### *Elezioni USA 2012, spunti di analisi di big data applicati a Twitter*

### Progetto per il corso di Big Data

### Master in Data Science, Università degli studi di Roma di Torvergata - 2024

Gruppo: Gisella Coppola[[1]](#footnote-1) e Salvatore dello Russo[[2]](#footnote-2)

### **Introduzione**

### Il presente report si propone di utilizzare le metodologie e strumenti acquisiti durante il corso Big Data per l’analisi dei dati, in particolare dei *big data*[[3]](#footnote-3), provenienti dai social media applicati al contesto delle elezioni del Presidente degli Stati Uniti d’America.

### A poco più di tre mesi dalle elezioni 2024, la tematica scelta è stata selezionata in virtù: (i) dell’ampia letteratura sul ruolo chiave dei social media in contesti elettorali, (ii) della prossimità temporale alle prossime elezioni americane, quale uno degli eventi politici più attesi dell’anno e determinanti a livello geopolitico mondiale e (iii) della significativa disponibilità di dataset sull’argomento.

### Nello specifico per il data fetching abbiamo utilizzato la piattaforma Kaggle. Il data set considerato riguarda i testi dei tweets ed è in formato json.

### Il titolo del data set è “Twitter 2012 USA presidential election”, ed è reperibile al seguente link: <https://www.kaggle.com/datasets/jgoodman8/twitter-2012-presidential-election>.

Il ruolo strategico dei social media in campo politico-istituzionale, e in particolar modo di Twitter, ha acquisito negli anni sempre più rilevanza. Attualmente non è pensabile escludere questi mezzi di comunicazione dalle attività politiche e di rappresentanza, e ancora di più dall’organizzazione di una campagna elettorale. L’esponenziale crescita, con rispettive ripercussioni, dell’utilizzo dei social media, hanno fatto sì che in occasione delle ultime elezioni americane del 2020 per la prima volta gli stessi social media sono arrivati ad applicare misure di intervento e di contrasto al deepfake[[4]](#footnote-4). Il presente elaborato volge lo sguardo ancora più indietro, sino al 2012, anno in cui, contestualmente alla campagna elettorale per il 44° Presidente degli Stati Uniti, Twitter raddoppiava il numero degli utenti giornalieri. Fu l’anno in cui per la prima volta un presidente americano comunicò la sua vittoria con un tweet. In occasione della rielezione di Obama, venne infine registrato il primato del suo tweet "four more years" per record di ricondivisioni fino a quel momento.

### Le elezioni americane del 2012 hanno di fatti visto la vittoria di [Barack Obama](https://it.wikipedia.org/wiki/Barack_Obama), [Presidente uscente](https://it.wikipedia.org/wiki/Presidenti_degli_Stati_Uniti_d%27America) sostenuto dal [Partito Democratico](https://it.wikipedia.org/wiki/Partito_Democratico_(Stati_Uniti_d%27America)), che ha sconfitto [Mitt Romney](https://it.wikipedia.org/wiki/Mitt_Romney), candidato per il [Partito Repubblicano](https://it.wikipedia.org/wiki/Partito_Repubblicano_(Stati_Uniti_d%27America)). Nel dataset oggetto del presente studio, ci troviamo in un contesto temporale di piena campagna elettorale; più precisamente, siamo a meno di due mesi dalle elezioni e a due settimane dal primo confronto televisivo tra i due candidati.

**DESCRIZIONE QUALITATIVA DEL DATA SET**

### Il data set consiste in una raccolta di tweet riguardanti le elezioni presidenziali statunitensi del 2012 che sono stati estratti tra il 13 ed il 14 settembre 2022. Piu specificatamente, il data set contiene 1,000,000 di tweets (a cui ci riferiremo come secondari) che includono sia dei retweet che delle repliche ad altri tweets. In esso, sono presenti anche i tweets sorgente (o source) che vengono retwettati. Per questi, il dataset fornisce le stesse informazioni presenti per i tweets secondari, i.e. possiamo sapere se un tweet source replica a un altro tweet o meno, ecc.

### Ispezionando i tweets a disposizione (secondari e sorgenti), catturati nelle poco più di 24 ore che intercorrono nell’arco temporale a disposizione, la presente analisi è stata sviluppata intorno ai seguenti obiettivi:

### Individuare le principali tematiche circolanti sul social a proposito di entrambi i candidati nello specifico contesto storico di riferimento;

### Analizzare (tramite pyspark) il sentiment, l’entità e la viralità dei tweet che presentano un esplicitato livello di engagement verso i due candidati, tramite la presenza di hashtags a loro dedicati. Un focus specifico sarà dedicato alle tematiche con sentiment positivo del candidato vincente Obama;

### Analizzare le reti di utenti tramite il modulo graphframe di Spark; in particolare, abbiamo considerato la rete dei retweet (archi) tra utenti (nodi), essendo la ricondivisione una delle funzionalità più d’impatto dei social media in generale.

Con questi ambiziosi obiettivi di analisi, il presente report si propone di applicare le tecniche acquisite sul trattamento dei Big Data per trovare spunti e campi di applicazione nello studio dell’evoluzione delle campagne elettorali attuate sui social media.

### **Preparazione dei dati**

### Il data set considerato è contenuto in file con estensione json. Per caricarlo in memoria e renderlo disponibile, come un oggetto DataFrame, per l’applicazione sviluppata tramite pyspark è stato utilizzato il metodo *read.json* relativo alla SparkSession (cfr. “Script1.py”)[[5]](#footnote-5).

### A seguito del suo caricamento, è stata analizzata la struttura tabellare originaria che presenta le seguenti caratteristiche:

### Arco temporale coperto: 13-09-2022 03:30:22 / 14-09-2022 05:25:11

### N°. totale di colonne: 271, di cui

### N°. di colonne primarie: 23;

### N°. di sottocolonne: 248;

### N°. totale di righe: 1,000,000 tweet, di cui

### N°. 114,457 (11%[[6]](#footnote-6)) tweet pubblicati in replica ad altri tweet;

### N°. 444,457 (44%) tweet pubblicati per la condivisione di altri tweet;

### N°. 440,759 (44%) tweet pubblicati senza replica o condivisione.

### Lo schema del dataframe utilizzato è riportato nel file allegato (“Schema.txt”).

### Ai fini delle analisi sono stati poi individuati, esplorati e processati, principalmente, i seguenti campi e sottocampi:

### created\_at (data e ora di creazione del tweet);

### entities.hashtags.text (lista di stringhe contenente gli hashtags);

### user.name (nome utente del tweet secondario);

### in\_reply\_to\_screen\_name (nome utente dell’utente a cui ha eventualmente risposto il tweet secondario);

### text (testo del tweet secondario);

### user.time\_zone (fascia di fuso orario);

* retweet\_status.user.name (nome dell’utente il cui tweet[[7]](#footnote-7) è stato retwettato da un tweet secondario);
* retweet\_status.in\_reply\_to\_screen\_name (nome dell’utente a cui il tweet source eventualmente rispondeva);
* retweet\_count (numero di retweet del tweet sorgente).

### Le attività preliminari di pulizia dati svolte sono state (le operazioni di seguito sono da intendersi applicate in maniera sequenziale):

### Eliminazione dei duplicati di riga. Risultato: n°. 327 righe rimosse;

### Eliminazione di righe corrispondenti a tweet senza testo (NULL o campo vuoto). Risultato: n°. 0 righe rimosse;

### Delimitazione geografica[[8]](#footnote-8) del dataset tramite la sotto-colonna “user.time\_zone”. Risultato: n°. 593,029 totale di righe rimosse per i tweet non contenuti la stringa “%US & Canada%” nella cella corrispondente alla sotto-colonna “user.time\_zone”;

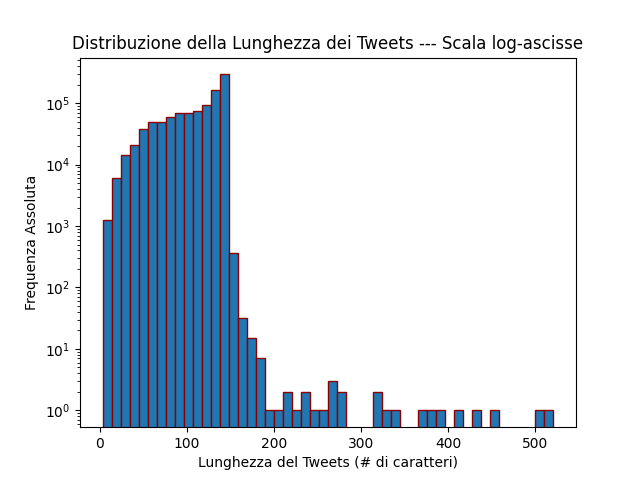
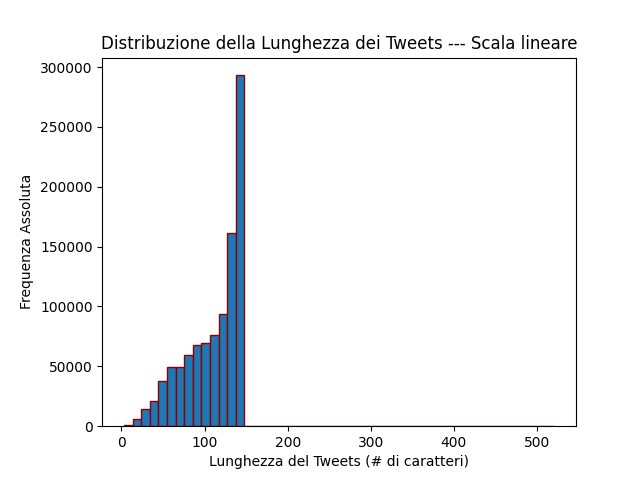
### Filtraggio dei tweet con hashtag per concentrare le analisi sui soli tweet che presentano questa forma di engagement. Risultato: n°. 305,662 di righe eliminate.

Come risultato delle operazioni di filtraggio sopra citate, il numero di tweets contenuto nel dataframe arriva ad essere n°. 100,982[[9]](#footnote-9), cioè circa il 10% di quello iniziale (1,000,000).

**VISUALIZZAZIONE DELLA LUNGHEZZA DEI TWEETS**

È stato graficato l’istogramma della lunghezza dei testi dei tweets (misurata in singoli caratteri). Di seguito sono riportate due versioni dello stesso istogramma, di cui una con scala delle ordinate logaritmica[[10]](#footnote-10).

Figura 1: Istogramma delle lunghezze dei tweets in caratteri, riferite al dataset in cui sono stati eliminati sono i duplicati.



In particolare, abbiamo dedotto che la lunghezza media dei tweets (nelle stesse condizioni degli istogrammi) arrotondata per difetto è 110 caratteri e che sono molto più presenti i tweets dai 150-160 caratteri in giù.

### Successivamente alla prima fase di esplorazione e pulizia dati, sono state implementate ulteriori attività di data cleaning che saranno implementate a seconda della specifica tipologia di analisi (i.e. calcolo frequenze, wordcloud, sentiment analysis ecc.):

### Conversione del testo in minuscolo e suddivisione in parole;

### Applicazione filtri per l’eliminazione delle stopwords;

### Rimozione di URL, mentions, hashtags e caratteri speciali;

### Conversione della lista di stringhe (campo user.hashtags.text) che contiene gli hashtags in una stringa che contiene la sequenza degli hashtags separati da “|”.

### **Analisi dei dati**

Principalmente sono state effettuate analisi di tipo testuale (cfr. “Script2.py”) e di networks (cfr. “Script3.py”).

**DELIMITAZIONE GEOGRAFICA DEL DATA SET**

Con l’intenzione di focalizzare l’analisi sulla popolazione statunitense, cioè il corpo elettorale americano, in fase di esplorazione dati è stata riscontrata l’inaccuratezza dei sottocampi della colonna primaria “geo”. Infatti, tali sott campi presentavano un valore diverso da NULL per un numero residuale di tweets rispetto al dataframe originale. Come riferimento per la geolocalizzazione degli utenti, è stata selezionata pertanto il sottocampo “user.time\_zone” (avente un numero rilevante di celle non NULL). In particolare, dall’ultimo sottocampo citato, supponendolo attendibile, possiamo evincere che i tweets del data set originale sono relativi ad utenti distribuiti in tutto il mondo (cfr. file “First 100 rows of time\_zone column.txt”).

### In termini di esplorazione dati, abbiamo quindi osservato quanto segue: i tweet che presentano un campo “user.time\_zone” che non NULL sono

### di cui quelli che contengono la stringa “%US & Canada%” sono

### La dimensione del dataset ridotto è stata pertanto ritenuta appropriata per lo svolgimento dell’analisi computazionale dei contenuti testuali.

### **ANALISI DESCRITTIVA DEL DATA SET USA**

### Come risultato della fase preliminare di pulizia dati il dataframe relativo agli USA risulta costituito da:

### Totale di 23 colonne principali;

### Totale di 406,664 (# di righe) tweet, di cui

### N°. 39,765 (10%) tweet pubblicati in replica ad altri tweet;

### N°. 192,714 (47%) tweet pubblicati per la condivisione di altri tweet;

### N°. 174165 ( 43%) tweet pubblicati senza replica o condivisione;

### Lunghezza media dei tweet: 113 caratteri. Non c’è una differenza significativa in lunghezza rispetto ai tweet del data set in cui sono stati rimossi soltanto i duplicati.

**SUDDIVISIONE E ANALISI DESCRITTIVA DEL DATAFRAME PER CIASCUN CANDIDATO**

È stato applicato un filtro sul sottocampo “hasthags.entities.text” per ottenere i tweets con hashtag obama e non romney e viceversa. In questo modo dal data set USA sono stati estratti i tweets scritti da utenti che sono dalla parte di uno dei due candidati soltanto e che hanno un forte engagment con i temi politici. Sono stati creati due dataframe di cui sono state calcolate i seguenti parametri:

### 

* + 36,699 Tweets nel dataframe relativo a Obama:
    - N°. 15,879 ( 43%) tweets che retwettano;
    - N°. 2,453 ( 6.6%) tweets che replicano;
    - N°. 18,367 ( 50%) senza replica o condivisione;
    - 996,710 numero di RT totale[[11]](#footnote-11) dei tweets source;
  + 6,660 Tweets nel dataframe relativo a Romney:
    - N°. 2,943 ( 44%) tweets che retwettano;
    - N°. 453 ( 6.8%) tweets che replicano;
    - N°. 3,264 ( 49%) tweets senza replica o condivisione;
    - 221,170 numero di RT totale dei tweets source.

Le proporzioni tra le tipologie di tweets sono praticamente le stesse.

**SENTIMENT ANALYSIS**

Sono state svolte operazioni di pre-process dei testi dei tweets: eliminazione di link, delle menzioni, degli hashtags, della punteggiatura e riduzione in caratteri minuscoli. Ciò è funzionale al poter effettuare opportunamente l’analisi del sentiment e delle frequenze.

È stato calcolato il sentiment per i tweets secondari che presentano hashtags esclusivamente riguardanti Obama o Romney. Nello specifico, per questi due sottoinsiemi, sono stati classificati i tweet in base a tre livelli di sentiment: positivo, negativo e neutro.

Per l’analisi del sentiment sono state utilizzate una lista di parole negative ed una lista di parole positive. Per le operazioni di calcolo, è stata applicata la seguente logica: se nel testo compaiono più parole di una delle due liste, viene assegnato al tweet il sentiment corrispondente a quella lista; il sentiment neutro viene invece assegnato in caso di parità tra il conteggio degli elementi appartenenti alle due liste.

|  |  |  |
| --- | --- | --- |
| **Sentiment** | Obama | Romney |
| **# Tweet neutri** | 17903 | 3001 |
| **# Tweet negativi** | 11798 | 2375 |
| **# Tweet positivi** | 6998 | 1284 |

Infine, è stato calcolato quanti tra i tweets secondari con sentiment positivo di ciascun candidato sono retweets e quanti repliche.

|  |  |  |
| --- | --- | --- |
| **Sentiment positivo** | Obama | Romney |
| **# Tweet di replica** | 118 | 42 |
| **# Tweet ricondivisi** | 2992 | 531 |

**ANALISI E VISUALIZZAZIONE DELLE FREQUENZE**

È stato realizzato uno script (cfr. “Script2.py”) che va a calcolare le frequenze delle parole nei testi ripuliti e diversi istogrammi e wordclouds.

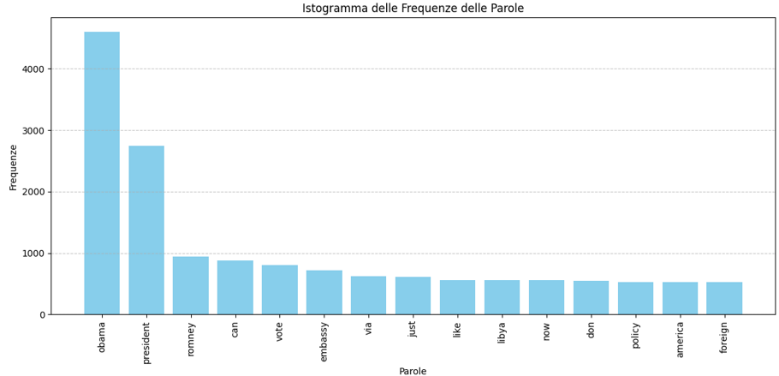
Considerando i dataframe che riguardano esclusivamente i due candidati, sono stati elaborati i seguenti grafici:

Figura 2– Istogramma delle frequenze assolute delle parole nei testi (ripuliti) dei tweets per il Dataframe su Obama, US & Canada time\_zone. La somma totale delle frequenze è: 15,712.

A graph with blue squares

Description automatically generated

Figura 3– Istogramma delle frequenze assolute delle parole nei testi (ripuliti) dei tweets per il Dataframe su Romney, US & Canada time\_zone. La somma totale delle frequenze è: 5,949.

Da una prima analisi della frequenza delle parole nei dataframe dei due candidati (pre-sentiment, con testi ripuliti), osserviamo che:

* I candidati sono reciprocamente citati nei tweet che hanno hashtags relativi all’altro candidato. In generale “obama” è la parola con più alta frequenza (2142 in Figura 3 e 4603 in Figura 2[[12]](#footnote-12)). Anche nei tweets riguardanti Romney viene citato molto di più Obama che egli stesso, in percentuale Obama viene citato di più nei tweets relativi a Romney (nota 11), quindi la portata mediatica di Obama è molto grande;
* La presenza del termine “can” che pone tipicamente enfasi su un messaggio positivo e affermativo e richiama strategicamente la tradizione liberal-progressista statunitense è molto più presente nel dataframe Obama (879, 5.5%) rispetto a quello di Romney (125, 2.1%);
* Se in entrambi i dataframe “president” si posiziona come secondo termine in ordine di frequenza, circa il riferimento alle elezioni per Romney si parla di “election” mentre compare il termine “vote” per il dataframe Obama. Per quest’’ultimo, ciò suggerisce la presenza di un linguaggio meno formale ed un utilizzo già consapevole della comunicazione sui social come forma di appello e slogan;
* Riferimenti alla politica estera, e in particolar modo alla Libia, a poco meno di un anno dall’uccisione di Gheddafi, sono presenti per entrambi i candidati.
* Rileva infine, la frequenza del temine “attack” riportata nei soli tweet riguardanti Romney.

Procedendo nell’analisi, passiamo al sentiment dei due candidati, i seguenti wordclouds restituiscono i termini più ricorrenti per i tweet di sentiment negativo per Romney ed Obama.

|  |  |
| --- | --- |
| **Wordclouds con Sentiment negativo – Testo ripulito** | |
| **Dataframe Obama** | **Dataframe Romney** |
|  |  |

Si osserva che:

* Obama è fortemente presente sulla scena di entrambi;
* Il linguaggio aggressivo e di odio è marcatamente più esplicito nel dataframe di Obama, presidente uscente;
* La presenza di riferimenti alla politica estera sembrano suggerire temi di apprensione comuni per i due candidati, con una più elevata ricorrenza di “Libya” e “foreign” per Romney, e viceversa un riferimento più frequente all’Islam per Obama;

A close up of words

Description automatically generatedInfine, soffermando l’analisi sul sentiment positivo rilevato per il candidato vincente, tramite il seguente wordcloud osserviamo che:

Figura 4 - Wordcloud dataframe Obama, Sentiment positivo, US & Canada time zone

* Esplicito riferimento al gradimento e al supporto nei termini di “like”, “support”, “trust”, “won” che suggeriscono la presenza una tribe particolarmente engaged. La ricorrenza di attributi di carattere positivo come “good”, “strongly”, “fascinating” sembra sottolineare la volontà descrittiva e persuasiva dei suoi elettori;
* Il riferimento a valori e ideali comuni come “justice”, “right”, “love”, “freedom” sui quali potrebbe poggiare il reciproco riconoscimento della tribe;
* I termini “won”, “better”, “stronger” sottolineano un livello di fiducia elevato risposto nel candidato;
* I riferimenti alla politica estera di sentiment positivo comprendono “Netanyau”, “embassy”, “intelligence”, “crises” suggerendo una percezione positiva dei risultati raggiunti durante il precedente mandato.

**Analisi delle reti**

Nell’ambito dell’analisi di rete abbiamo considerato la rete di retweet tra gli utenti, nello specifico:

1. Gli utenti rappresentano i nodi, 493,904 nodi in totale;
2. La relazione di retweet rappresenta un arco, 444,457 archi in totale.

In questo caso abbiamo considerato tutto il data set a disposizione, eliminando solo i duplicati.

Per l’analisi di rete è stato ritenuto opportuno e rilevante calcolare le seguenti misure di centralità:

* Degree centrality;
* Closeness centrality.

La degree centrality va a contare il numero di link di cui un nodo è un estremo (quando il grafo considerato è non diretto). Un nodo che ha un’alta centralità di questo tipo può facilmente diffondere un’informazione all’interno della rete. Tale aspetto risulta d’interesse per una strategia comunicativa efficace applicata ai social media. Rendere il profilo del candidato un nodo di questo tipo, o altresì coinvolgere ed ingaggiare altri nodi con elevato valore di degree, può contribuire significativamente al successo di una campagna mediatica.

Piu specificatamente, nel contesto di una campagna elettorale osservare ed applicare questo tipo di metriche nei social media può essere cruciale per il monitoraggio e tempestivo orientamento delle strategie comunicative dei candidati.

Nell’ambito del dataset osservato, e della rispettiva finestra temporale, le analisi di centralità sembrano rispecchiare ed anticipare i risultati delle elezioni. In particolare, la classifica dei primi dieci nodi della rete è:

|  |  |
| --- | --- |
| Utente | Degree |
| Barack Obama | 14358 |
| Shit Nobody Says | 9662 |
| Niall Horan | 8134 |
| Donald J. Trump | 7212 |
| THE X FACTOR (USA) | 6700 |
| Alfredo Flores | 2964 |
| Obama 2012 | 2613 |
| Kevin Eder | 2039 |
| Michelle Malkin | 1981 |
| Laura Ingraham | 1661. |

Abbiamo considerato i valori percentuali per avere una visione indipendente dalla dimensione del data set considerato.

|  |  |  |  |
| --- | --- | --- | --- |
| Utente | Degree | Degree (%) |  |
| Barack Obama | 14358 | 3.23 |  |
| Shit Nobody Says | 9662 | 2.17 |  |
| Niall Horan | 8134 | 1.83 |  |
| Donald J. Trump | 7212 | 1.62 |  |
| THE X FACTOR (USA) | 6700 | 1.51 |  |
| Alfredo Flores | 2964 | 0.67 |  |
| Obama 2012 | 2613 | 0.59 |  |
| Kevin Eder | 2039 | 0.46 |  |
| Michelle Malkin | 1981 | 0.45 |  |
| Laura Ingraham | 1661 | 0.37 |  |

Come si può osservare il primo nodo risulta essere lo stesso candidato Obama, che nel 2012 sembra aver prontamente intuito il potenziale di Twitter, facendone un protagonista nella sua strategia comunicativa.

Infine, la centralità di prossimità (closeness) è una misura utilizzata nell'analisi delle reti per determinare l'importanza relativa di un nodo all'interno di una rete. I nodi con alta centralità di prossimità hanno distanze medie più brevi rispetto a tutti gli altri nodi. Questi nodi possono interagire rapidamente con tutti gli altri, rendendoli influenti per la diffusione delle informazioni o per il controllo della comunicazione all'interno della rete.

La centralità di prossimità di un nodo u è definita come il reciproco della somma delle distanze dei percorsi più brevi da u a tutti gli altri nodi nella rete. Matematicamente, può essere espressa come:

dove:

* N è il numero totale di nodi nella rete.
* è la distanza del percorso più breve tra i nodi u e v.

L’algoritmo è stato implementato, tuttavia, a causa della limitatezza delle risorse computazionali a disposizione, non è stato possibile calcolare questo tipo di centralità sulla rete considerata, essendo troppo grande.

**Conclusioni**

Nel corso delle analisi descritte, abbiamo osservato la rilevanza delle metodologie e metriche applicate, intuendone il potenziale contributo informativo in una compagna mediatica elettorale consapevole e strategica. Rispetto al dataset, in assenza di metadata e nei limiti dei dati osservabili, siamo consapevoli delle possibili distorsioni nelle stime effettuate, ma con l’intenzione di identificare spunti utili per la definizione di una campagna efficace sui social media, le conclusioni dell’analisi dei tweet in tema elezioni USA del 2012 sono:

* Nel 2012, la strategia comunicativa su Twitter di Obama risulta vincente rispetto al suo avversario: l’alto degree, il linguaggio, il sentiment analizzato e ad ultimo l’elevato numero di RT rispetto a Romney (996,710 vs 221,170, rispettivamente), suggeriscono la maturità della sua campagna mediatica e la consapevolezza del potenziale del social media;
* La politica estera è un argomento fortemente presente e di rilievo nella campagna elettorale statunitense. Una visione di politica estera del candidato e i messaggi riguardo a eventi di attualità estera si prestano molto bene ad essere veicolati sottoforma di tweets;
* Dallo studio delle reti abbiamo classificato i nodi più centrali nello specifico contesto temporale del dataset. Nello specifico, osserviamo che nel 2012 Donald D. Trump risulta essere appena sotto il podio per il degree nella rete analizzata. Questa tipologia di analisi, quindi, risulta particolarmente strategica non solo per identificare i nodi “push factor” da potenzialmente ingaggiare per la crescita della tribe del candidato, ma anche per monitorare da una nuova angolazione la scena politica, intercettandone i potenziali futuri protagonisti;
* In considerazione delle analisi sopra descritte e della corrispondenza degli esiti con i risultati elettorali realmente registrati, è in definitiva enorme il potenziale strategico di un monitoraggio continuativo dei social media, in termini di contenuti, sentiment e rete, per avere il controllo e la gestione ottimale della campagna elettorale.

Riconoscendo la significatività dell’applicazione di queste analisi allo specifico contesto storico statunitense, riteniamo tali conclusioni valide per qualsiasi contesto elettorale odierno relativamente ai social media.

1. Email universitaria: gisella.coppola@students.uniroma2.eu [↑](#footnote-ref-1)
2. Email universitaria: salvatore.dellorusso@students.uniroma2.eu [↑](#footnote-ref-2)
3. Il termine "big data" si riferisce a dati informatici così grandi, veloci e/o complessi, difficili o impossibili da elaborare con i metodi tradizionali. Tale definizione non è statica, nel tempo ha subito un’evoluzione. [↑](#footnote-ref-3)
4. Il deepfake è una tecnica per la sintesi dell'[immagine](https://it.wikipedia.org/wiki/Immagine) umana basata sull'[intelligenza artificiale](https://it.wikipedia.org/wiki/Intelligenza_artificiale), usata per combinare e [sovrapporre](https://it.wikipedia.org/wiki/Sovrimpressione) immagini e [video](https://it.wikipedia.org/wiki/Video) esistenti con video o immagini originali, tramite una tecnica di [apprendimento automatico](https://it.wikipedia.org/wiki/Apprendimento_automatico), nota come [rete antagonista generativa](https://it.wikipedia.org/wiki/Rete_antagonista_generativa). [↑](#footnote-ref-4)
5. In allegato a questo documento ci sono tre codici python che sono numerati secondo l’ordine cronologico dell’analisi svolta (“Script1.py” – Analisi esplorativa e descrittiva, “Script2.py” Istogrammi di frequenza e wordclouds, “Script3.py” Analisi di rete). [↑](#footnote-ref-5)
6. Le percentuali sono calcolate rispetto ai dati senza duplicati (vedi seguito). [↑](#footnote-ref-6)
7. Source tweet. [↑](#footnote-ref-7)
8. Siamo consapevoli di aver a che fare con tweets non geolocalizzati; tuttavia, abbiamo provato a cercare di localizzare, grossomodo, geograficamente i tweets americani, dato che sono quelli scritti dagli elettori direttamente interessati e quindi che meglio si prestano agli scopi di analisi. [↑](#footnote-ref-8)
9. Si precisa che non sempre sono stati usati tutti i filtraggi elencati contemporaneamente nei codici pyspark sviluppati. [↑](#footnote-ref-9)
10. Per facilitare la visualizzazione delle barre dell’istogramma che corrispondono a lunghezze grandi. [↑](#footnote-ref-10)
11. Abbiamo assunto che i tweets che venivano retweettati fossero inerenti agli stessi argomenti dei tweets retweettanti, come ragionevolmente ci si può aspettare. Numero totale dei tweets retweettati da quelli presenti nel DataFrame Obama: 17,784, Romney: 3,581. [↑](#footnote-ref-11)
12. Le corrispondenti frequenze relative sono: 4603/15712 36% per il dataframe Obama e 2142/5949 29% per il dataframe Romney. [↑](#footnote-ref-12)