.05 e $< +0.05$	*Neutral sentiment*	16
Giorni con *compound medio* ≤ -0.05	*Negative sentiment*	2

3.7 Analisi di cui al 3.5 considerando solo gli *outlier*

3.7.1 Individuazione degli *outlier*

Concludiamo svolgendo un'analisi sugli *outlier*, partendo dagli scatterplot della sezione *Variazioni di score e compound medi degli utenti*.

- Per l'individuazione degli *outlier* abbiamo utilizzato la libreria *scikit-learn*, contenente funzioni per il machine learning.
- La funzione di nostro interesse è `IsolationForest()`, che permette l'applicazione dell'omonimo algoritmo.

Abbiamo recuperato il `dataset` utilizzato per la costruzione del dataset, limitandolo alle sole colonne contenenti le coordinate del punto sullo scatterplot:

- `diff_score` e `diff_compound` per gli utenti attivi in entrambe le finestre, e
- `average_score` e `average_compound` per gli utenti attivi in una sola delle due finestre.

Per mezzo del seguente codice otteniamo un array `predictions` dove si segnala, per ogni punto, se questo è un *outlier* (valore -1) o un *inlier* (valore $+1$).

```
1    IF = IsolationForest()
2    predictions = IF.fit_predict(dataframe)
```

Ottenuto questo array rimuoviamo dal `dataset` tutti gli *inlier*. In questo modo potremo generare degli scatterplot con solo gli *outlier* e concentrarci sul loro "comportamento" (gli *inlier* hanno un "comportamento" simile a quello globale, già analizzato in 3.5).

3.7.2 Utenti attivi in entrambe le finestre

Per prima cosa consideriamo gli utenti attivi in entrambe la finestre.

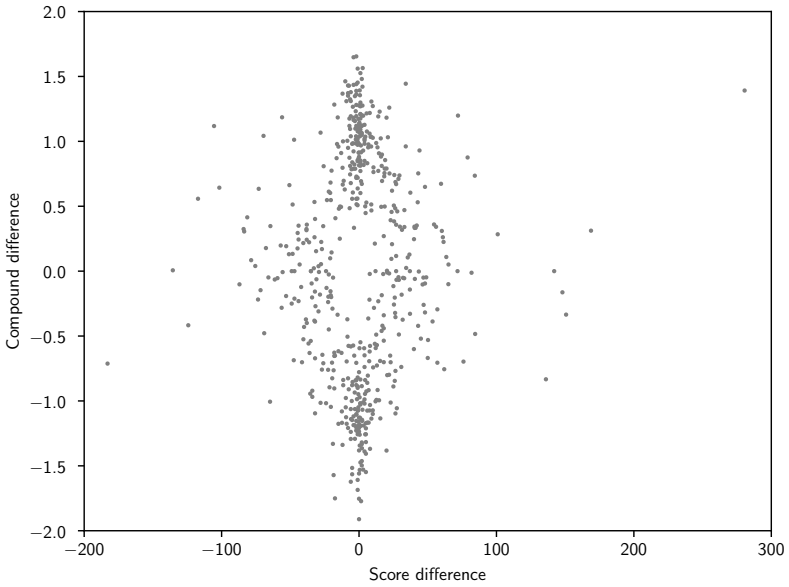

Figura 3.21: scatterplot con differenze di *score* e *compound* medi, solo *outliers*

Informazione	#
Outliers	572
Normal samples	3260
TOTALE	**3832**

Con il mantenimento dei soli *outlier* notiamo una maggiore polarizzazione, in quanto la maggioranza degli stessi ha un `diff_compound` $\geq +0.5$ e ≤ -0.75.

Informazione	#
Outliers con `diff_compound` $\geq +0.5$	205
Outliers con `diff_compound` ≤ -0.75	145
TOTALE OUTLIERS "ESTREMI"	350
Outliers rimanenti	222
TOTALE OUTLIERS	**572**

Utilizziamo nuovamente la funzione `IsolationForest()` per trovare gli *outlier* rispetto ai singoli subreddit.

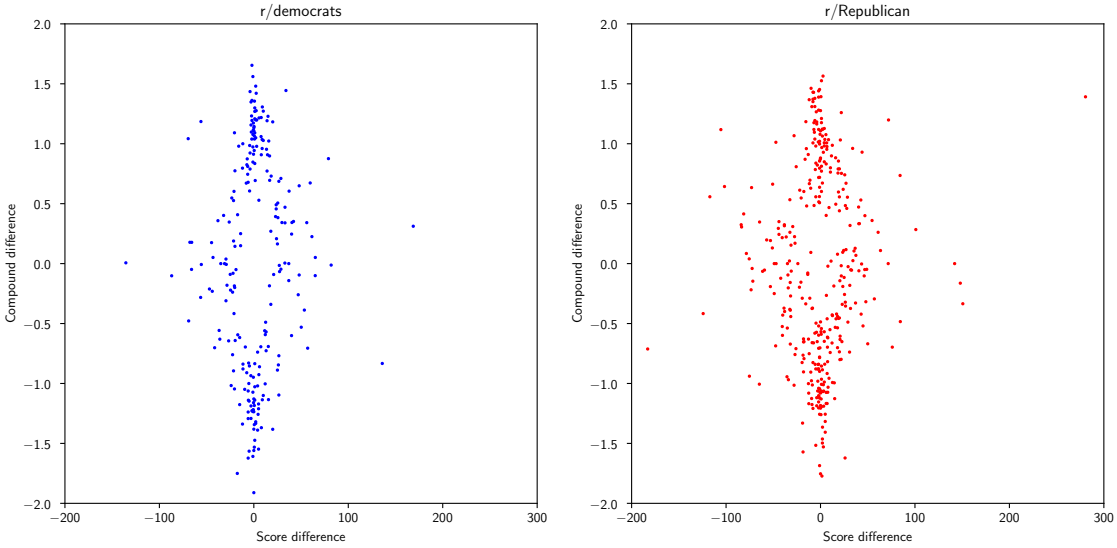

Figura 3.22: scatterplot con diff. di *score* e *compound* medi, *outliers* distinti per *subreddit*

Subreddit *r/democrats*	
Informazione	#
Outliers	233
Normal samples	1442
TOTALE	**1675**

Subreddit *r/Republican*	
Informazione	#
Outliers	366
Normal samples	1721
TOTALE	**2087**

Anche in questo caso assistiamo a una polarizzazione, per le stesse ragioni di prima.

Subreddit *r/democrats*	
Informazione	#
Outliers con `diff_compound` $\geq +0.5$	84
Outliers con `diff_compound` ≤ -0.75	65
TOTALE OUTLIERS "ESTREMI"	149
Outliers rimanenti	84
TOTALE OUTLIERS	**233**

Subreddit *r/Republican*	
Informazione	#
Outliers con `diff_compound` $\geq +0.5$	112
Outliers con `diff_compound` ≤ -0.75	88
TOTALE OUTLIERS "ESTREMI"	200
Outliers rimanenti	166
TOTALE OUTLIERS	**366**

3.7.3 Utenti attivi in una sola delle due finestre

Adesso consideriamo gli utenti che sono stati attivi in una sola delle due finestre.

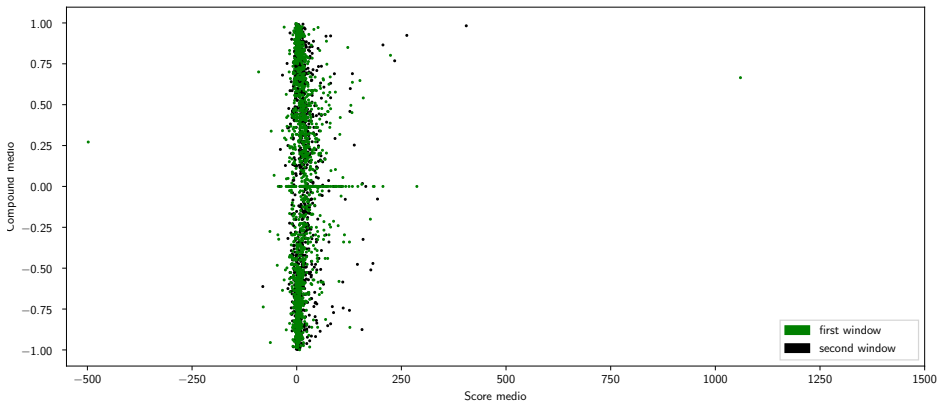

Figura 3.23: scatterplot con diff. di *score* e *compound* medi, *outliers* distinti per *subreddit*

Informazione	#
Outliers	3212
Normal samples	14961
TOTALE	**18173**

L'immagine mostra una parvenza di polarizzazione, visti gli elementi rimossi! Per avere una visione più chiara poniamoli in scatterplot differenti i punti della prima e della seconda finestra temporale.

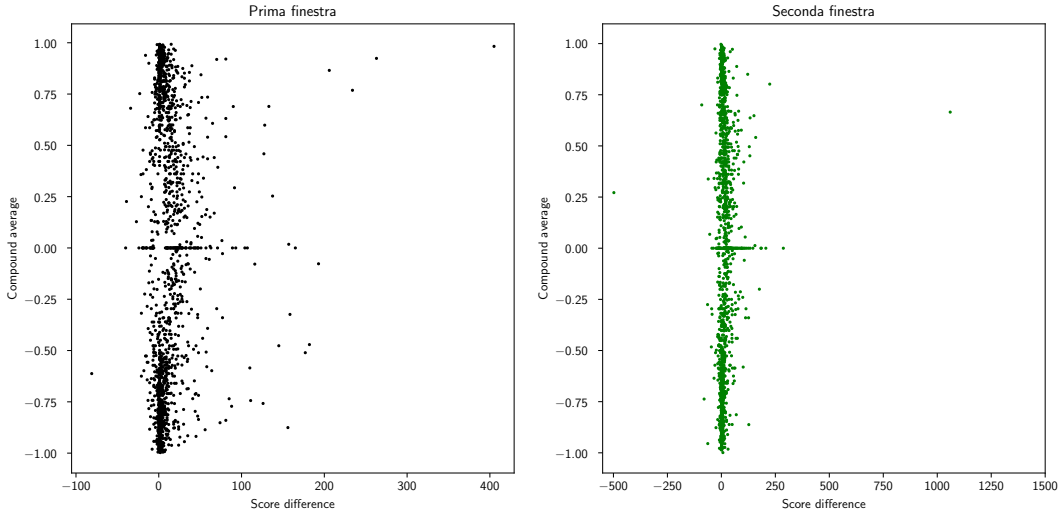

Figura 3.24: scatterplot con diff. di *score* e *compound* medi, *outliers* distinti per finestra

Prendiamo a riferimento i *threshold* $+0.5$ e -0.5: si conferma la parvenza di prima.

Prima finestra temporale		**Seconda finestra temporale**	
Informazione	#	**Informazione**	#
Outliers con `average_compound` $\geq +0.5$	457	*Outliers con* `average_compound` $\geq +0.5$	497
Outliers con `average_compound` ≤ -0.5	503	*Outliers con* `average_compound` ≤ -0.5	622
TOTALE OUTLIERS "ESTREMI"	960	**TOTALE OUTLIERS "ESTREMI"**	1119
Outliers rimanenti	646	*Outliers rimanenti*	487
TOTALE OUTLIERS	**1606**	**TOTALE OUTLIERS**	**1606**

Capitolo 4

Conclusioni

L'esito delle studio offre dei risultati non estremamente netti. Non sono emerse correlazioni dirette tra *score* e *vader compound*, correlazioni che abbiamo cercato di individuare in due contesti:

- rispetto ai singoli commenti (3.4), analizzando per ogni commento *score* e *compound*
- rispetto agli utenti (3.5), analizzando:
  - `diff_score` e `diff_compound` per gli utenti che hanno pubblicato almeno un commento in entrambe le finestre temporali;
  - `average_score` e `average_compound` per gli utenti che hanno pubblicato commenti solo in una delle due finestre temporali.

Lo studio ha permesso, nonostante ciò, di individuare un aumento della polarizzazione nella seconda finestra temporale (e quindi una maggiore polarizzazione a seguito dell'*election day*).

1. L'evoluzione di *score* e *compound* medi giornalieri (3.6) indica un range più grande di valori assunti nella seconda finestra temporale

Informazione	**FW**	**SW**
Score medio	$[+4.0, +7.5]$	$[\approx +3.40, \approx +8.28]$
Compound medio	$[-0.02, +0.10]$	$[\approx -0.035, +0.10]$

   All'aumento dello *score medio massimo* non è seguito un aumento del *compound medio massimo*, mentre a una diminuzione dello *score medio minimo* è seguita una diminuzione del *compound medio minimo* (aumenta il numero di commenti classificati come *neutral sentiment* e *negative sentiment* - come visto a 3.3.2). Risultati simili, anche se ciascuno con le proprie peculiarità, possono essere visti rispetto ai subreddit `r/democrats` e `r/Republican`

2. Si ha una maggiore polarizzazione dei cosiddetti *utenti outlier* (3.5): sia gli utenti attivi in entrambe le finestre temporali, sia gli utenti attivi in una sola di esse. Tale polarizzazione si manifesta sia nella prima che nella seconda finestra temporale ed è maggiore nella seconda: differenza dettata non tanto dalla natura "straordinaria" dell'esito elettorale, ma dalla discussione dei risultati elettorali (polarizzazione che si potrebbe definire "fisiologica" nel dibattito politico - digitale o meno che sia)

La tematica può definirsi tutt'altro che chiusa: da una parte evidenziamo la complessità del comportamento delle persone, che risulta difficile definire in maniera algoritmica (assenza di correlazioni dirette tra *score* e *compound*); dall'altra osserviamo come la polarizzazione, che coinvolge soprattutto gli *outlier* (3.7), ricade sul "comportamento globale" del *subreddit* (3.6) e quindi dell'intera community di Reddit.

Suggerimenti per futuri studi Il presente studio vuole porsi come base per future riflessioni. Si segnala come possibili idee:

- l'analisi di un maggior numero di *subreddit*, scelta che dovrà tenere conto della loro attività (in termini di *submissions* e *comments*) nelle finestre elettorali considerate;
- l'analisi delle *submissions* e l'individuazione di possibili correlazioni tra <*score*,*compound*> delle *submission* con <*score*,*compound*> dei relativi *comments*
- l'individuazione di correlazioni tra <*score*,*compound*> di *comments* e *submissions* con particolari eventi politici avvenuti durante le due finestre temporali;
- studio sulle attività dei *subreddit* `r/democrats` e `r/Republican` nei giorni successivi alla seconda finestra temporale, con particolare riferimento all'assedio del Campidoglio del 6 gennaio 2021.

Capitolo 5

Bibliografia

[1] Lorenzo Cima, Amaury Trujillo Larios, Marco Avvenuti, and Stefano Cresci. The great ban: Efficacy and unintended consequences of a massive deplatforming operation on reddit. *arXiv preprint arXiv:2401.11254*, 2024.

[2] Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225, 2014.

[3] RaiderBDev stuck_in_the_matrix, Watchful1. Reddit comments/submissions 2005-06 to 2023-12 (`https://academictorrents.com/details/9c263fc85366c1ef8f5bb9da0203f4c8c8db75f4`).

[4] Amaury Trujillo and Stefano Cresci. Make reddit great again: assessing community effects of moderation interventions on r/the_donald. *Proceedings of the ACM on Human-computer Interaction*, 6(CSCW2):1–28, 2022.