La Comunicazione - Note, Recensioni e Notizie n. 64

Anno 2021

Una piattaforma di Big Data per la stima del numero di utenti attivi e l'analisi della qualità dell'informazione dei social network

A Big Data Platform for estimating the number of active users and analysing the quality of information of Social Networks

Giambattista Amati♦, Simone Angelini♦, Antonio Cruciani□, Gianmarco Fusco●, Giancarlo Gaudino●, Daniele Pasquini○, Paola Vocca○

♦ Fondazione Ugo Bordoni, Roma

□ Istituto di Scienze del Gran Sasso, L'Aquila

● DGTCSI-ISCTI Ministero dello Sviluppo Economico

○ Università della Tuscia, Viterbo

Sommario

Il primo obiettivo di questo lavoro di ricerca è fornire una stima accurata del numero di utenti italiani attivi su Twitter lo scorso anno. La stima è stata possibile filtrando in modo opportuno tutto il flusso in lingua italiana di Twitter. Con il filtro utilizzato, abbiamo stimato che l'86.60% dei tweet selezionati risultano essere effettivamente di lingua italiana. Grazie a questa stima di accuratezza, siamo stati anche in grado di fornire il numero effettivo di utenti attivi (AU) di Twitter. Il secondo risultato è la definizione di una metodologia per effettuare il *clustering* massivo dei testi, veloce, facilmente scalabile e applicabile a Twitter. La tecnica proposta è basata sugli algoritmi di *community detection* CoDA e Louvain, applicati al grafo pesato degli *hashtag*. Una volta raggruppati gli *hashtag* in *cluster (comunità)*, sia i cluster più numerosi che gli hashtag sono stati associati ad argomenti di interesse generale, come lo sport, la politica, la salute ecc. Mediante tale metodologia è stato così possibile analizzare e fornire statistiche significative sugli argomenti più popolari toccati in Twitter lo scorso anno.

Abstract

We first introduce an accurate estimate of the number of users that were active on Twitter last year. This estimate was made possible by appropriately filtering the entire Twitter stream in Italian. We estimated that 86.60% of the selected tweets were actually Italian. Thanks to this accurate estimate, we were also able to provide Twitter's actual number of active users

(AU). Then, we introduce a methodology to perform massive clustering of texts, that is fast, easily scalable and applicable to Twitter. It relies on the CoDA and Louvain community detection algorithms for the weighted hashtag graph. Once the hashtags were grouped into clusters, both the larger clusters and the hashtags were associated with topics of general interest, such as sports, politics, health, etc. Using this methodology, it was thus possible to provide meaningful statistics on the most significant topics touched on Twitter last year.

1 - Introduzione

In questo lavoro, verrà presentata l'infrastruttura *hardware* e *software* del laboratorio di *Big Data* di ISCOM e FUB, grazie alla quale è stato possibile introdurre una nuova metodologia per selezionare e analizzare l'intero flusso di *Twitter* in lingua italiana (il *Firehose* italiano), relativamente ai dati nel periodo dal 25 Maggio 2020 al 25 Aprile 2021. In particolare, abbiamo stimato il numero di utenti attivi italiani di *Twitter*, aggregato i loro tweet in argomenti di interesse, e fornito delle statistiche accurate relative all'incidenza di questi argomenti nel flusso completo italiano. *Twitter* definisce due tipologie di utenti attivi: ***mDAU*** (*Monetary Daily Active Users*) e il ***MAU*** (*Monthly Active Users*). *mDAU*, che è una misura introdotta dall'azienda dal 2019, fornisce il numero di utenti che in un singolo giorno siano stati esposti ad almeno una pubblicità di *Twitter* eseguendo il *login* tramite l'applicazione ufficiale, o accedendo a applicazioni di terze parti. Il *MAU* che non è più in uso, descrive invece gli utenti attivi registrati che hanno fatto l'accesso su *Twitter*, tramite applicazione *web* o mobile, nei 30 giorni precedenti. Esistono diverse stime sul numero di utenti attivi su *Twitter*, spesso però con valori discordanti. Il valore ufficiale *mDAU* fornito da *Twitter* è di 2,8 milioni di utenti attivi italiani[1]. Anche se non è possibile conoscere il numero esatto di utenti attivi, è comunque possibile in linea di principio derivare il sottoinsieme degli utenti italiani, che hanno scritto un *tweet* (o eseguito un *retweet*) almeno una volta in un determinare periodo temporale. Per ottenere

[1] Vedi per esempio il report Digital 2021 https://datareportal.com/reports/digital- 2021- italy

tale stima è necessario però filtrare tutto il flusso italiano, ed effettuare una stima dei falsi positivi e dei falsi negativi del filtro. Per catturare tutto il flusso italiano, e fornire delle statistiche più precise sul numero di utenti attivi, si è deciso di attivare un filtro testuale contenente le parole funzionali più frequenti (*stop-words*) della lingua italiana [2-5]. Per non eccedere i limiti imposti da *Twitter*, si è inoltre attivato un secondo filtro di controllo fornito da *Twitter* che assegna una lingua al *tweet*. Ovviamente, questo filtro linguistico nativo di *Twitter* non è preciso e quindi è stato necessario valutarne la sua efficacia. In questo modo sono stati recuperati ben 1,8 milioni di tweet al giorno.

A valle di ciò, in questo lavoro si propone inoltre un approccio altamente scalabile per effettuare il *clustering* dei *tweet* e determinare così gli argomenti principali coperti dai *tweet* (*Topic Modelling*). Lo stato dell'arte propone diversi algoritmi per eseguire tali task: LSH (*Locality-Sensitive Hashing*) [8-12], LDA (*Latent Dirichlet Allocation*) [13], *k-Means*, *k-Medians* [11].

Nonostante *LSH* sia un algoritmo lineare, non è molto efficace su testi con numero limitato di caratteri come i *tweet*, poiché richiede di impostare una soglia bassa di similarità per evitare che le classi contengano un numero ristretto di *tweet*. D'altro canto, *LDA* è intrattabile poiché è un algoritmo *NP-hard* in presenza di un numero elevato di *feature* [13]. Infine, gli algoritmi che richiedono invece di impostare in anticipo un numero ideale di *cluster* secondo i quali suddividere i testi, presentano problemi intrinsechi per eseguire un *clustering* massivo efficace, in quanto, in generale, il numero ottimale di *cluster* non è noto a priori.

La tecnica che abbiamo utilizzato, invece, è basata su algoritmi di *community detection* sui grafi e, nel caso specifico, sul grafo degli *hashtag*. Le co-occorrenze degli *hashtag* di ogni *tweet* definiscono un grafo non diretto pesato: esiste un arco non diretto tra due *hashtag* se occorrono nello stesso *tweet*. Di conseguenza, ogni arco è pesato con il numero totale di co-occorrenze tra tutti i *tweet*. Il grafo degli *hashtag* è stato già utilizzato in [15] per l'analisi del *sentiment*, mentre in [9] è stato utilizzato l'algoritmo di *Louvain* sul grafo delle parole, ma relativamente a dimensioni di due ordini di grandezza inferiori al *Firehose* italiano. Il *topic*

modelling di [14], tenta l'aggregazione degli *hashtag* utilizzando una tecnica di *pooling*, mentre l'algoritmo di *community detection* di [6] usa il grafo dei termini.

In questo articolo, per eseguire il *Clustering* del grafo degli *hashtag* si sono applicati gli algoritmi di complessità temporale lineare *CoDA* (*Communities through Directed Affiliations*) [16] e *Louvain* [7]. Poiché gli *hashtag* sono di fatto dei meta-tag generati dagli utenti, non è necessario eseguire una *feature selection*, in quanto rappresentano già dei concetti e le loro statistiche possono essere usate così come sono ovvero senza l'applicazione di funzioni che normalizzino il loro peso semantico, come *TF-IDF*[2] o mediante altre tecniche di *smoothing*. Comunque, la scalabilità effettiva di *CoDA* è limitata dal fatto che è un algoritmo al momento non distribuibile, cioè ne esiste solo una versione non parallela e *standalone*. *CoDA* prova a determinare il numero ottimale di classi da generare, e questa fase di ottimizzazione può richiedere un tempo computazionale enorme soprattutto in presenza di un numero elevato di *cluster*. Al contrario, *Louvain* è distribuibile ed è in grado di indentificare il numero ottimale di classi in modo molto veloce ed efficace.

Per stimare la distribuzione dei *tweet* nelle varie classi (argomenti o *topic*), valutiamo gli *hashtag* contenuti nei tweet. A loro volta gli hashtag sono stati raggruppati in comunità e a ciascuna comunità è stato assegnato un argomento specifico da una lista predefinita di categorie giornalistiche, partendo dalle comunità più grandi. Si è poi effettuata la stessa classificazione con i primi 1.000 *hashtag* più comuni. Grazie a queste due classificazioni è stato possibile assegnare degli argomenti alla maggior parte dei 600 milioni di *tweet* filtrati.

[2] TF-IDF sta per "Term Frequency-Inverse Document Frequency". È una tecnica di *Information Retrieval* e *Text Mining* utilizzata per quantificare una parola all'interno di documenti.

Tabella 1 - Statistiche relative ai tweet filtrati nell'anno. Gli utenti che hanno prodotto almeno un post in italiano sono 3,1 milioni. Il numero di account attivi italiani è ottenuta utilizzando la stima che solo il 67,5% degli utenti attivi posti almeno un tweet nell'ultimo anno [10].

Tipo	Numero
Tweet filtrati	614.825.151
Falsi positivi	152.658.036
Falsi negativi	36.590.058
Stream Italiano di Twitter	570.282.182
Media giornaliera Tweet Italiani	1.749.332
Tweet Italiani filtrati	533.692.124
Dizionario (parole uniche)	488.625.358
Occorrenze	10.237.354.591
Stima Autori Italiani	3.070.912
Stima Autori Italiani filtrati	3.030.556
Account Attivi Italiani	4.548.927

2 – Architettura Hardware e Software della Piattaforma Big Data

Le elaborazioni sono state svolte dal *Laboratorio Big Data* di DGTCSI-ISCTI, dedicato alle analisi di grandi quantità di dati, all'implementazione di software e realizzazione di servizi di tipo Big Data e Data Streaming a disposizione del Ministero, anche su tematiche di tipo IoT o di monitoraggio di flussi quali quelli delle piattaforme sociali. Il laboratorio è stato inizialmente allestito nel 2010, ma la sua attuale configurazione in termini sia di *hardware* sia di *tecnologie software* è nato nel 2015. L'attuale configurazione del laboratorio Big Data è basata principalmente su tre gruppi di *server* di 8 macchine ciascuna ai quali afferiscono 3 diverse generazioni di macchine, per una RAM complessiva di 584 GB e 548 core. Per quanto riguarda l'elaborazione massiva dei dati viene utilizzato *Spark* come unico ecosistema per implementare gli algoritmi di *Data Mining* e sperimentare le nuove metodologie su *Big Data*.

Oltre a questa componente di estrazione di informazione, la piattaforma comprende:

- La componente di filtraggio, memorizzazione e indicizzazione dello streaming.
- La componente di analisi e visualizzazione real-time del flusso corrente.
- La componente *batch* per la consultazione e visualizzazione dei risultati.
- La componente di visualizzazione e navigazione dei grafi delle interazioni.

3 – Indicizzazione dei dati e analisi

Il monitoraggio è iniziato il giorno 25 maggio 2020. Il processo contenuto in questo report si riferisce a dati collezionati fino al giorno 25 aprile 2021. Dal momento che non è possibile filtrare l'intero *stream* della lingua italiana senza sottomettere un'interrogazione, si è deciso di impostare un filtro con le 400 parole di maggiore uso nella lingua italiana, di tipo funzionale e non semantico (*stop-word*), anche per evitare di introdurre dei *bias* nel filtraggio. Allo stesso tempo è stato attivato il filtro nativo di *Twitter* specifico della lingua italiana per ridurre il rumore, cioè per escludere i *tweet* con *stop-word* italiane che potessero essere usate anche in altre lingue.

3.1 – Dati collezionati

Dal 25 maggio 2020 al 25 Aprile 2021, sono stati recuperati 614.825.151 *tweet*, contenenti 488.625.358 entità (con *account*, citazioni, *hashtag*, parole, *emoji*, paesi, città e URL). Il numero totale di queste entità è pari a 10.237.354.591, ovvero la lunghezza media di ogni *tweet* è di 16,65 entità. Grazie alla stima del numero di falsi positivi e negativi, è possibile stabilire quanto è grande il flusso italiano di *Twitter*: ci sono 570.282.182 *tweet* di cui 533.692.124 sono stati filtrati, 152.658.036 sono i falsi positivi mentre 36.590.058 non sono stati filtrati. La media giornaliera è di 1.749.332 *tweet* (vedi Tabella 1). Gli autori di *tweet* nell'anno, cioè utenti attivi che hanno postato almeno un *tweet* nell'ultimo anno, sono stati circa 3,1 milioni e gli utenti attivi sono invece 4,5 milioni. Il numero di account attivi italiani è stato ottenuto utilizzando la stima che solo il 67,5% degli utenti attivi posti almeno un tweet nell'ultimo anno [10].

3.1.1 – Stima dei falsi positivi del filtro.

Abbiamo valutato il numero di post erroneamente filtrati come italiani ma attribuibili ad altre lingue. Questa valutazione è necessaria per ottenere delle statistiche sul reale numero di utenti attivi in Italia e, quindi, per determinare la percentuale d'uso di questo social network. Di seguito illustreremo la metodologia usata per ottenere le seguenti stime:

- Numero di utenti italiani attivi su *Twitter* lo scorso anno
- Volume giornaliero del sottoinsieme italiano del *Firehose* di *Twitter*

Abbiamo estratto casualmente 1.000 account dai 18 milioni di utenti filtrati, abbiamo stabilito la lingua usata da questi utenti considerando i loro tweet prodotti e, per ciascun account del campione, abbiamo considerato i tweet che sono stati filtrati e attribuiti alla lingua italiana dal classificatore di Twitter. In questo modo è stato ottenuto un campione di 67.118 tweet altamente significativo del *Firehose* italiano. Successivamente è stata valutata la lingua usata da ciascun account. Abbiamo così stabilito l'accuratezza ottenuta dal classificatore, che è stata pari all' 86,8% (vedi Tabella 3).

3.1.2 – Stima dei falsi negativi

Per stabilire la percentuale dei *tweet* non catturati dal filtro, sono stati monitorati indipendentemente alcuni account italiani e quindi è stata calcolata la percentuale di dispersione (falsi negativi) del numero di *tweet* italiani. Il numero di falsi negativi è stato di 140 tweet mancati ogni 2042 tweet filtrati (pari al 6,86%), con l'intervallo di confidenza [5,76%, 7,95%].

3.2 – Numero di autori e utenti attivi italiani

Non è possibile, ad oggi, sapere gli utenti di Twitter che abbiano eseguito l'accesso alla piattaforma almeno una volta nell'ultimo anno. Possiamo però monitorare alcune interazioni. Ad esempio, è possibile stimare gli utenti attivi che abbiano prodotto almeno un tweet. Esistono diverse stime sul numero di utenti attivi italiani, estremamente discordanti tra loro. Per esempio, Twitter dichiara che ci sono 2,8 milioni di utenti mensili raggiunti dalle loro

pubblicità, ma il report Digital 2021[3] afferma che la piattaforma ha una share del 32,8%, pari a 12,4 milioni di utenti di età compresa tra i 13 e i 64 anni. *Statcounter*[4] fornisce invece un dato più conservativo: lo share è del 4,39% tra tutte le piattaforme social. L'Agenzia italiana delle Comunicazioni (AGCOM) ha affermato che sono presenti circa 8,2 milioni di utenti unici su *Twitter* nel 2017 [1], con una share del 9,9%. La nostra stima relativa all'anno 2021 è invece più vicina ai valori MAU di *Twitter* e alla stima di *Statcounter*.

3.2.1 – Metodologia di stima

A seguire verranno elencati i passi adottati per stabilire il numero di utenti:

- Il *database* è composto da 18.451.666 *account*. Ogni autore ha prodotto almeno un *tweet* catturato dal nostro filtro e attribuito alla lingua italiana dal classificatore automatico di *Twitter*. Per stabilire il numero corretto di autori italiani, è stato valutato l'errore compiuto dal classificatore ovvero sono stati stabiliti il numero di falsi positivi (errore di tipo 1) e negativi (errore di tipo 2). Per raggiungere tale obiettivo, è stato estratto un campione di 1.000 account sul quale è stata eseguita una valutazione manuale della lingua usata nei tweet. Grazie a questa valutazione, e non considerando l'errore di tipo 2, siamo stati in grado di affermare che numero di account attivi italiani è il 16,38% con un intervallo di confidenza di [14,3% 18,55%]. Gli account italiani nel nostro database, dunque, sono 3.030.556. L'errore di tipo 1 per la stima del numero di utenti italiani, comunque, non è l'errore di tipo 1 del classificatore. Può succedere infatti che, per un utente straniero, solo un *tweet* (o pochi *tweet*) siano stati erroneamente attribuiti alla lingua italiana. Dal momento che l'errore del classificatore è pari a 16,38%, la probabilità che due *tweet* dallo stesso account straniero siano stati erroneamente attribuiti alla lingua italiana scende al 2,8%, per tre *tweet* allo 0,44%, e così via.

[3] https://datareportal.com/reports/digital- 2021- italy

[4] https://gs.statcounter.com/social-media-stats/all/italy

- Ai 3.030.556 di *account* filtrati italiani occorre aggiungere i falsi negativi ovvero gli account non catturati dal filtro. Osserviamo che, così come alcuni *tweet* sono stati attribuiti alla lingua italiana, alcuni *tweet* italiani sono stati attribuiti ad altre lingue e quindi non sono presenti nella base dati. Inoltre, le parole chiave funzionali usate potrebbero non essere state sufficienti a caratterizzare e filtrare i testi in italiano. Si è notato che il numero di tweet italiani non filtrati è stato pari al 6,86%. È comunque necessario calcolare quanto questa percentuale influisce sui falsi negativi per gli utenti attivi. Ovviamente la percentuale di errore del filtro non influisce sul numero di utenti attivi allo stesso modo perché, come già spiegato, solo un ridotto numero di *account* italiani (quelli che hanno prodotto uno o al più due *tweet*) potrebbero essere soggetti all'errore del classificatore e quindi assenti nella base dati. Si è quindi calcolata una stima della dispersione sul campione: solo il 6,12% degli *account* che hanno prodotto un singolo *tweet* sono italiani, e quindi gli *account* italiani che hanno prodotto un singolo *tweet* sono 588.618, pertanto i falsi negativi sono stati pari a 40.356 (6,86%). Il numero di account con due o più tweet prodotti non è stato incluso nel calcolo perché trascurabile.

3.2.2 – Stima finale degli autori e utenti attivi con i falsi positivi e negativi

La stima finale per gli utenti italiani attivi è 3.070.912 di cui 3.030.556 filtrati e inclusi nel database. Infine, tenendo conto che il 32,5% degli utenti attivi non posta *tweet* nell'anno [10] otteniamo una stima di circa 4,5 milioni di utenti attivi su *Twitter*.

4 - Identificazione degli argomenti trattati da Twitter Italiano

Affrontiamo ora il problema di classificare i *tweet* secondo categorie informative e argomenti (*topic*). In *Machine Learning* e *Information Retrieval* questo problema è noto come *topic modelling*. L'estrazione di *topic* da una elevata quantità di dati presenta due problematiche da risolvere:

(1) Occorre innanzitutto individuare una tecnica di classificazione o di clustering scalabile. Ad esempio, la tecnica maggiormente utilizzata su dati di dimensione piccola, LDA, è *NP-hard* in presenza di un elevato numero di termini, come nel nostro caso [13]. Per questa ragione, gli unici algoritmi scalabili per i *Big Data* sono quelli lineari. Al momento si potrebbero considerare tre algoritmi lineari. Il primo, LSH, elabora gli oggetti complessi suddividendoli in insiemi di frammenti, che sono definiti in qualche modo in base al problema che deve essere modellato, con il fine di mantenere la struttura originaria. Nel nostro caso, per esempio, si potrebbero considerare gli insiemi di tutte le possibili co-occorrenze di parole o di *hashtag*. Per questo insieme, si estraggono delle firme usando delle funzioni *hash*. Ogni oggetto è quindi descritto da un vettore di valori numerici di dimensione predeterminata che preserva le proprietà geometriche dell'oggetto. È possibile determinare la distanza tra due oggetti semplicemente controllando il numero di valori *hash* condivisi. Nonostante risulti un algoritmo interessante, in quanto lineare nel numero di dati, ovvero che ogni dato che ogni oggetto è processato individualmente e inserito in una o più classi, in LSH il numero di classi può potenzialmente essere molto elevato. Il numero di classi dipende dal valore minimo di similarità all'interno di ogni classe richiesto. Inoltre, a causa della dimensione ridotta dei *tweet*, si otterrebbero molte classi aventi pochi tweet simili tra loro, a meno che non si decida di ridurre troppo la soglia di similarità. In base alla nostra esperienza, l'algoritmo non è molto scalabile a causa dell'enorme quantità di memoria richiesta per gestire grandi quantità di classi.

(2) Il secondo problema è l'identificazione dei termini o caratteristiche, *feature,* usati nell'algoritmo (***feature selection***). Solo gli *hashtag* hanno statistiche comparabili tra loro senza che ci sia bisogno di normalizzarle. Per le parole comuni, infatti, si pone il problema delle parole funzionali (*stop-word*) o delle parole che hanno valori qualitativi non caratterizzanti, o che caratterizzano dei concetti quando combinate con altre parole. Il problema dell'estrazione delle parole chiave (*feature*), su cui eseguire la classificazione, potrebbe essere insormontabile rendendo l'approccio basato su *clustering*/classificazione intrattabile. Per questa ragione si è deciso di non procedere con le tecniche note di *clustering*/classificazione come LDA o LSH.

Tabella 2 - Dati da un campione di 1.000 account. La percentuale di account italiani è il 16,38% (intervallo di confidenza [14,3%, 18,55%]).

Lingua dell'account	%
Spagnolo	26,67
Portoghese	19,64
Inglese	18,01
Italiano	16,38
Francese	4,12
Coreano	2,40
Filippino	1,72
Indonesiano	1,54
Turco	1,46
Hindi	1,29
Giapponese	1,11
Altri	5,66

Tabella 3 - Sono 67.118 i tweet prodotti dal campione dei 1.000 account. L'accuratezza del classificatore di Twitter la lingua italiana è del 86,8% [85,73%, 86,26].

Lingua	Tweet filtrati per account	% Tweet
Italiano	318	86,80
Spagnolo	16	7,00
Inglese	8	2,31
Portoghese	7	2,19
Coreano	8	0,32
Francese	5	0,32
Hindi	8	0,18
Filippino	5	0,16
Giapponese	7	0,15
Catalano	19	0,08
Turco	3	0,07

4.1 – Clustering massivo con Louvain

Abbiamo considerato due algoritmi per l'individuazione delle comunità nei grafi di comunicazione: CoDA [16] e Louvain [7]. Louvain ha anche una versione distribuita in *Spark*, facilmente integrabile nella nostra piattaforma per l'analisi dei grafi. Il nostro approccio non è stato ancora utilizzato per effettuare il clustering massivo di *Twitter*. L'algoritmo adottato è il seguente:

- Sono stati considerati gli hashtag come nodi di un grafo, ed è stato costruito il grafo pesato delle co-occorrenze degli hashtag. Non abbiamo incluso le singole parole a causa del problema della loro selezione e normalizzazione già discussa precedentemente. Il numero di *hashtag* (nodi del grafo) è stata pari a 2.726.034 con 34.769.762 co-occorrenze (archi non diretti pesati del grafo).
- Si sono quindi eseguiti I due algoritmi di CoDA e Louvain per identificare le comunità sugli insiemi degli hashtag. Nonostante CoDA sia lineare, è di fatto scalabile solo se il numero di classi generate non sia elevato (l'algoritmo prova a determinare il numero ottimale di classi a cui assegnare i nodi, fino ad un massimo di K classi stabilito a priori). Si sono impostati valori crescenti di K ma il numero ottimale di classi ottenuto è sempre stato lo stesso K, pertanto si è impostato un K elevato (100.000 topic) al fine di determinare il valore ottimo. Per valori inferiori di 10.000 classi, l'algoritmo ha impiegato un tempo sufficientemente ragionevole, mentre per valori superiori è di fatto impraticabile. In assenza di una versione distribuita, CoDA è a tutti gli effetti non scalabile nel nostro caso, anche se arriva a terminare il calcolo. Come CoDA, anche Louvain prova a determinare il numero di classi ottimale ma, a differenza del primo algoritmo, è in grado di terminare in tempi eccezionalmente brevi. L'algoritmo di Louvain ha determinato 64.338 argomenti (comunità di nodi utilizzando la terminologia dei grafi sociali).

- Identificate le comunità (64.338) di *hashtag*, si sono assegnate manualmente a argomenti scelti da una lista di categorie predefinite (questo processo può essere sostituito in futuro da un classificatore automatico, attualmente è di nostro interesse presentare solo un approccio generale). Sono state utilizzate come categorie quelle usate abitualmente a livello giornalistico per catalogare gli articoli (politica, esteri, sport, cinema, televisione, musica, pubblicità, scienza, tecnologia, economia, salute, ecc.), a cui sono state aggiunte quelle relative ai social network (*influencer*, *blogger*, *bot*, ecc.).
- Grazie a questa partizione in classi, ogni *hashtag* è stato associato ad una *topic*, ovvero la categoria associata al *cluster* a cui appartiene, e questa topic è stata poi associata a tutti i *tweet* che contengono quell'*hashtag*. Allo stesso tempo, si è stilata manualmente una classifica dei 1.000 hashtag più frequenti. Nel caso in cui diverse topic sono associate allo stesso *tweet*, le *topic* sono state processate come nell'algoritmo *Random Forest* scegliendo quella con il numero maggiore di hashtag (la topic che contiene più citazioni con gli *hashtag*). Una volta terminata questa valutazione, siamo stati in grado di assegnare una *topic* alla maggioranza dei *tweet* in esame.

5 - Gli argomenti trattati su Twitter

Dopo aver valutato manualmente ~~i~~primi 1.000 *hashtag* più frequenti e i *cluster* con il maggior numero di elementi così come identificati dall'algoritmo di Louvain, si è stati in grado di stimare il numero di citazioni delle categorie (vedi Tabella 4) e anche le categorie più frequenti (Tabella 5).

Tabella 4 - Le topic individuate nel flusso di Twitter italiano. Twitter rimane una piattaforma social fortemente legata al contesto televisivo o usata per il live sharing e la discussione durante gli eventi di maggiore rilevanza.

Etichetta	Nr hashtag	Citazioni	% citazioni
Tv Radio	54.092	52.257.298	35,09%
Politica	478.869	34.433.579	23,12%
Sport	223.304	17.727.610	11,90%
Salute	33.538	7.666.946	5,15%
Arte Cultura	104.351	5.801.901	3,90%
Scienza & Tecnologia	182.622	5.169.674	3,47%
Esteri	7.666	3.441.604	2,31%
Agenzie notizie	1.012	4.159.098	2,79%
Musica	17.982	2.940.716	1,97%
Turismo	20.888	2.566.472	1,72%
Economia	10.915	2.468.186	1,66%
Sesso	48.307	1.905.758	1,28%
Giochi	49.062	1.255.619	0,84%
Cinema Teatro	8.461	996.074	0,67%
Cibo	22.148	765.366	0,51%
Social Network	6.791	722.901	0,49%
Educazione	7.452	700.120	0,47%
PA	2.741	579.316	0,39%
Pubblicità	7.652	394.309	0,26%
Traffico e meteo	6.604	373.956	0,25%
Ambiente	5.954	243.520	0,16%
Fakenews	1.828	135.520	0,09%
Altro	72	2.236.543	1,50%

Tabella 5 - Le prime 50 topic della Tabella 4

Topics	Nr hashtag	Citazioni	%
Grande Fratello VIP	436	31.439.876	21,11
Temi politici	441.498	20.875.461	14,02
Calcio	149.152	15.004.881	10,07
Daydreamer (serie TV)	11.322	7.412.160	4,97
Covid	27.858	6.868.005	4,61
Scienza Tecnologia	182.616	4.982.114	3,34
Arte Cultura	104.295	4.370.711	2,93
Governo	94	3.512.824	2,36
Turismo	20.888	2.566.472	1,72
Lega (Partito politico)	8.879	2.284.622	1,53
Musica	16.809	2.015.732	1,35
TV (Altro)	33.240	1.939.294	1,30
Sesso	48.307	1.905.758	1,28
Regioni italiane	60	1.590.631	1,07
Economia	10679	1.564.782	1,05
M5S (Partito politico)	145	1.502.137	1,01
Altre news	896	1460429	0,98
Esteri	6083	1.438.721	0,96
Motori	45.133	1.436.897	0,96
Amici (TV)	4	1287795	0,86
Giochi	49.061	1240421	0,83
Giustizia	17.388	977.596	0,65
Altre serie TV	3.393	893.481	0,60
Italia Viva (Partito politico)	1.022	815.262	0,55
Mondo	20	799.689	0,54
PD (Partito politico)	8.507	768.294	0,52
Cucina	22.148	765.366	0,51
Altri sport	29.002	723.084	0,49
Altro Musica	1.170	694.210	0,47
Libri	43	683.042	0,46

Topics	Nr hashtag	Citazioni	%
Istituzioni europee	19	682.621	0,46
Economia	128	656.381	0,44
Uomini e donne (TV)	2	601.579	0,40
Temptation island (TV)	1	550.421	0,37
Educazione	7.449	544.248	0,36
Immigrazione	1.253	523.975	0,35
Ministero della Salute	2	503.891	0,34
Vip Amici (TV)	1	489.177	0,33
X Factor (TV)	33	447.294	0,30
Cinema	8.457	439.153	0,29
Il collegio (TV)	2	432.496	0,29
Referendum	4	412.189	0,28
Advertisement	7.651	394.203	0,26
Formula 1	11	386.516	0,26

6 – Conclusioni

È stato introdotto un nuovo approccio di *topic modelling* per la *Twitter*-sfera, basato su algoritmi di *community detection* applicati al grafo degli *hashtag*. Si è mostrato che nel caso specifico di testi di dimensioni ridotte come i *tweet*, questo approccio è di gran lunga più efficiente rispetto agli approcci di *topic modeling* convenzionali basati sulla *feature selection* oppure agli algoritmi di *clustering* come *LDA*. L'algoritmo di *clustering* massivo, introdotto in questa ricerca, è invece facilmente applicabile e scalabile sui *social network* come *Twitter*. Una volta suddivisi gli *hashtag* in classi di similarità, i *cluster* con il maggior numero di elementi e gli *hashtag* sono stati associati alle *topic* di interesse generale (come lo sport, la politica, la salute, ecc). In questo modo si è stati in grado di fornire statistiche sulle *topic* presenti su *Twitter* lo scorso anno. In futuro potrebbe essere interessante validare l'approccio con CoDA e con un sottoinsieme del flusso di Twitter per altre lingue con un valore di mDAU comparabile, e sviluppare un classificatore automatico per associare i *cluster* di hashtag con le *topic* di interesse generale.

8 – Bibliografia

[1] - AGCOM. [n.d.]. "Osservatorio sulle comunicazioni" n. 4/2017. https://www.agcom.it/documents/10179/3316522/Studio-Ricerca+18-01- 2018/f1a47c94- 2c53- 4006- 8ce7- 4c595c9ffa9d?version=1.0

[2] - G. Amati, S. Angelini, F. Capri, G. Gambosi, G. Rossi, P. Vocca. 2017. "On the Retweet Decay of the Evolutionary Retweet Graph". In *Smart Objects and Technologies for Social Good: Second International Conference, GOODTECHS 2016*, Venice, Italy, November 30 – December 1, 2016, Proceedings. Springer International Publishing, Cham, 243–253. https://doi.org/10.1007/978- 3- 319- 61949- 1_26

[3] - G. Amati, S. Angelini, G. Gambosi, G. Rossi and P. Vocca, "Estimation of distance-based metrics for very large graphs with MinHash Signatures," *2017 IEEE International Conference on Big Data (Big Data)*, 2017, pp. 536-545, doi: 10.1109/BigData.2017.8257969.

[4] - G. Amati, S. Angelini, G. Gambosi, G. Rossi, and P. Vocca. 2019. "Influential users in Twitter: detection and evolution analysis". *Multimedia Tools and Applications* 78, 3 (2019), 3395–3407.

[5] - G. Amati, D. Pasquini, S. Angelini, G. Rossi, G. Gambosi, and P. Vocca. 2018. "Twitter: Temporal events analysis". *ACM International Conference Proceeding Series* (2018), 298–303.

[6] - S. Barbon, G.Kido, G. Tavares. "Artificial and Natural Topic Detection in Online Social Networks". *Revista Brasileira de Sistemas de Informatycs*, 10, (2017), 80–98. https://doi.org/10.5753/isys.2017.329

[7] - V. Blondel, J.L. Guillaume, R. Lambiotte, E. Lefebvre. 2008. "Fast unfolding of communities in large networks". *Journal of Statistical Mechanics: Theory and Experiment* 2008, 10 (2008), P10008. http: //stacks.iop.org/1742- 5468/2008/i=10/a=P10008

[8] - A. Gionis, P. Indyk, and R. Motwani, "Similarity Search in High Dimensions via Hashing". In *Proc. 25th Int. Conf. on Very Large Data Bases*. Edinburgh, (1999), UK, 518–529.

[9] - K. H. Lim, S. Karunasekera, A. Harwood, "ClusTop: A clustering-based topic modelling algorithm for twitter using word networks". In 2017 IEEE International Conference on Big Data (Big Data). 2017, 2009–2018.

[10] - Y. Liu, C. Kliman-Silver, A. Mislove, "The Tweets They Are a-Changin': Evolution of Twitter Users and Behavior". *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 2014, 305-314.

[11] - J. MacQueen, "Some methods for classification and analysis of multivariate observations". In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1. Oakland, CA, USA, 1967, 281–297.

[12] - A. Rajaraman, J. Leskovec, and J. Ullman, "Mining of Massive Datasets". *Cambridge University Press*, 2014.

[13] - D. Sontag, D. M. Roy, "Complexity of Inference in Latent Dirichlet Allocation"*, (NIPS'11). Curran Associates Inc., Red Hook*, NY, USA, 2011, 1008–1016.

[14] - A. Steinskog, J. Therkelsen, B. Gambäck, "Twitter Topic Modeling by Tweet Aggregation". In *Proceedings of the 21st Nordic Conference on Computational Linguistics. Association for Computational Linguistics*, Gothenburg, Sweden, 2017,77–86.

[15] - X. Wang, F. Wei, X. Liu, M. Zhou, M. Zhang, "Topic Sentiment Analysis in Twitter: A Graph-based Hashtag Sentiment Classification Approach". In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management* (Glasgow, Scotland, UK) (CIKM '11). ACM, New York, NY, USA, 2011, 1031–1040.

[16] - J. Yang, J. McAuley, and J. Leskovec, "Detecting cohesive and 2-mode communities indirected and undirected networks". *WSDM 2014 - Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, 2014, 323–332.