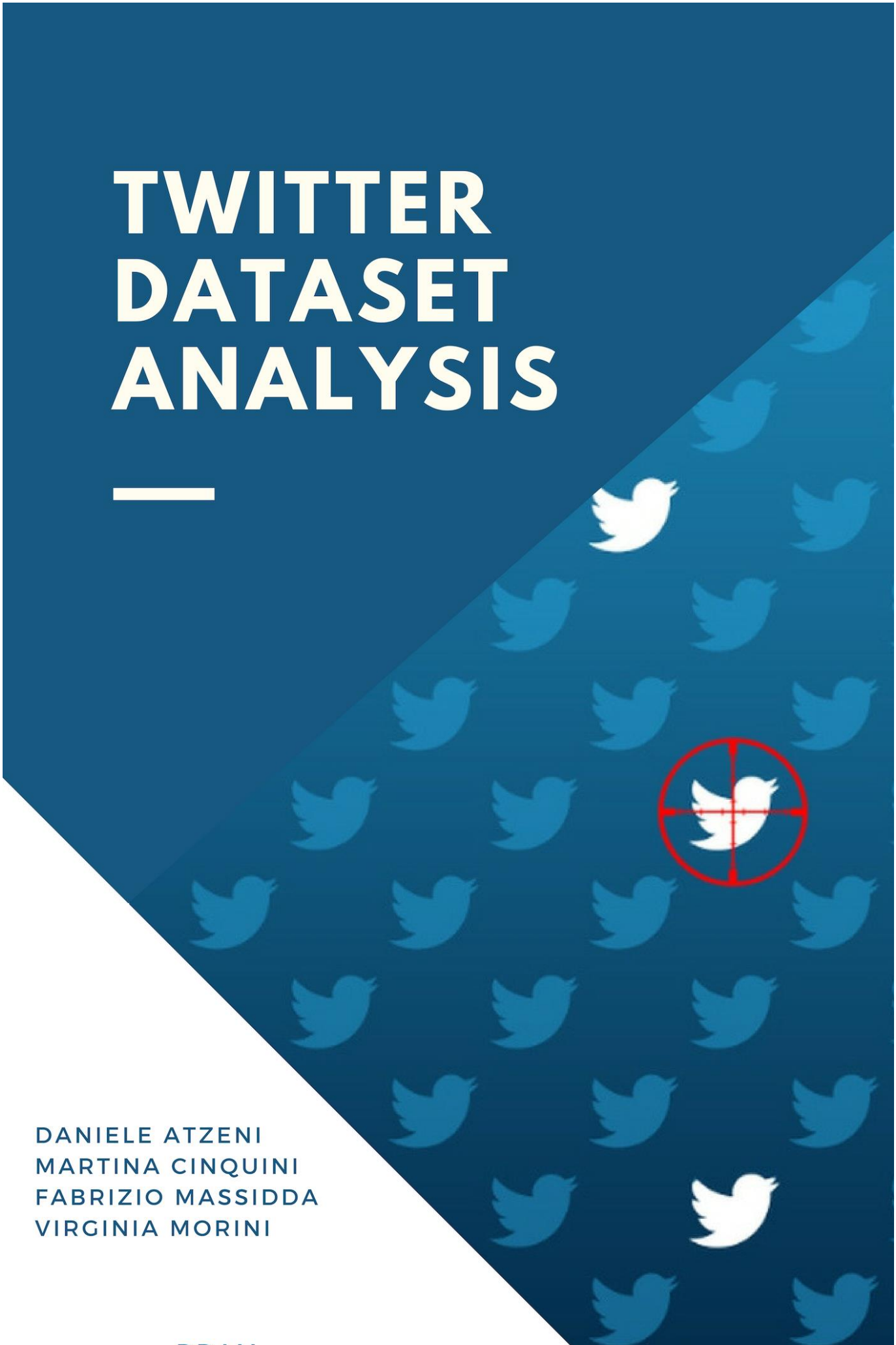
TWITTER DATASET ANALYSIS



DANIELE ATZENI
MARTINA CINQUINI
FABRIZIO MASSIDDA
VIRGINIA MORINI

Business Understanding

Il dataset preso in considerazione per l'analisi contiene oltre 6994108 tweet geolocalizzati in 84874 città del mondo. I tweet, realizzati da oltre 2 milioni di utenti, coprono un periodo compreso tra il 21 e il 29 novembre 2015. Inoltre, i dati forniti dalle API di Twitter restituiscono i tweet distribuiti in 9 file JSON, uno per ogni giorno preso in considerazione.

Nell'analisi e nella formulazione degli obiettivi del progetto è stato necessario tener presente che il seguente dataset, oltre a coprire un periodo ristretto, presenta un *bias* piuttosto alto, in quanto non offre un quadro generale di tutti i tweet realmente postati in una determinata area; sono disponibili infatti solo i tweet degli utenti che hanno volontariamente deciso di condividere la loro posizione.

Il progetto nasce quindi con l'obiettivo di effettuare una profilazione degli utenti che utilizzano Twitter analizzandone gli spostamenti, le attitudini e il linguaggio. Per quanto riguarda gli spostamenti dell'utente l'analisi mira a:

1. Tracciare gli spostamenti degli utenti nel periodo di tempo analizzato
2. Categorizzare gli utenti in base alla tipologia/numero di spostamenti
3. Effettuare una sentiment analysis sulle suddette categorie con l'obiettivo di capire se il modo di esprimersi di 'utente sedentario' e di un 'utente dinamico' cambiano o meno.

Invece per quanto riguarda il singolo tweet l'indagine ha l'intenzione di:

1. Individuare i trending topics di ogni giorno in base alla popolarità di un tweet
2. Studiare gli hashtag in base alla loro frequenza e alla loro varianza con il fine di determinare eventi quotidiani di risonanza mondiale

Data Understanding

Il dataset si compone di 27 attributi di primo livello descritti in Tabella 1.

Tabella 1. Descrizione e analisi degli attributi di primo livello

Attributo	Tipologia	Descrizione
created_at	Stringa	Indica la data e l'ora di creazione del tweet.
id	Intero	Identificatore univoco del tweet.
id_str	Stringa	Identificatore univoco del tweet in formato stringa.
text	Stringa	Testo presente nel tweet.
source	Stringa	Indica la tipologia di dispositivo da cui è stato postato il tweet.
truncated	Booleano	Indica se il testo del tweet è presentato in versione integrale o meno.
in_reply_to_status_id	Intero	Se il tweet è una retweet, l'id indica l'id del tweet originale.
in_reply_to_status_id_str	Stringa	Se il tweet è una retweet, l'id indica l'id del tweet originale (in formato stringa).
in_reply_to_user_id	Intero	Se il tweet è una retweet, l'id indica l'id dell'autore del tweet originale.
in_reply_to_user_id_str	Stringa	Se il tweet è una retweet, l'id indica l'id dell'autore del tweet originale (in formato stringa).
in_reply_to_screen_name	Stringa	Se il tweet è una retweet, viene indicato il nome dell'autore del tweet originale.
user	Struct	Indica l'utente che ha postato il tweet. La struct contiene numerose informazioni come l'id dell'utente, il nome, l'url del profilo, il numero di follower e la lingua utilizzata.
coordinates	Struct	Indica la geolocalizzazione del tweet (Indicata dall'applicazione utilizzata dall'utente). La struct contiene le coordinate di latitudine e longitudine.
place	Struct	Indica la posizione inserita dall'utente all'interno del tweet. Può non riferirsi al luogo effettivo in cui si trova. Contiene le coordinate, lo stato, la città, l'id.
quoted_status_id	Intero	Se il tweet è un 'quote tweet', l'id indica l'id del tweet citato.
quoted_status_id_str	Stringa	Se il tweet è un 'quote tweet', l'id indica l'id del tweet citato (in formato stringa).
is_quote_status	Booleano	Indica se il tweet è un 'quote tweet' oppure no.
quoted_status	Struct	Struct che contiene il tweet originale che è stato citato.
retweet_count	Intero	Indica il numero di retweet del tweet.
favorite_count	Intero	Indica il numero di like ricevuti dal tweet.
entities	Struct	Entità eliminate dal testo del tweet. È una struct contenente eventuali hastag, link, foto, simboli inclusi nel tweet.
extended_entities	Struct	Quando le entità inserite nel tweet sono più di una vengono inserite in questa struct.
favorited	Booleano	Indica se il tweet ha ricevuto like o meno.
retweeted	Booleano	Indica se il tweet è stato retwittato o meno.
possibly_sensitive	Booleano	Indica se il tweet contiene un link.
lang	Stringa	Indica la lingua utilizzata nel testo del tweet.
contributors	Stringa	Se presenti, indica i co-autori di un tweet.

In questa fase, con il fine di comprendere se i dati a disposizione risultassero utili per gli obiettivi preposti si è deciso, a causa della grande quantità di dati, di analizzare un singolo file JSON composto da 856457 tweet.

Per quanto riguarda la geolocalizzazione dell'utente è necessario sottolineare che nel dataset iniziale sono presenti due attributi che descrivono, in termini di coordinate, 'country' e 'city', la localizzazione del tweet: la feature *coordinates* indica la reale geolocalizzazione del tweet rilasciata dall'applicazione, mentre *place* indica il luogo inserito manualmente dall'utente, identificato da quattro coppie di coordinate. Analizzando la distribuzione dei missing values (Figura 1) si è notato che solo il 9,5% dei tweet ha la feature *coordinates* diversa da null, mentre place ha solo 1,2% di valori nulli.

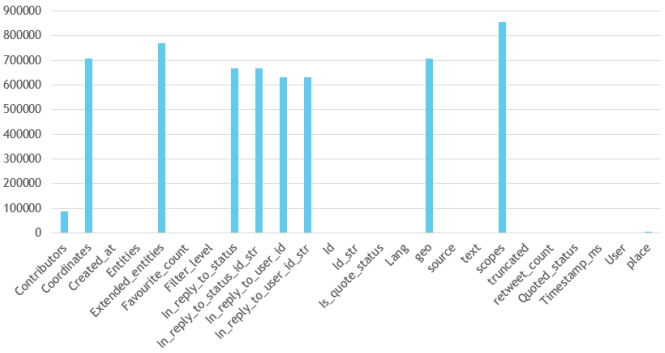


Figura 1. Distribuzione missing values

Figura 2. Localizzazione dei tweet

Per questo motivo è stato controllato che le coordinate di *coordinates* fossero comprese tra i valori minimi e massimi assunti dalle coordinate dei punti che descrivono l'attributo *place*. Avendo ottenuto un match del 99,8% si è deciso di prendere in considerazione solamente *place*. Per verificare che i tweet provenissero da diverse aree geografiche si sono plottate le coordinate di place nella seguente Bubble Map (Figura 2).

Per quanto riguarda l'analisi dei tweet, nello specifico l'individuazione dei trending topics in base alla popolarità del tweet, si è concluso che non è possibile portare a termine il task in quanto i valori delle features *retweet_count* e *favourites_count* sono sempre settati a 0. Risulta invece possibile studiare la frequenza e la varianza degli hashtag (vedi sezione 'Analisi dei cluster e conclusioni').

Data Preparation

In primis, al fine di poter svolgere i suddetti obiettivi si sono uniti i 9 dataset in un unico globale. Per effettuare la profilazione degli utenti si è deciso di creare due nuovi dataset: il primo contenente le informazioni relative agli spostamenti di ogni utente, il secondo contenente dati riguardanti le attitudini degli utenti su Twitter. Entrambi i dataset presentati di seguito descrivono il comportamento di oltre 2 milioni di utenti.

Il dataset 'Spostamenti utente'

'Spostamenti utente' è formato esclusivamente da feature create a partire dall'attributo *place*. Per la creazione degli attributi relativi agli spostamenti è stato calcolato il baricentro di ogni place, cioè il punto del piano le cui coordinate sono la media delle coordinate dei quattro punti che rappresentano l'attributo. Successivamente, per ogni utente, è stata creata una lista dei baricentri dei posti visitati, ordinata in ordine cronologico tramite l'attributo *created_at*. A partire da questa lista sono stati creati, per ogni utente, gli attributi in Tabella 2.

Tabella 2. Feature creation 'Spostamenti utente'

Attributo	Tipologia	Descrizione
Distanza totale percorsa	float	Rappresenta la somma delle distanze euclidee tra un punto della lista e il suo successivo
Numero di spostamenti	intero	Indica il numero di volte in cui un elemento della lista è diverso dal successivo
Numero di posti visitati	intero	Rappresenta la lunghezza della lista dopo aver rimosso i duplicati

Il dataset 'Tipologia utente'

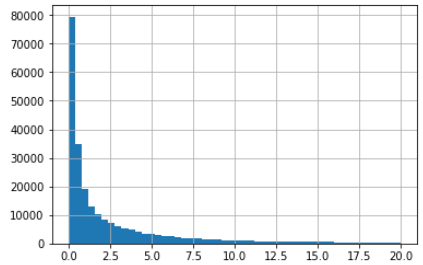
Il secondo dataset è stato creato utilizzando 7 attributi già presenti e 3 attributi creati a partire dalle caratteristiche dei tweet (indicati in rosso in Tabella 3). Per quanto riguarda gli attributi selezionati, dal momento che sono in gran parte variabili nel tempo, si è deciso di raggruppare per user id, quindi selezionarne il massimo.

Tabella 3. Feature creation e Feature selection 'Tipologia utente'

Attributo	Tipologia	Descrizione
friends_count	fintero	Indica il numero di following dell'utente
follower_count	intero	Indica il numero di follower dell'utente
favourites_count	intero	Rappresenta il numero di tweet a cui l'utente ha messo mi piace da quando ha un profilo twitter
listed_count	intero	Indica il numero di liste pubbliche di cui l'utente è membro
statuses_count	intero	numero di tweet emessi dall'utente da quando ha un profilo twitter
verified	booleano	Rappresenta il badge blu di account verificato su Twitter che permette alle persone di sapere se un account di interesse pubblico è autentico
default_profile_image	booleano	Se True vuol dire che l'utente ha l'immagine del profilo di default
avg_media	float	Numero medio di foto per tweet
avg_hashtag	floatf	Numero medio di hashtag per tweet
avg_url	loat	Numero medio di link per tweet

Data Modeling: Clustering

La fase di clustering è stata effettuata parallelamente sia sul dataset *Spostamenti utente* sia su *Tipologia utente* utilizzando l'algoritmo K-Means. Il primo dataset è stato inizialmente suddiviso in due parti, la prima composta dagli utenti che non hanno effettuato alcuno spostamento (1.634.771 record), la seconda formata dai restanti utenti che si sono effettivamente spostati. Si è ritenuto opportuno applicare l'algoritmo di clustering esclusivamente sulla seconda parte. Prima di applicare il K-Means sono state studiate le distribuzioni dei singoli attributi e gli scatter plot bidimensionali.



Dal momento che ognuno dei 3 attributi presenta una distribuzione esponenziale, si è deciso di pre-processare i dati calcolandone il logaritmo e solo successivamente procedere con la ricerca del numero ideale di cluster. Quest'ultima è stata effettuata avvalendosi dell'aiuto del grafico dell'SSE in funzione del numero di cluster ed ha portato alla scelta del valore 3.

Figura 3. Distribuzione distanza totale percorsa

Per il secondo dataset sono stati effettuati due metodi di pre-processing dei dati diversi, di cui poi si è scelto quello che portava a risultati più convincenti. In entrambi i casi gli attributi booleani sono stati mappati in 1 per i record che assumono valore True e 0 per i record con valore False.
Nel primo metodo si è utilizzata la min-max Normalization, si è poi trovato il numero ideale di cluster utilizzando nuovamente il grafico dell'SSE. Tuttavia, a seguito di un'analisi più approfondita dei centroidi, si è capito che così facendo si sarebbe data troppa importanza agli attributi booleani e agli attributi discreti che assumono pochi valori distinti, rendendo quasi insignificanti gli attributi *friends_count*, *follower_count* e

favourites_count. Per questo motivo si è deciso di effettuare un altro tipo di normalizzazione, la standardizzazione. Infine, si è proceduto nuovamente al calcolo del numero ideale di cluster, che ha dato come risultato il valore 6. L'analisi dei risultati ottenuti verrà riportata nel paragrafo successivo.

Analisi dei cluster e conclusioni

Il clustering effettuato sul dataset 'spostamenti utente', data l'alta correlazione tra gli attributi, ha portato esclusivamente ad una stratificazione degli utenti, producendo quattro cluster etichettabili come:

- Utenti sedentari
- Utenti mediamente mobili
- Utenti dinamici
- Utenti super-dinamici

Di seguito vengono analizzati i cluster risultanti dal dataset 'tipologia utenti', assegnando a ciascuno di essi un'ipotetica etichetta in base alle peculiarità che li contraddistinguono.

Utenti medi
CLUSTER 0-1-2: Cluster composti rispettivamente da 528635, 213036 e 1147294 utenti, che possono essere classificati come utenti medi di twitter. Sono caratterizzati da un numero di followers e following nella media. Le uniche differenze significative tra questi cluster si riscontrano nel numero di url per tweet, molto più elevato nel cluster 0, e nel numero di hashtag per tweet, molto più elevato nel cluster 2, con una media di 4.5. Al cluster 2 inoltre appartengono presumibilmente gli utenti meno attivi, come si può notare dal numero basso di *statuses*.

Popolari
Cluster composto da 5492 utenti che possono essere considerati medio/popolari in quanto il loro numero medio di followers si aggira intorno ai 90000 mentre il numero di following intorno ai 2500. Questi utenti sono inoltre al 99% verificati.

Utenti automatizzati
Cluster composto da 118573 utenti, di cui probabilmente fanno parte bot e pagine che si occupano di pubblicità. Il numero di following e followers per questi utenti è medio/alto, mentre è estremamente elevato il numero di *favourites count* ed il numero di *statuses*.

Celebrità
Cluster di dimensioni ridotte (53) di cui fanno parte le celebrità, con un numero di following medio/alto, intorno ai 60 mila, ed un numero molto alto di followers, intorno ai 4,5 milioni. Questi utenti hanno un numero molto alto di *listed*, e sono per la maggior parte verificati.

Incrociando i dati risultanti dalle due clusterizzazioni si può notare che:

1. La percentuale di utenti del cluster 2 (cluster comprendente gli utenti poco attivi), scende dal 60% nel cluster degli utenti sedentari fino al 37.3% nel cluster degli utenti super-dinamici.
2. La percentuale di utenti popolari, di cui fanno parte solo lo 0.2% degli utenti del dataset, sale fino a 0.8% nel cluster degli utenti super-dinamici.
3. Gli utenti automatizzati compongono il 13.2% del cluster degli utenti super-dinamici e il 10.7% del cluster degli utenti dinamici, mentre solamente il 4.8% del cluster degli utenti sedentari.

Tutte queste considerazioni sono coerenti con la categorizzazione dei cluster effettuata in precedenza.

Successivamente è stata effettuata una Sentiment Analysis sui testi dei tweet, sfruttando due indicatori: *polarity* e *subjectivity*. Questi assegnano al testo un valore compreso tra -1 e 1, positivo se il testo è associabile a sentimenti positivi e a considerazioni soggettive, negativo altrimenti. I risultati ottenuti sono stati incrociati con le clusterizzazioni ma non hanno evidenziato differenze apprezzabili.

Infine sono stati analizzati gli hashtag sia per frequenza sia per varianza. I risultati della prima analisi sono mostrati nella Word Cloud in Figura 4; la seconda analisi, il cui scopo era quello di identificare eventi occasionali associabili ad una determinata area geografica in modo da poter individuare variazioni negli spostamenti degli utenti ad essi relativi, non ha portato ai risultati sperati.

Figura 4. Wordcloud degli hashtag più frequenti