Università
Ca'Foscari
Venezia

Corso di Laurea Magistrale

in

Marketing a Comunicazione

Tesi di laurea

Analisi statistica di dati testuali:
la percezione del Pride month in Italia e all'estero

Relatore
Ch. Prof. Debora Slanzi

Laureando
Alberto Brescancin
Matricola 887994

Anno Accademico
2021 / 2022

Indice

Ringraziamenti

Ringrazio la Professoressa Debora Slanzi per essere stata flessibile con le mie esigenze, per essermi stata di supporto nelle mie difficoltà.

Ringrazio la mia famiglia, per avermi dato la possibilità di studiare e intraprendere nuovamente un percorso universitario. Mamma e papà, senza di voi non sarei stato qui oggi.
A mia sorella Federica, che mi sprona sempre a dare il meglio di me, che mi spinge a non mollare mai. Fede tu per me ci sei sempre stata dal giorno in cui ho preso vita, sei la mia sicurezza.
A Ketty e Lucrezia, per avermi sempre supportato e sopportato le mie lamentele sull'università, ci siete sempre per me.
Ringrazio questa laurea magistrale per avermi permesso di essere in JEVE e aver conosciuto tante persone, su tutte le mie Viperelle, Carlotta e Maddalena, con voi ho condiviso tanti momenti, mi auguro di condividerne tanti altri dopo questo traguardo, nonostante la lontananza.
Un grazie speciale anche a Laura, Sofia, Michela, Lorenzo e Luca.
Ai miei compagni di corso Giorgia, Elena e Nicola con cui ho diviso tanti lavori di gruppo e non solo, è stato bello condividere questo percorso con voi.
Ai miei amici di Trieste, che sono con me da quando ho iniziato l'università, vi sono grato dei momenti passati sia durante che dopo la fine della triennale, in particolare Veronica F. con cui ci siamo supportati nei nostri momenti di sconforto.
Ai miei amici dell'infanzia e gruppo Esclusi, con cui ho passato dei momenti bellissimi ed indimenticabili negli ultimi due anni tra un esame e l'altro.
A Federica, Agnese e Valeria che ci sono dal primo giorno di superiori, vi auguro di raggiungere questo traguardo al più presto.
A Mattia S. che mi ha sempre aiutato durante i momenti di difficoltà, ti sarò sempre grato.
A tutte le persone che mi hanno supportato e sopportato, grazie di cuore.

A me stesso, per essere stato più forte del pensiero di lasciare questo percorso, per aver trovato la voglia e motivazione quando queste mancavano.
In questi due anni sono successe tante cose che non avrei nemmeno immaginato, ho cambiato casa, continente, abitudini, conosciuto nuove persone e stretto nuove amicizie. Nonostante tanti squilibri e poche certezze, sono riuscito a non perdere la strada.

Ed ancora una volta,

a chi è stato,
a chi c'è,
a chi sarà.

Introduzione

L'utilizzo dei social media negli ultimi anni ha subito una crescita esponenziale in tutto il mondo ed è parte quotidiana della maggior parte degli individui. Ciò ha reso accessibile un'enorme mole di dati sia di carattere orizzontale con la geolocalizzazione, sia di carattere verticale come dati anagrafici, socio-economici, abitudini quotidiane ecc.
La possibilità di interagire con altri individui comporta anche la condivisione e la partecipazione a più discussioni, rese disponibili per differenti individui.
Una chiave che rende disponibile l'utilizzo dei social media riguarda la facilità di accesso agli stessi.
La possibilità di condividere un'opinione ha rivoluzionato diversi settori, non riguarda solo temi sociali, ma anche politici ed economici.
Le aziende hanno la possibilità di utilizzare i social media come strumento di comunicazione del proprio prodotto e servizio e soprattutto capire in modo diretto le esigenze dei clienti e le loro opinioni.
Pertanto, si è iniziato a studiare i bisogni, gli atteggiamenti e le opinioni degli utenti attraverso l'analisi dei Social media e ciò ha determinato la nascita di nuove metodologie per la raccolta e l'analisi dei dati estratti dai social network, grazie alla facilità di ottenere informazioni con velocità e minor costo di realizzazione.
Ad oggi, i principali social media presenti nel web sono Facebook, Instagram, Tik Tok e Twitter. Questi social media hanno come vantaggi principali il costo pressoché nullo e la rapidità di ottenimento di informazioni; d'altra parte, ci sono svantaggi nell'utilizzo di questi strumenti al fine di ottenere dati, quali riconoscere le caratteristiche delle persone, come età, sesso, geolocalizzazione, porta spesso ad ottenere un campione non rappresentativo o mancante di informazioni utili per i successi: le analisi. Una volta ottenuti i dati, esistono poi varie metodologie di analisi che si differenziano per gli obiettivi specifici da raggiungere.

Tra le recenti metodologie sviluppate, vi è la Sentiment Analysis, il cui scopo è quello di comprendere e classificare i sentimenti, cioè le opinioni su un particolare tema attraverso l'analisi dei testi e le singole parole che li compongono.

Il presente elaborato si propone di presentare alcune metodologie di classificazione testuale applicate ad un caso studio specifico, al fine comprendere l'opinione sul tema

analizzato presente sui social network, per il quale l'obiettivo sarà anche valutare la percezione degli utenti italiani nei confronti degli utenti stranieri.

Nel capitolo 1 verranno descritte la nascita dei social media e verranno presentati i diversi social media, attuali e passati, per capire l'importanza dell'era digitale.
Nel capitolo 2 verranno descritte le sfere sociali, economiche e politiche che sono influenzate dai social media e verranno presentati dei casi studio per ogni sfera descritta.
Il terzo capitolo riguarderà l'analisi testuale e le metodologie che possono produrre la stessa, in particolare verrà illustrata la metodologia utilizzata per la Sentiment Analysis.
Nell'ultimo capitolo, viene presentato il caso studio proposto: nella prima parte viene raccontata la storia del *Pride month*, dalla sua nascita fino ai giorni nostri, mentre nella seconda parte verranno applicate le tecniche precedentemente introdotte per analizzare il sentimento espresso sul tema, sia da utenti italiani che stranieri.
L'elaborato si concluderà con un confronto dei risultati ottenuti e delle considerazioni personali.

CAPITOLO 1: I SOCIAL MEDIA

I social media sono un fenomeno che ha trasformato l'interazione e la comunicazione degli individui in tutto il mondo. Negli ultimi tempi, i social media hanno avuto un impatto su molti aspetti della comunicazione umana, influenzando così anche il business. I social network sono diventati una pratica quotidiana nella vita di alcuni utenti. In questo capitolo, verranno descritte le caratteristiche e l'evoluzione dei social media, compresi i principali siti di social networking nati nel XXI secolo. Alcuni dei siti trattati sono Facebook, YouTube, Twitter, MySpace, CyWorld e LunarStorm.

Prima di spiegare l'evoluzione dei social media, è bene dare una definizione a questi. Il dizionario Merriam-Webster definisce i social media come "forme di comunicazione elettronica (come i siti Web per il social networking e il blogging) attraverso le quali gli utenti creano comunità online per condividere informazioni, idee, messaggi personali e altri contenuti (come i video)".

1.1 La nascita di internet ed il web 2.0

Negli anni '60 le persone hanno assistito all'avvento della posta elettronica (Borders, 2010). Tuttavia, Internet non fu disponibile al pubblico fino al 1991. In origine l'e-mail era un metodo per scambiare messaggi da un computer all'altro, ma entrambi i computer dovevano essere online. Oggi, i server di posta elettronica accettano e archiviano i messaggi, consentendo ai destinatari di accedervi a loro piacimento. Nel 1969 è stata sviluppata ARPANET, creata dall'Advanced Research Projects Agency (ARPA), un'agenzia governativa statunitense. ARPANET è stata una "prima rete di computer in time-sharing che ha costituito la base di Internet".
La rappresentazione di ARPANET è mostrata in figura 1.1.

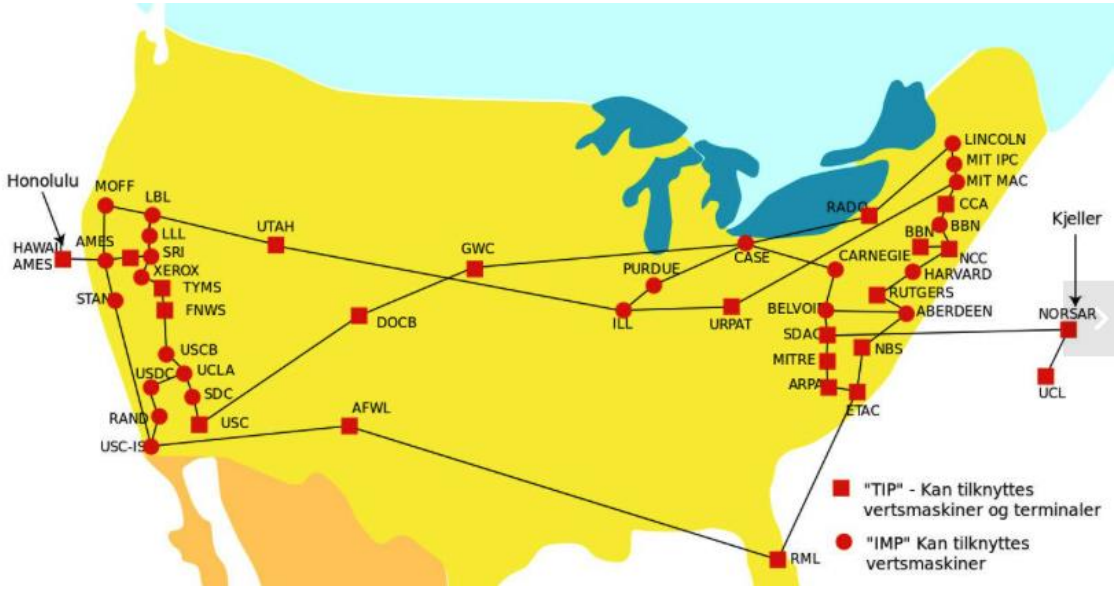

Figura 1.1 Rappresentazione di ARPANET, fonte: web

Anche CompuServe, il terzo sviluppo degli anni '60, fu creato nel 1969 con la missione di fornire servizi di condivisione del tempo affittando il tempo sui suoi computer. Con tariffe molto elevate, questo servizio era troppo costoso per molti (Rimskii, 2011; Ritholz, 2010).

I social media si sono sviluppati negli anni Settanta. Il MUD, originariamente noto come Multi- User Dungeon, Multi-User Dimension o Multi-User Domain, era un mondo virtuale in tempo reale con giochi di ruolo, fiction interattiva e chat online. Il MUD è principalmente basato sul testo e richiede agli utenti di digitare comandi utilizzando un linguaggio naturale.

La schermata di MUD è raffigurata in Figura 1.2

Figura 1.2 Rappresentazione MUD, fonte: web

Le BBS sono state create nel 1978, lo stesso anno di MUD. BBS è un sinonimo di *bulletin board system*. Gli utenti si collegano al sistema per caricare e scaricare software, leggere notizie o scambiare messaggi con altri utenti. Nei primi anni, le bacheche erano accessibili tramite un modem attraverso una linea telefonica da una persona alla volta. All'inizio le bacheche non avevano colori o grafica, dopodiché sono state i predecessori del World Wide Web. La rappresentazione di BBS è raffigurata in Figura 1.3.

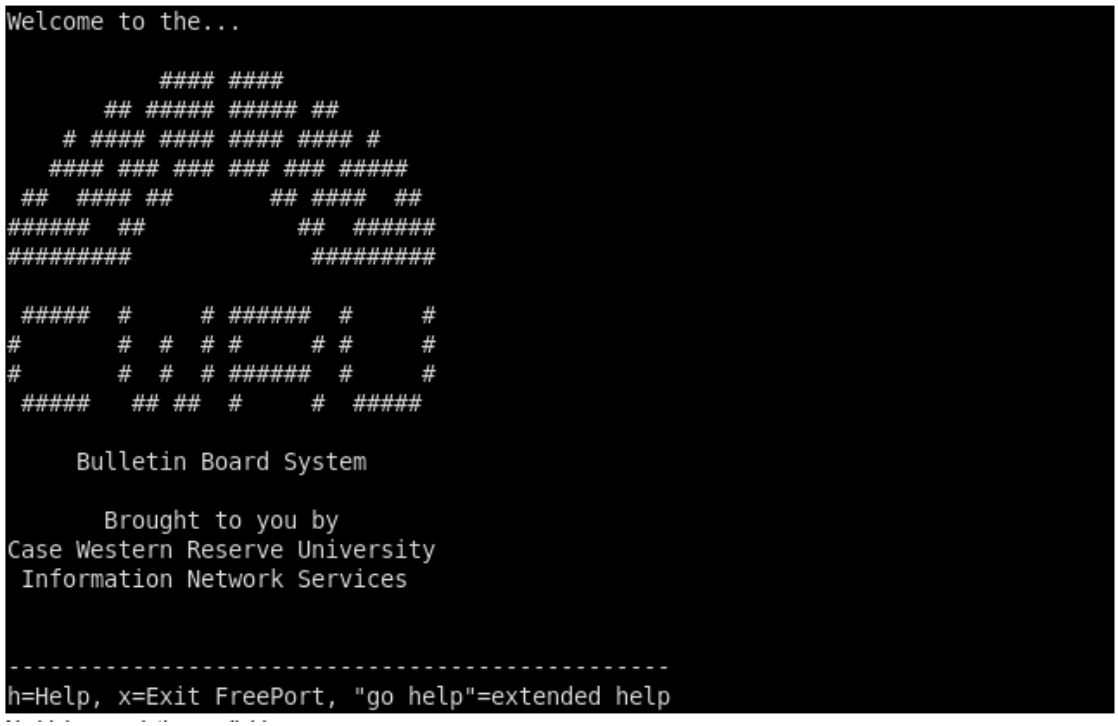

Figura 1.3 Rappresentazione BBS, fonte: web

Concepita nel 1979 e fondata nel 1980, Usenet è simile a una BBS. Usenet è un sistema per pubblicare articoli o notizie. La differenza rispetto a una BBS è che Usenet non ha un server centrale o un amministratore dedicato: i messaggi vengono inoltrati a vari server tramite feed di notizie (Ritholz, 2010).
La rappresentazione del diagramma di Usenet è rappresentato in Figura 1.4.

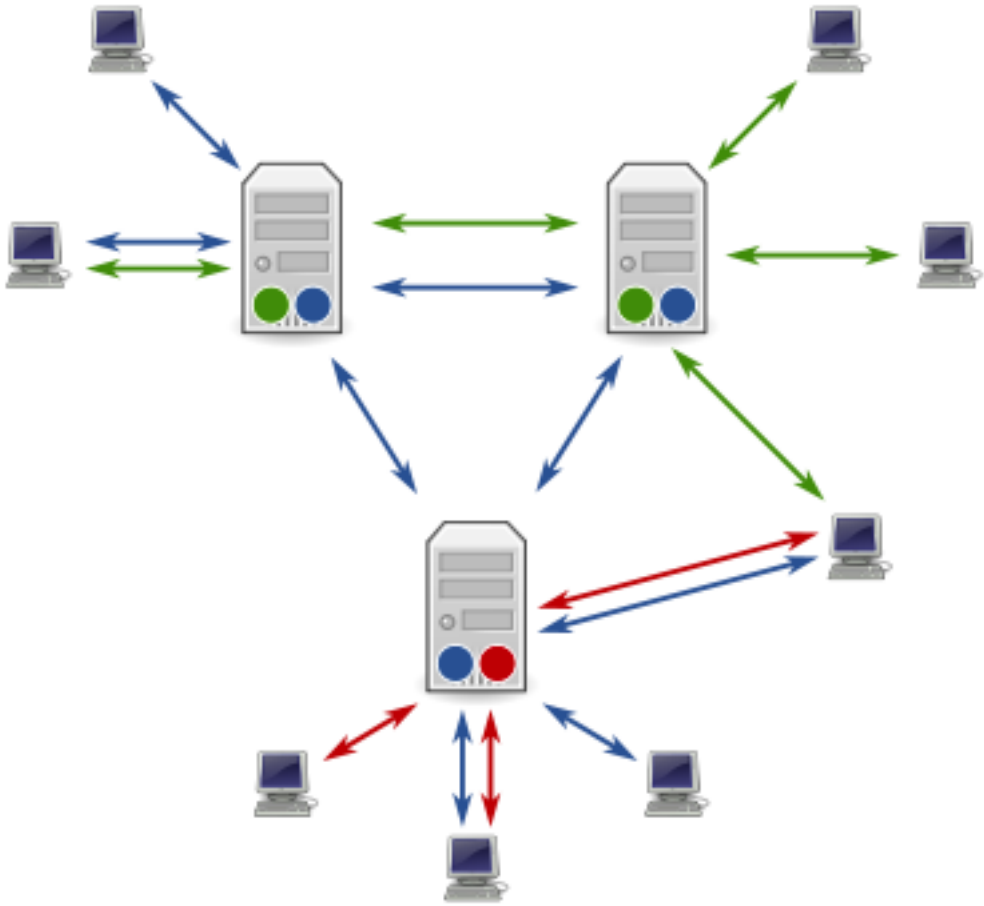

Figura 1.4 Diagramma di usenet servers e clienti, fonte: web

Negli anni '80 furono introdotti WELL, GEnie, Listserv e IRC. Il WELL, nato come BBS, è l'abbreviazione di Whole Earth "Elettronic Link". È stata fondata a Sausalito, in California, da Stewart Brand e Larry Brilliant ed è una delle più antiche comunità virtuali in attività.
Il logo di WELL è rappresentato nella figura 1.5.

Figura 1.5 Rappresentazione logo WELL, fonte: web

GEnie è l'acronimo di General Electric Network for Information Exchange. Era un servizio online che utilizzava il linguaggio ASCII ed era considerato un concorrente di CompuServe. General Electric Information Services (GEIS) gestiva GEnie sui computer mainframe in time-sharing durante le ore non di punta. Inizialmente GEIS si rifiutò di espandere la rete per consentire a GEnie di crescere.
La schermata di GEnie è raffigurata nella figura 1.6.

Figura 1.6 Rappresentazione GEnie, fonte: web

Listserv, lanciato nel 1986, è stata la prima applicazione software per mailing list elettroniche. Prima della sua creazione, le liste di posta elettronica dovevano essere gestite manualmente. Il software consente al mittente di inviare un solo messaggio di posta elettronica per raggiungere più persone. Originariamente Listserv era freeware, ma ora viene venduto a livello commerciale. Attualmente è disponibile una versione gratuita per un limite di dieci liste con non più di 500 iscritti.

La schermata di Listserv è rappresentata nella figura 1.7.

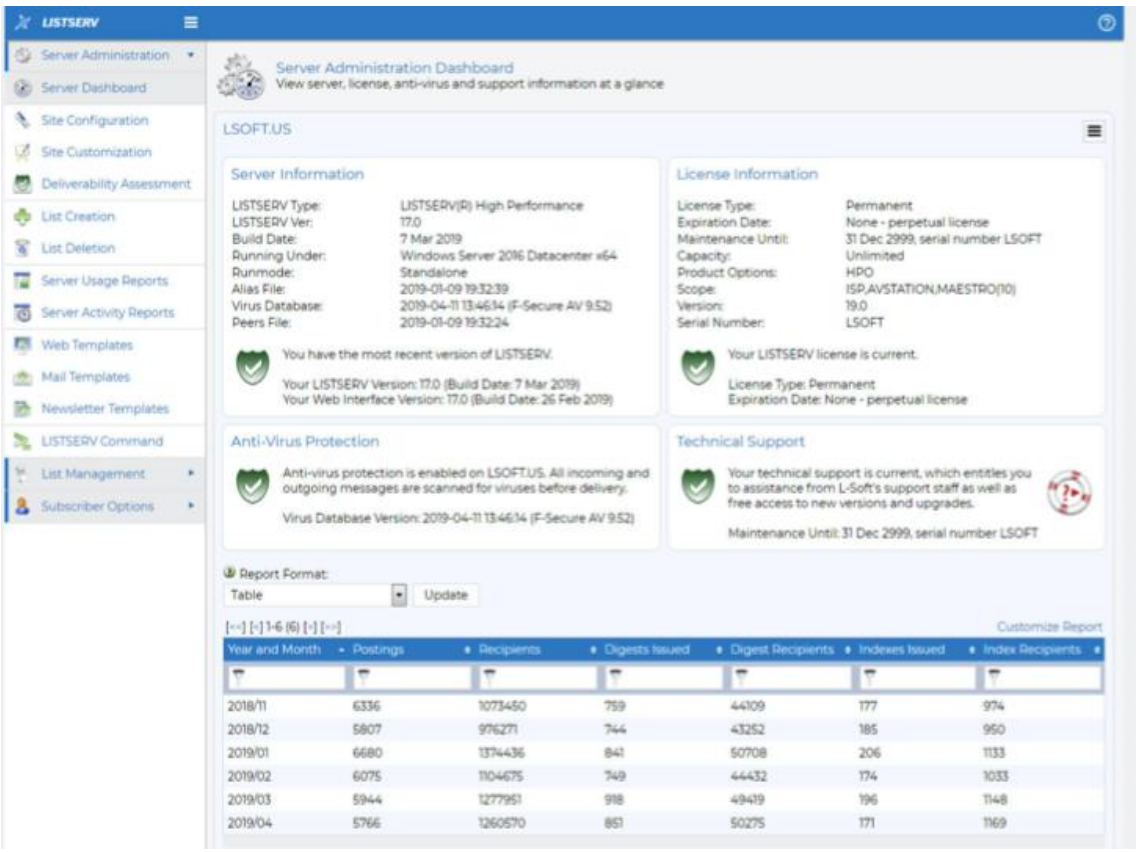

Figura 1.7 Rappresentazione grafica LISTSERV, fonte: web

Un altro esempio è IRC, Internet Relay Chat, progettato per la comunicazione di gruppo. È una forma di chat in tempo reale, nota anche come messaggistica di testo su Internet o conferenza sincrona. Lo scopo principale di IRC è la comunicazione di gruppo, ma consente messaggi privati, chat e trasferimenti di dati tra due utenti (Ritholz, 2010).
La grafica di IRC è raffigurata nella figura 1.8.

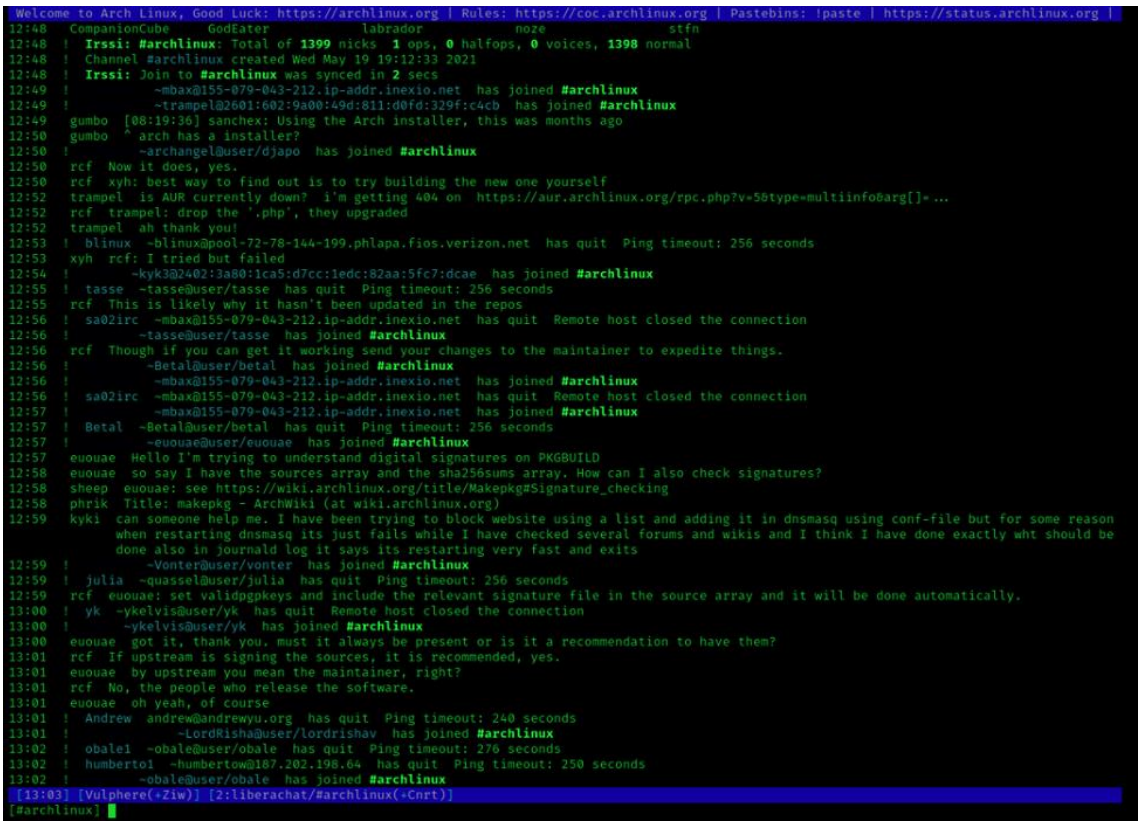

Figura 1.8 Rapprentazione grafica IRC, fonte: web

1.2 L'età dei social media

Molti siti di social networking sono stati creati negli anni Novanta. Alcuni esempi sono Six Degrees, BlackPlanet, Asian Avenue e MoveOn. Questi sono, o sono stati, siti sociali di nicchia online in cui le persone possono interagire, compresi siti per la difesa delle politiche pubbliche e una rete sociale basata su un modello di rete di contatti. Inoltre, sono stati creati servizi di blogging come Blogger ed Epinions.
Epinions è un sito in cui i consumatori possono leggere o creare recensioni di prodotti. ThirdVoice e Napster sono due applicazioni software create negli anni '90 che sono state rimosse dal mercato. ThirdVoice era un plug-in gratuito che permetteva agli utenti di inserire commenti nelle pagine web. Gli oppositori del software sostenevano che i commenti erano spesso volgari o diffamatori. Napster era un'applicazione software che permetteva la condivisione di file peer-to-peer. Gli utenti potevano condividere i file musicali aggirando i normali metodi di distribuzione, il che alla fine fu considerato una violazione delle leggi sul copyright (Ritholz, 2010).

La definizione più usata di Web 2.0 è di Tim O'Reilly: insieme dei servizi sopravvissuti al crollo borsistico del 2000 e di quelli nati successivamente. Il termine Web 2.0 diventa

da quel momento l'etichetta per tutto il mondo del social network e dei nuovi servizi e piattaforme in rete, con protagonisti prima Google e YouTube, poi la messaggistica istantanea e infine i social network veri e propri. Caratteristiche di questi social network sono multimedialità, facilità d'uso, possibilità per l'utente di inserire contenuti e renderli visibili, creazione di una reputazione on line indipendente dall'appartenenza a grandi marchi di comunicazione, ma costruita con successo quotidiano. Ciò permette la diffusione di una rete virtuale che mantiene gli individui costantemente in contatto. Il termine Web 2.0 è una parola che cerca di etichettare il fenomeno evidente e riconosciuto senza precisi confini. Lo stesso fenomeno è indicato attraverso varie definizioni incrociate che cercano di coglierne alcuni aspetti, senza la pretesa di essere esaustive.

Dal grafico sottostante rappresentato in Figura 1.9, abbiamo un'istantanea dell'utilizzo dei social network in Italia, comparando la fascia giovani (14-29 anni) e il totale della popolazione (fonte: Censis, 2022).

Come si può osservare dalla Figura, i 5 social network più utilizzati dal totale della popolazione sono WhatsApp, YouTube, Facebook, Instagram ed Amazon.

Entrando meglio nel dettaglio, comparando la fascia giovani con il totale della popolazione, i giovani utilizzano molto di più WhatsApp, YouTube, Instagram, TikTok e Spotify rispetto all'utilizzo di Facebook, che sta passando ad essere un social utilizzato prettamente da persone adulte.

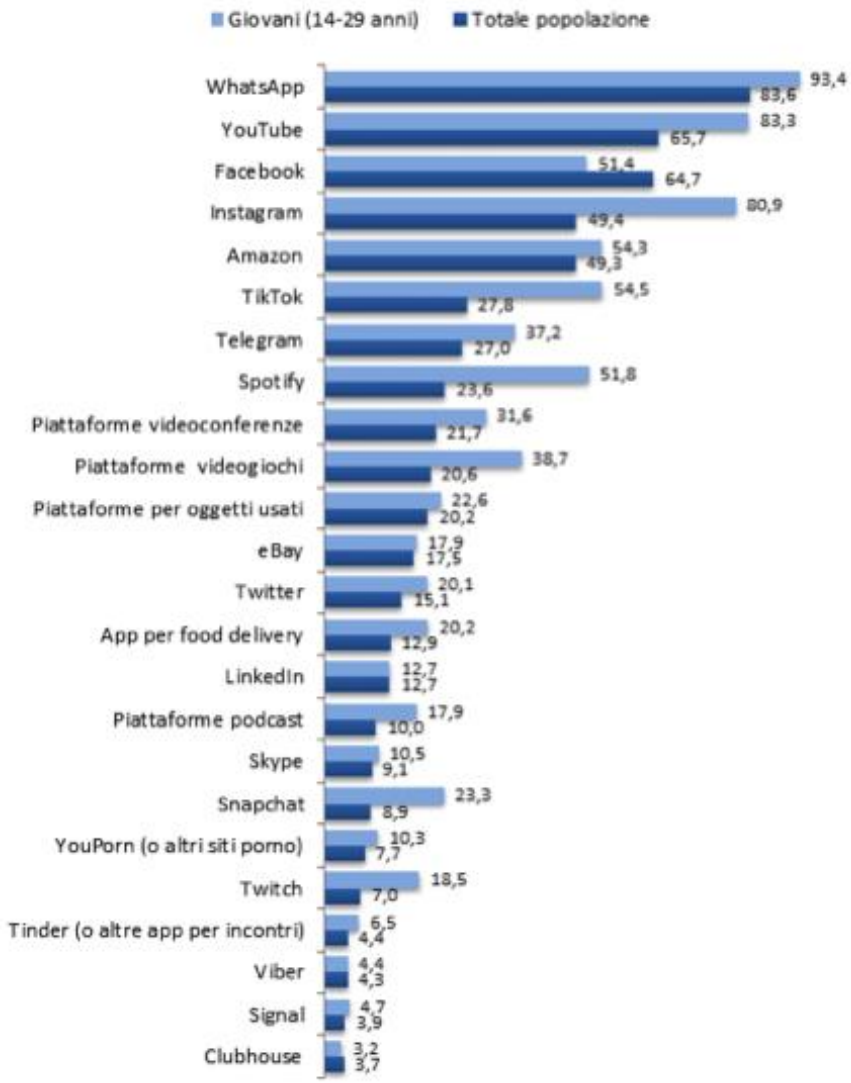

Figura 1.9 Utenza complessiva di social network, fonte: Censis, 2022

1.3 Definizione e caratteristiche dei social network

Nel 2000 i social media hanno ricevuto un grande impulso con la nascita di molti siti di social networking. Questo ha potenziato e trasformato l'interazione tra individui e organizzazioni che condividono interessi comuni in materia di musica, istruzione, film e amicizia, basandosi sul social networking. Tra i siti lanciati vi sono LunarStorm, six degrees, cyworld, ryze e Wikipedia. Nel 2001 sono stati lanciati fotolog, sky blog e Friendster, e nel 2003 MySpace, LinkedIn, lastFM, tribe.net, Hi5 ecc. Nel 2004 si sono sviluppati nomi popolari come Facebook Harvard, Dogster e Mixi. Nel 2005 sono emersi

nomi importanti come Yahoo!360, YouTube, cyword e Black planet (Junco, Heibergert, & Loken, 2011).

Differenza tra Social media e social network

Secondo Daniel Nations (2010), i social media sono difficili da definire e sono una strada a doppio senso che offre la possibilità di comunicare. Questo significa che un social media è uno strumento di comunicazione, proprio come qualsiasi altro social network? Esistono differenze tra questi due concetti?
I social media possono essere definiti una strategia e uno sbocco per la diffusione, mentre i social network sono uno strumento e un'utilità per connettersi con gli altri (Cohen, 2009; Stelzner, 2009). Inoltre, Cohen (2009) riferisce che "la differenza non è solo semantica, ma nelle caratteristiche e nelle funzioni inserite in questi siti web dai loro creatori, che dettano il modo in cui devono essere utilizzati".
In effetti, ci sono diverse differenze tra social media e social network (Hartshorn, 2010). La prima potrebbe essere la definizione: i social media sono ancora un mezzo di comunicazione utilizzato principalmente per trasmettere o condividere informazioni con un ampio pubblico, mentre il social networking è un atto di coinvolgimento in quanto le persone con interessi comuni si associano e costruiscono relazioni attraverso la comunità (Cohen, 2009; Hartshorn, 2010).

Di seguito verranno portati alcuni esempi di Social network e social media più influenti.

LunarStorm

LunarStorm accessibile all'indirizzo www.LunarStorm.se, è un sito virtuale commerciale disponibile in lingua svedese. In realtà, LunarStorm è nato nel 1996 ed è stato progettato da Rickard Ericsson. Si trattava di un sito web di social networking per adolescenti ed è stata la prima comunità digitale online europea. LunarStorm è stato lanciato ufficialmente nel 2000.

Nel 2001, LunarStorm era cresciuta fino a raggiungere oltre 600.000 membri, ma continuava a incontrare difficoltà economiche. Fin dall'inizio, LunarStorm è stato finanziato da banner e altre pubblicità sul sito web, ma presto si è evoluto fino a includere servizi a pagamento via SMS. Un primo esempio fu la carta prepagata "Vrål" ("Bawl") di

LunarStorm. Nel 2002 è stata introdotta "Kolla" ("Guarda" o "Controlla"), che ha permesso agli utenti di visitare LunarStorm dai loro telefoni cellulari. Nello stesso anno, i membri hanno potuto passare allo status di "pro" e ottenere un accesso illimitato a una serie di servizi a pagamento. LunarStorm Pro era estremamente popolare tra i membri e ha migliorato notevolmente la situazione economica del sito (Goma, 2001).

Questo social network si è diffuso non solo in Svezia, bensì anche nel Regno Unito fino al 2007. LunarStorm è stato abbandonato nel 2010, in quanto il numero di utenti è sceso da 600 mila a solamente 1500. Raffigurazione del Log In di LunarStorm in figura 1.10.

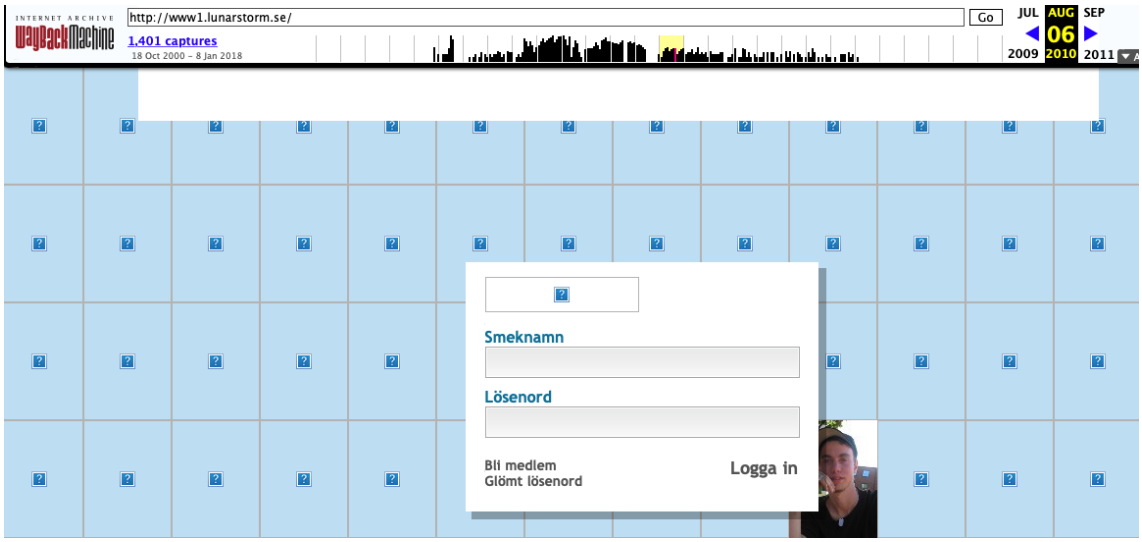

Figura 1.10 Immagine di Lunar Storm, fonte: web

MYSPACE

La nuova generazione (quella della fase espressiva) dei social network comincia con la nascita di MySpace nel 2003 quando Tom Anderson e Chris DeWolfe, i suoi fondatori, crearono questo sistema di condivisione di file audio e video, oltre al servizio di comunicazione tra utenti con i messaggi, già ampiamente sperimentato con i precedenti network.

La possibilità di personalizzare il proprio profilo con suoni e clip video fu il motivo che decretò la fortuna del social network. Infatti, molte band emergenti e cantanti sconosciuti scelsero MySpace come supporto mediatico per pubblicizzare i propri lavori. Nel 2005 venne acquistato dalla News Corporation di Rupert Murdoch per 580 milioni di dollari e

ampliò la possibilità d'espressione e comunicazione degli utenti includendo nuove funzioni come l'invio dei messaggi con i telefoni cellulari e l'uso di alcuni gadget virtuali predefiniti (i widget), che potevano essere usati per abbellire la pagina del profilo personale, offrendo in questo modo un diverso modo di presentarsi all'utenza.

Il sito di MySpace è ancora utilizzabile al giorno d'oggi, è stato il primo sito a raggiungere l'audience globale, comportando un'influenza sulla musica e la cultura pop.

Tuttavia la personalizzazione del profilo di ciascun utente, ha comportato la difficoltà della gestione per i creatori del social, in quanto il caricamento di file mp3 ha reso non scorrevole l'utilizzo di MySpace.

Rappresentazione della home di MySpace in figura 1.11

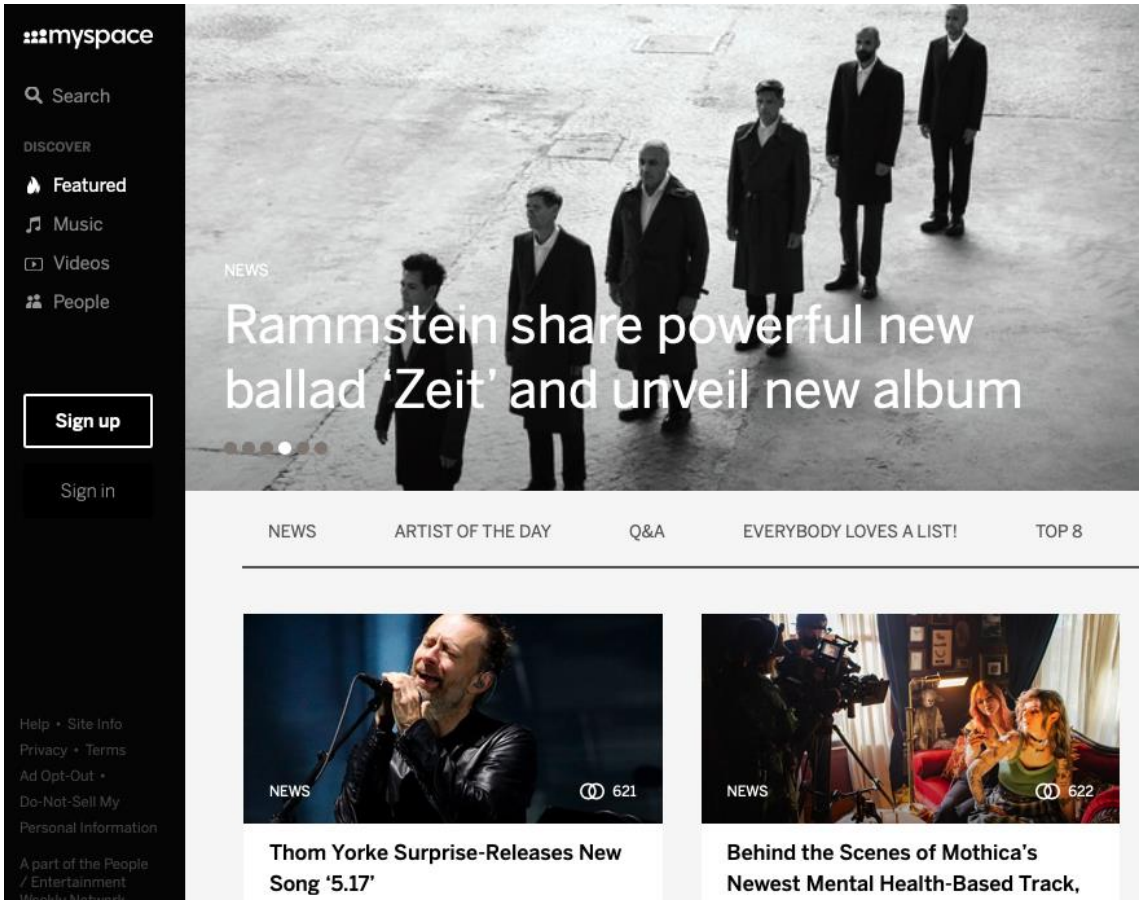

Figura 1.11 Immagine di My Space, fonte: web

Facebook

Facebook è un sito web di social networking lanciato nel febbraio 2004 e gestito privatamente da Facebook, Inc. (Facebook, 2004). Facebook è stato fondato da Mark Zuckerberg e altri quando era uno studente di Harvard; tuttavia, quando il sito è stato lanciato inizialmente, era riservato solo agli studenti di Harvard.

In seguito, il privilegio è stato esteso agli studenti delle scuole superiori e poi a tutti coloro che avevano almeno 13 anni (Boyd, 2007). Nel gennaio 2009, Facebook è stato classificato come il social network più utilizzato al mondo. Inoltre, nel maggio 2010, Google ha annunciato che più persone hanno visitato Facebook di qualsiasi altro sito web al mondo. Dichiara che questo dato è stato scoperto grazie ai risultati ottenuti su 1.000 siti in tutto il mondo. (TIMES, 2010). A luglio 2010, Facebook conta più di 500 milioni di utenti attivi.

Gli utenti possono creare un profilo personale, aggiungere altri utenti come amici e scambiare messaggi, comprese notifiche automatiche, foto e commenti quando aggiornano il loro profilo. Inoltre, gli utenti di Facebook possono unirsi a gruppi di utenti di interesse comune, organizzati per luogo di lavoro, scuola, università o altre caratteristiche. Facebook consente a chiunque abbia almeno 13 anni di diventare un utente registrato del sito.

Ogni giorno il traffico verso la rete Facebook è in aumento. Facebook è diventato anche il primo social network in otto mercati asiatici: Filippine, Australia, Indonesia, Malesia, Singapore, Nuova Zelanda, Hong Kong e Vietnam. Il 24 ottobre 2007, Microsoft ha annunciato di aver acquistato una quota dell'1,6% di Facebook per 240 milioni di dollari, dando a Facebook un valore implicito totale di circa 15 dollari.

L'acquisto da parte di Microsoft comprendeva i diritti di inserire annunci internazionali su Facebook; altre aziende hanno seguito lo stesso esempio (STONE, 2007). Per esempio, proprio durante la Coppa del Mondo di calcio FIFA 2010, Nike ha fatto un annuncio con Facebook e, in pochi minuti, una media di 8 milioni di spettatori si è registrata su Facebook (kevthefont, 2010).

Rappresentazione della copertina di Facebook in Figura 1.12.

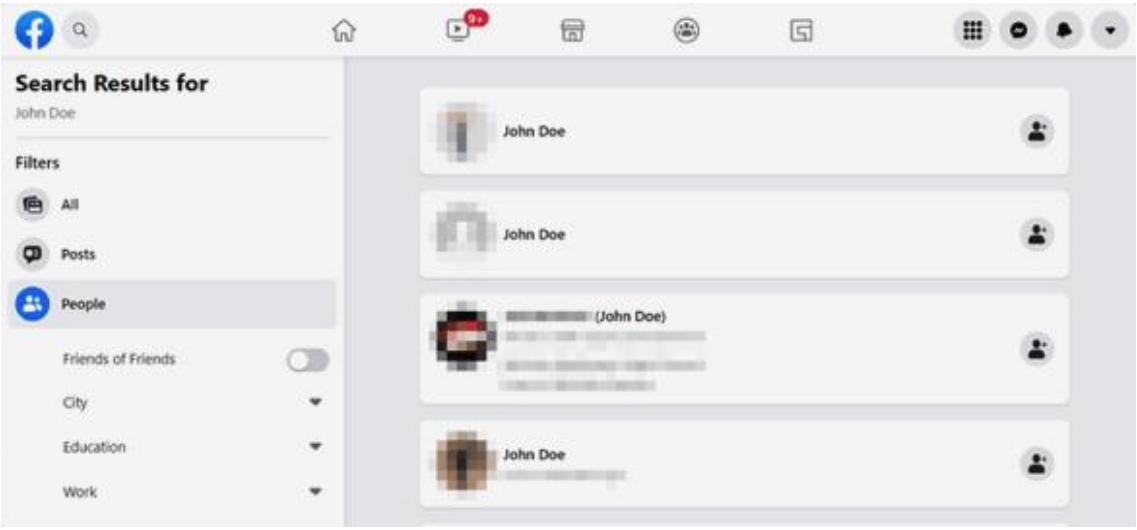

Figura 1.12 Immagine copertina Facebook, fonte: web

Dalla rappresentazione sottostante, Figura 1.13, si può osservare quanti sono i milioni di utenti nel 2022 secondo Statista: L'India è il primo paese con quasi 330 milioni di utenti, seguito da Stati Uniti con 180 milioni ed Indonesia con 130 milioni.

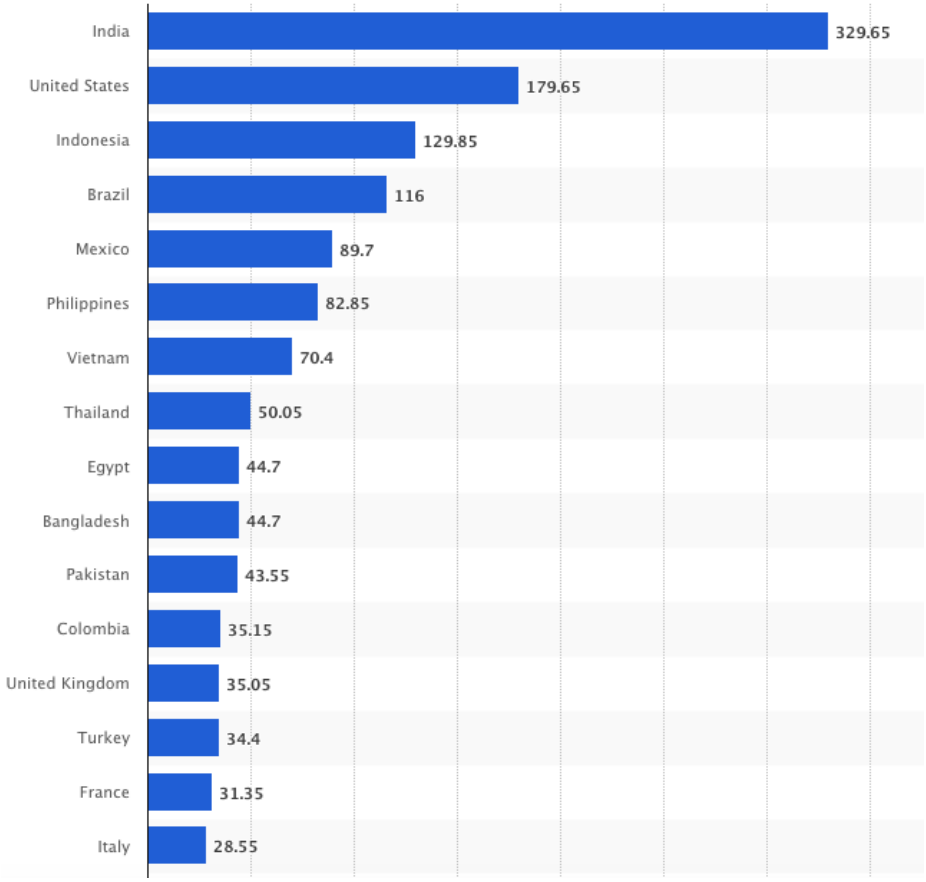

Figura 1.13 Utenti Facebook (in milioni) a gennaio 2022, fonte: Statista

Dalla rappresentazione sottostante, Figura 1.14, si può osservare come i ricavi e gli utili di Facebook sono incrementati esponenzialmente dal 2018 in poi.

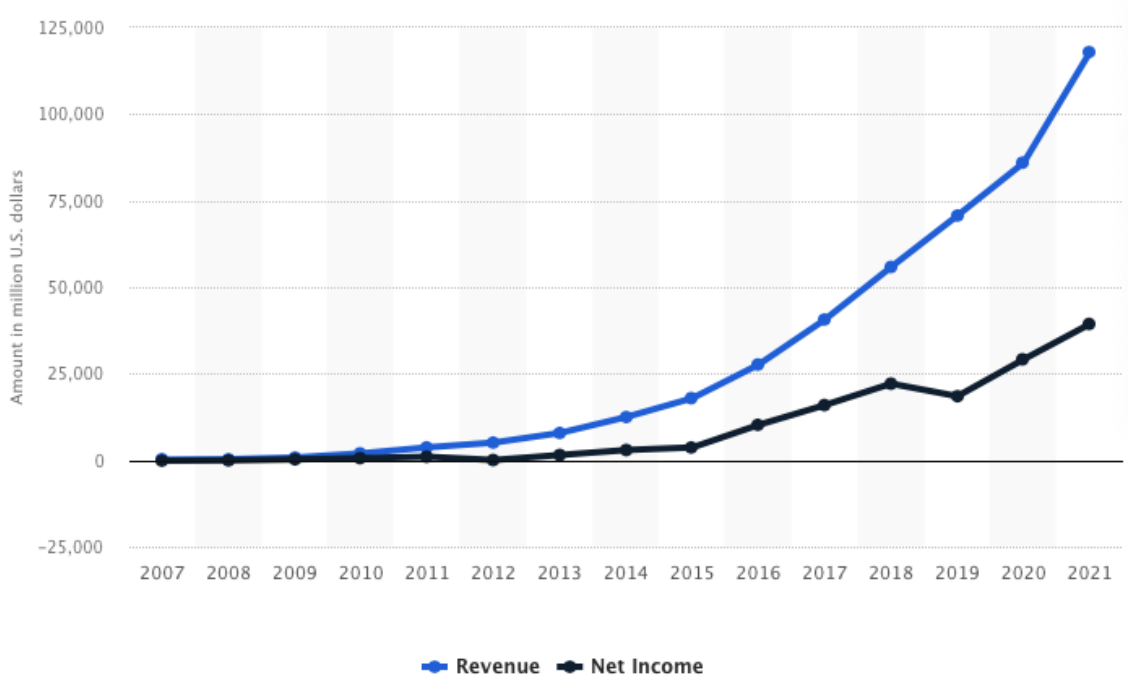

Figura 1.14 Fatturato e ricavi di Facebook 2007-2021, fonte: Statista

YouTube

YouTube, fondato nel 2005, è la comunità video online più popolare al mondo, dove milioni di persone possono scoprire, guardare e condividere video originali (YouTube, 2005). YouTube offre alle persone un forum per connettersi, informare e ispirare altre persone in tutto il mondo e funge da importante piattaforma di distribuzione per i creatori di contenuti originali e gli inserzionisti, grandi e piccoli. YouTube ha sede a San Bruno, in California, e utilizza la tecnologia Adobe Flash Video per visualizzare un'ampia varietà di contenuti video generati dagli utenti, tra cui filmati, clip televisive e video musicali, oltre a contenuti amatoriali come video blogging e brevi video originali. Nel novembre 2006, a un anno dal lancio, YouTube è stato acquistato da Google Inc. in una delle acquisizioni più chiacchierate di sempre. YouTube ha stretto una serie di rapporti di partnership con fornitori di contenuti come CBS, BBC, Universal Music Group, Sony Music Group, Warner Music Group, NBA, The Sundance Channel e molti altri (YouTube, 2005). YouTube ha offerto al pubblico una versione beta del sito nel maggio 2005, sei mesi prima del lancio ufficiale nel novembre 2005. Il sito è cresciuto

rapidamente e nel luglio 2006 la società ha annunciato che ogni giorno venivano caricati più di 65.000 nuovi video e che il sito riceveva 100 milioni di visualizzazioni al giorno (YouTube, 2005).

Rappresentazione grafica della schermata home di Youtube in Figura 1.15

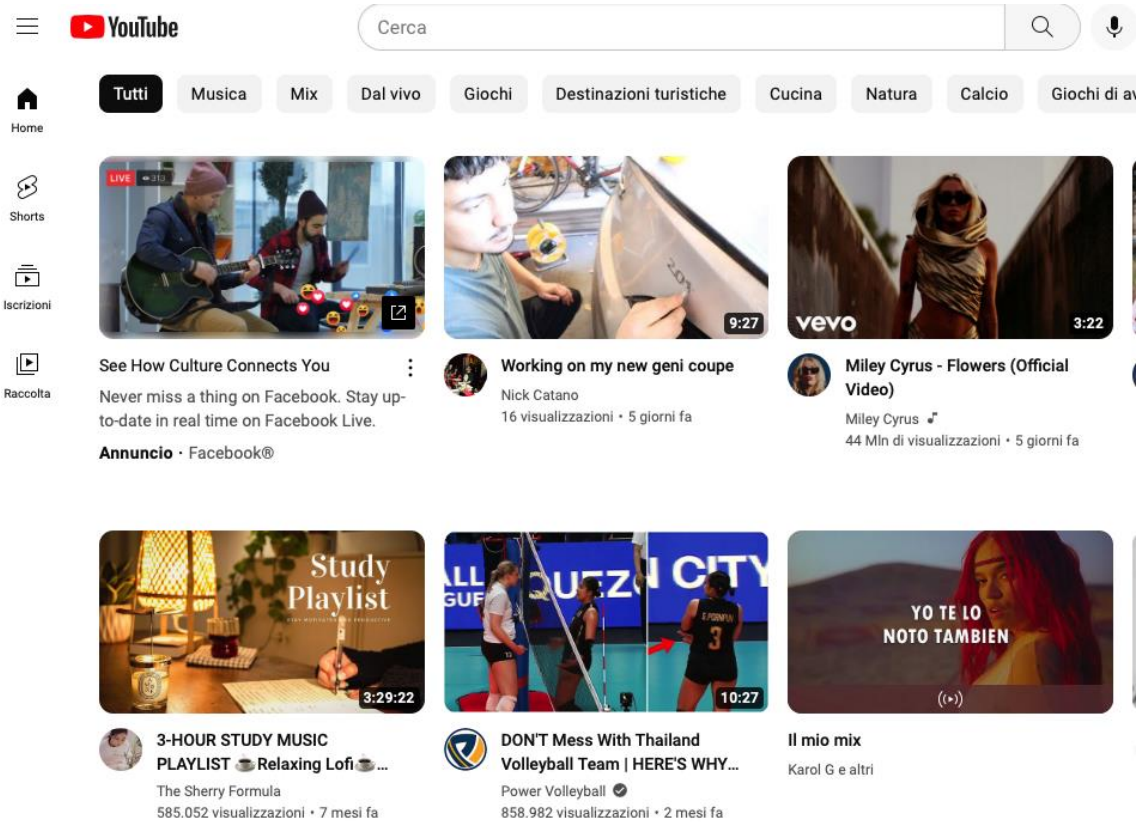

Figura 1.15 Immagine di YouTube, fonte: web

Dalla rappresentazione in Figura 1.16, si può notare come YouTube abbia incrementato il fatturato nella finestra temporale dal 2017 al 2021.

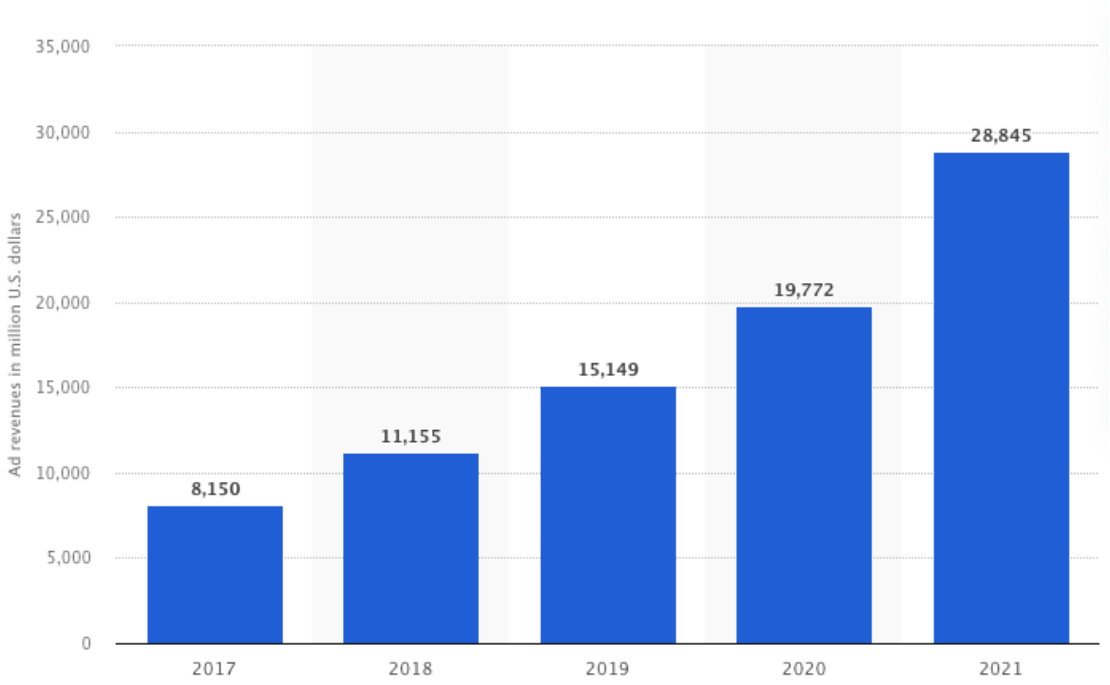

Figura 1.16 Fatturato di Youtube in milioni di USD, fonte: Statista

Twitter

All'inizio non fu progettato per essere un social network, infatti il suo ideatore Jack Dorsey lo pensò per farlo funzionare come piattaforma di comunicazione per dispositivi mobili. Col passare del tempo, però, la sua sopravvivenza dipese sempre più dagli adattamenti informatici dei suoi programmatori, che per farlo diventare più competitivo tra i vari sistemi di comunicazione, lo fecero diventare un vero e proprio social network. Più precisamente Twitter è un servizio di microblogging, cioè un network basato sullo scambio di messaggi brevissimi, più corti di un Sms (Short Message Service).

Nacque nel 2006 dall'idea di J. Dorsey e altri due colleghi, Evan Williams e Biz Stone, che avevano già realizzato insieme una piattaforma di comunicazione per blog. Cominciarono a sviluppare un software per pubblicare messaggi più brevi di un sms. Il gruppo di programmatori si dedicò intensamente allo sviluppo del servizio con l'intento di adattarlo all'uso del cellulare, ma nei primi mesi dalla sua pubblicazione non ebbe grande successo, poiché veniva utilizzato solo da una nicchia di appassionati intorno all'area di San Francisco.

Nel 2007 diventò sempre più simile a un social network e si diffuse in tutto il mondo. Il suo nome deriva dalla definizione della parola inglese “twitter”, che significa “cinguettare”. Lo scambio delle informazioni su questo social network avviene tramite i “tweet” (cinguettii), dei messaggi molto brevi (massimo 140 caratteri), come cinguettii di uccelli appunto, che sono la caratteristica principale del network. Furono gli utenti stessi ad inventare il “Retweet” (ri-messaggio), un modo di segnalare e riproporre i messaggi scritti da altri utenti, premettendo al testo le lettere “RT” seguite dal nome dell’autore. In seguito, il Retweet è diventata una funzionalità supportata dal social network facendo aumentare notevolmente l’interazione tra i suoi utenti. Successivamente la diffusione di Twitter fu vertiginosa: passò da 105 milioni di utenti dell’aprile 2011 ai 200 milioni alla fine dello stesso anno. Se nel 2006 solo Twitter consentiva di accedere e utilizzare un social network mediante telefono cellulare, oggi questo è possibile con tutti i principali networks. Rappresentazione della schermata home di Twitter in Figura 1.17.

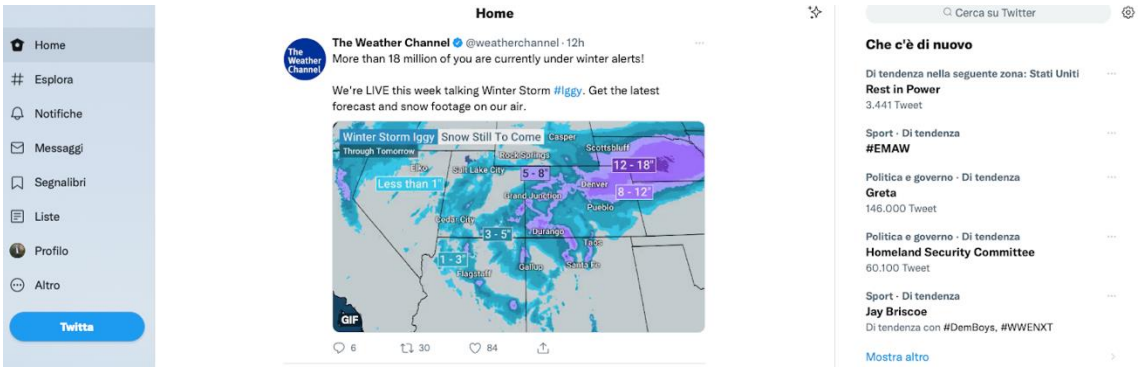

Figura 1.17 Immagine copertina di Twitter, fonte: web

La rappresentazione sottostante in figura 1.18 raffigura la percentuale di ricavi che si dividono tra annunci pubblicitari e licenze di dati dal 2010 al 2021.

Come si può notare, i ricavi per Twitter nel 2010 erano prevalentemente derivati da licenze di dati (74% contro il 26 per gli annunci pubblicitari), mentre ora la situazione si è capovolta, nel 202 l’89% dei ricavi per Twitter derivano da annunci pubblicitari.

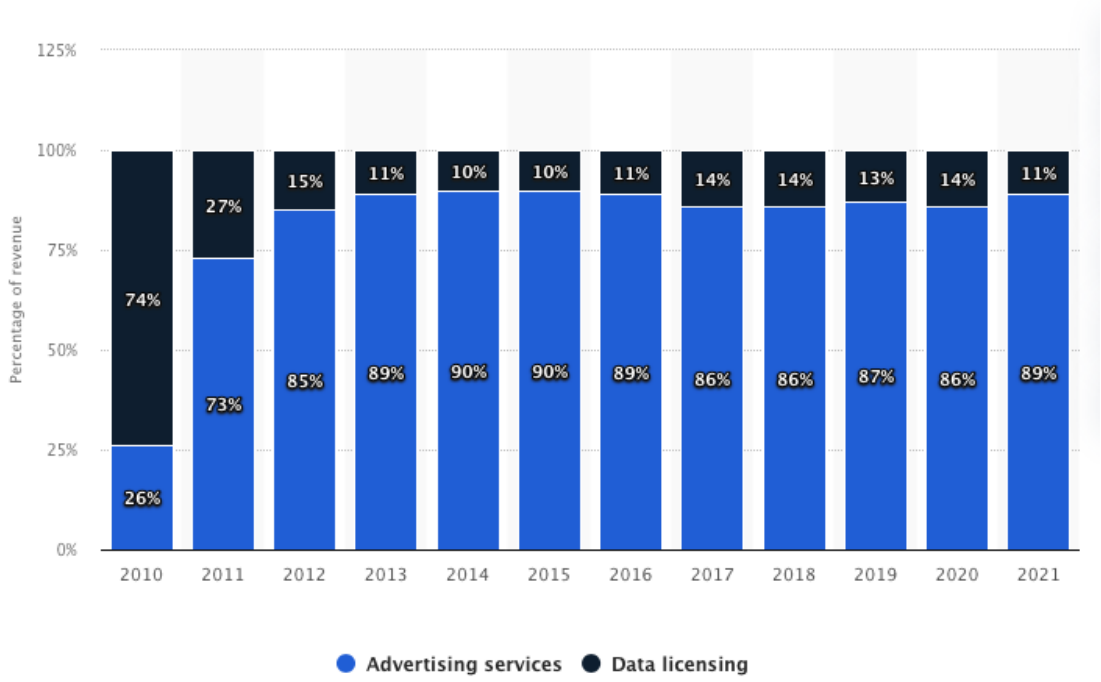

Figura 1.18 Twitter ricavi annuali 2010-2021, fonte: Statista

La figura 1.19 rappresenta l'arco temporale dell'ammontare degli utenti giornalieri di Twitter nel mondo dal 2017 al secondo trimestre del 2022. È interessante vedere che dal terzo trimestre del 2020 ci sia crescita costante fino a metà 2022, superando i 200 milioni di utenti giornalieri dal secondo trimestre del 2021.

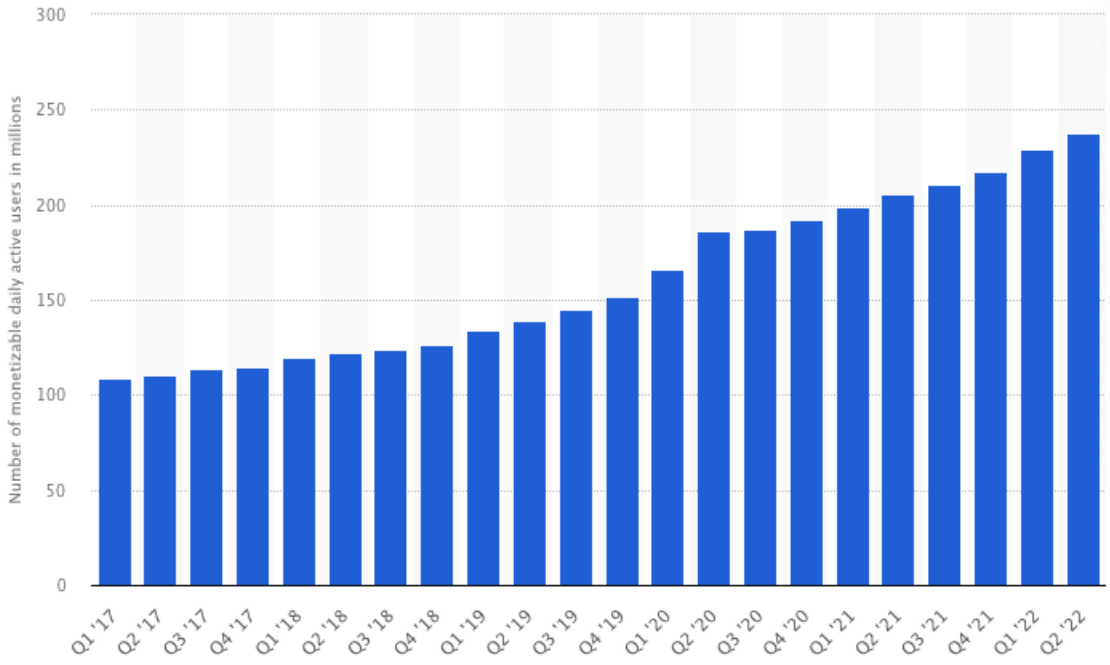

Figura 1.19 Utenti giornalieri Twitter worldwide 2017-2022, fonte: Statista

PINTEREST

Pinterest, creato nel 2010 da Paul Sciarra, Evan Sharp e Ben Silbermann (Wikipedia), è uno strumento di bookmarking visuale che aiuta gli utenti a scoprire e salvare le idee creative. Gli utenti inseriscono le immagini nelle bacheche, che sono raccolte curate intorno a temi o argomenti particolari. Questo grafo di immagini e bacheche curato dall'utente contiene una ricca serie di informazioni sulle immagini e sulle loro relazioni semantiche reciproche. Ad esempio, quando un'immagine viene appuntata su una bacheca, implica un "collegamento curatoriale" tra la nuova bacheca e tutte le altre in cui l'immagine compare. I metadati, come le annotazioni sull'immagine, possono essere propagati attraverso questi collegamenti per formare una ricca descrizione dell'immagine, della bacheca e degli utenti.

Poiché l'immagine è il fulcro di ogni pin (cioè di ogni contenuto postato su Pinterest), le caratteristiche visive giocano un ruolo importante nel trovare contenuti interessanti, stimolanti e rilevanti per gli utenti. Rappresentazione del feed di Pinterest in figura 1.20.

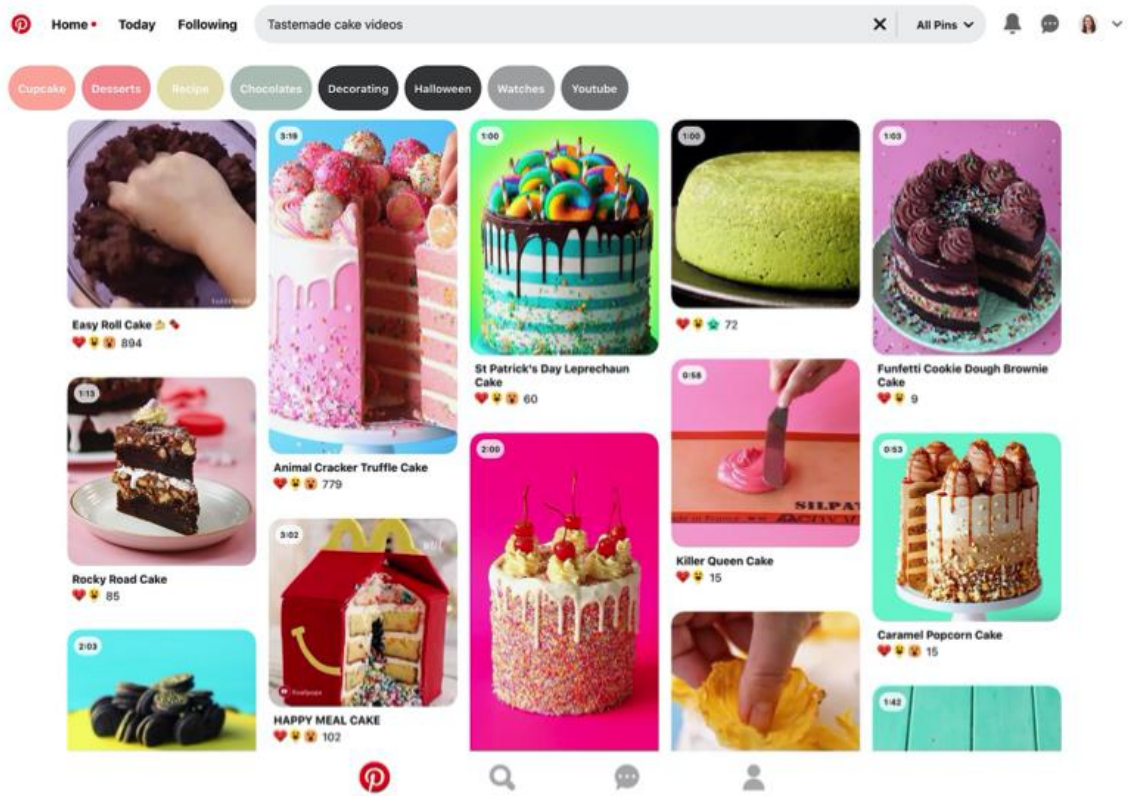

Figura 1.20 Immagine copertina di Pinterest, fonte: web

Nel 2022, secondo Statista, il paese con maggior utenti attivi su Pinterest sono gli Stati Uniti con 86 milioni, seguito da Brasile (27 milioni) e Messico (18 milioni). L'Italia si trova all'ottavo posto con 8.5 milioni di utenti (fonte: Statista, 2022).

Instagram

Instagram è un Social Network nato nel 2010 ideato da Kevin Systrom, Mike Krieger per la pubblicazione di foto e video. Chiunque lo utilizza cerca di mostrare la propria persona decidendo di apparire con una versione di sé stesso più o meno veritiera rispetto alla vita reale.

Il sociologo Erving Goffman (1959) esaminò le relazioni e le interazioni sociali, i comportamenti ricorrenti arrivando così ad utilizzare la metafora del teatro per descrivere le interazioni. I due luoghi immaginari di cui si parla sono la ribalta e il retroscena, descrivendo il momento della ribalta come il momento in cui si compie l'interazione in sé, quando l'attore recita la propria parte sul palcoscenico; al contrario il retroscena è il momento in cui si organizza la recita teatrale, in cui si producono i repertori.

Questo è proprio ciò che è presente dietro ad una pubblicazione di una fotografia su Instagram o su qualsiasi altro social network: la pubblicazione della foto più bella della giornata o del momento più divertente è il momento della ribalta, che i followers vedranno, commenteranno e a metteranno likes, il momento del retroscena è invece il tempo trascorso a scattare la foto, a scegliere la migliore, a decidere la caption ecc.

Instagram ha già da qualche anno aggiunto una nuova funzione, le IG stories: quest'ultime sono foto o video che vengono pubblicate ma che al contrario del post hanno durata esclusivamente di ventiquattro ore. Le IG stories possono essere commentate dai followers attraverso reazioni (date in automatico da Instagram) oppure attraverso messaggi privati. Data la loro durata predeterminata, danno modo di condividere la propria vita in tempo reale o non, con meno "serietà" di quanta ne abbia una foto/video pubblicato sul profilo che rimarrà lì fino a quando il proprietario dell'account non deciderà di cancellarlo.

Questo binomio ribalta/retroscena è stato però un po' allentato su Instagram con l'introduzione della possibilità di aggiungere persone ad una lista di amici stretti (Close friends).Se, da una parte, l'utente ha la possibilità di condividere stories solamente con questa lista di persone, dall'altra sempre più persone hanno iniziato a creare profili spam, cioè profili privati in cui si hanno pochi followers e si pubblicano parti di retroscena, momenti di vita quotidiana in cui non ci si deve mettere nella miglior luce possibile, anzi, sono perlopiù utilizzati per far ridere, per pubblicare tutto ciò che non verrebbe mai pubblicato su un profilo pubblico. Rappresentazione di un feed di Instagram nella figura 1.21

Figura 1.21 Immagine feed di Instagram, fonte: web

Nella rappresentazione in figura 1.22, possiamo vedere come gli utenti mensili su Instagram siano cresciuti del doppio dal 2018 a dicembre 2021. Probabilmente questo cambiamento è dovuto dalla migrazione da Facebook verso Instagram per gli utenti di fascia più giovane.

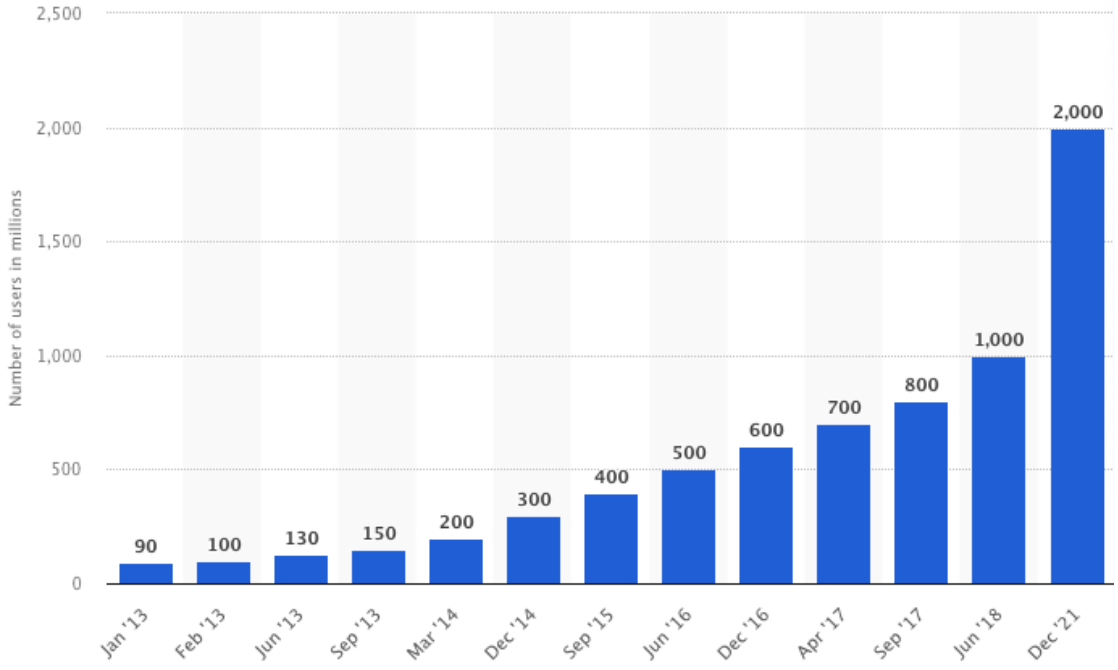

Figura 1.22 Utenti mensili di Instagram 2013-2021, fonte: Statista

LINKEDIN

LinkedIn è il frutto del lavoro di Reid Hoffman e di tre suoi colleghi ed è stato fondato nel 2022.

È una piattaforma sociale dedicata ai professionisti, per dar loro occasione di incontrarsi e discutere di lavoro e business in rete, di pubblicare i loro curricula, i loro progetti ed eventualmente trovare collaboratori. L'obiettivo del social network è quello di creare una rete di professionisti e aziende, in cui avviene uno scambio reciproco di contatti e informazioni.

Gli utenti dopo la registrazione (gratuita) possono pubblicare e aggiornare il curriculum vitae e una lettera di presentazione; possono leggere gli annunci di lavoro organizzati per settori, iscriversi a gruppi e seguire i profili delle imprese di proprio interesse. Il funzionamento di LinkedIn è strutturato secondo tre livelli diversi. Ogni livello è contraddistinto dalle competenze attuali di ogni utente, dalle esperienze lavorative, dall'evoluzione del loro curriculum e dal tipo di contatti che si hanno con gli altri utenti nel network. Così le aziende hanno a disposizione strumenti utili per scegliere i candidati più adatti alle loro aspettative e possono selezionare meglio il personale, mentre i

candidati possono cercare i lavori che più si addicono alle loro competenze professionali. Ogni utente può allargare la propria rete cercando nuovi contatti che potrebbero essere utili ai fini professionali.

LinkedIn oggi è presente in 200 paesi raggiungendo più di 200 milioni di iscritti ed è disponibile in 11 lingue: francese, giapponese, inglese, italiano, portoghese, rumeno, russo, spagnolo, svedese, tedesco e turco. Il nostro Paese è quello che in Europa fa registrare il maggiore tasso di crescita. Infatti nel 2011 è stata aperta una sede del social network a Milano. Con una capitalizzazione di circa 18 miliardi di dollari, nello stesso anno, è stato quotato in borsa ed ha assunto la leadership dei social network professionali.

Rappresentazione della pagina di Linkedin in Figura 1.23.

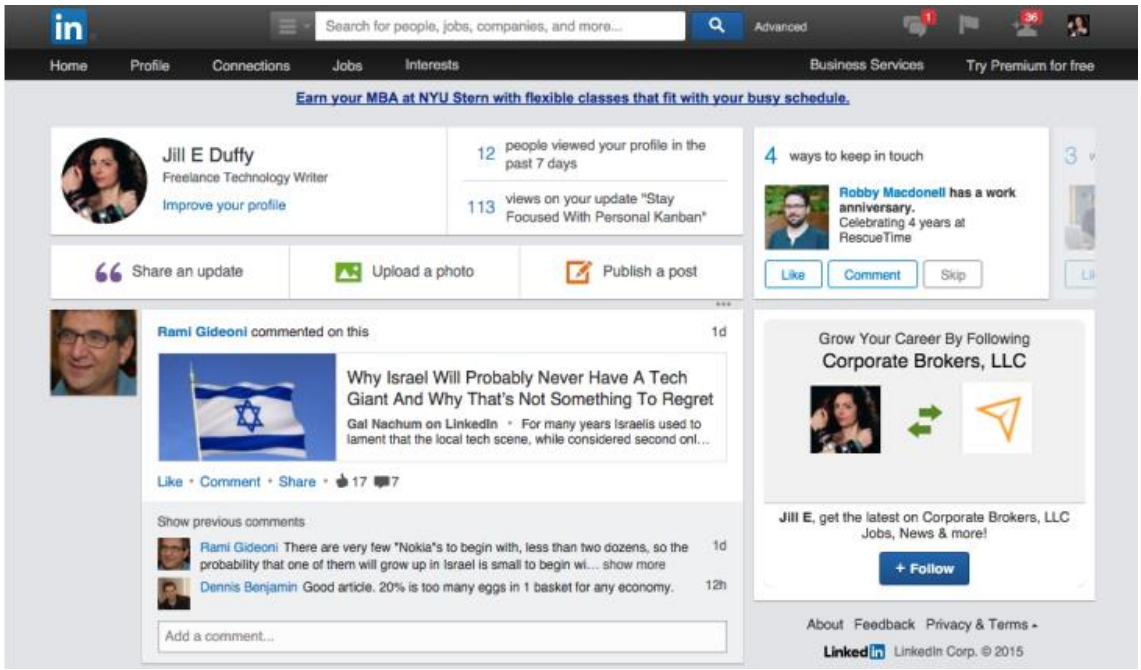

Figura 1.23 Immagine copertina di LinkedIn, fonte: web

Secondo Statista, i ricavi di LinkedIn sono cresciuti dal 2017 al 2022, da 2,27 miliardi a 13,8 miliardi nel 2022.

TIKTOK

TikTok, conosciuta come DouYin in Cina, è un'applicazione di social media molto popolare fra generazione Z e millennial.

TikTok com'è ora nasce dall'unione di due app: Musical.ly e TikTok. Musical.ly, rilasciata nel 2014, fu progettata per essere una piattaforma di sincronizzazione labiale o lip-syncing: gli utenti potevano selezionare una canzone e diverse opzioni di velocità per registrare un breve video in cui cantavano in playback e/o ballavano. Nel settembre 2016 dalla società cinese ByteDance venne rilasciata DouYin o TikTok, un'altra app per la creazione di video con sincronizzazione labiale. ByteDance decise di provare ad acquistare Musical.ly per ottenere l'esperienza del team nel gestire un'app all'estero e per acquisire decine di milioni di utenti internazionali. Alla fine del 2017, ByteDance raggiunse un accordo e acquisì l'app per 800 milioni di dollari, accedendo così agli oltre 60 milioni di utenti attivi mensilmente di Musical.ly negli USA e in Europa.

Viene utilizzata per creare e condividere brevi video. Le clip variano dai 15 ai 60 secondi (da luglio 2021 possono arrivare a 3 minuti) in cui gli utenti possono ballare, improvvisare scenette divertenti, sincronizzare il labiale in playback a delle registrazioni, mostrare la propria giornata o condividere un proprio talento. Gli utenti possono anche partecipare a sfide o "challenge" e duettare i video degli altri utenti.

Nel primo quadrimestre del 2021 TikTok è la app più scaricata con più di 58 milioni di download superando Facebook, Youtube, Instagram e WhatsApp (SensorTower, Chan, 2021).

Secondo Cloudflare, una compagnia che si occupa di web security e performance, TikTok risulta essere il sito web più popolare del 2021 e ha superato anche Facebook come social network più popolare. A settembre 2021, inoltre, TikTok ha superato un miliardo di utenti attivi mensili.

In Italia, secondo le stime più recenti di WeAreSocial, il 23,9% degli utenti di età compresa fra i 16 e i 64 anni utilizza TikTok. L'anno precedente la percentuale era la metà.

Qui sotto la rappresentazione del feed di TikTok in figura 1.24.

Figura 1. 24 Immagine copertina TikTok, fonte: web

Nella figura sottostante (Figura 1.25) si può notare come il download di TikTok sia esploso durante lo scoppio della Pandemia nel 2020, fino a diminuire nel anno successivo.

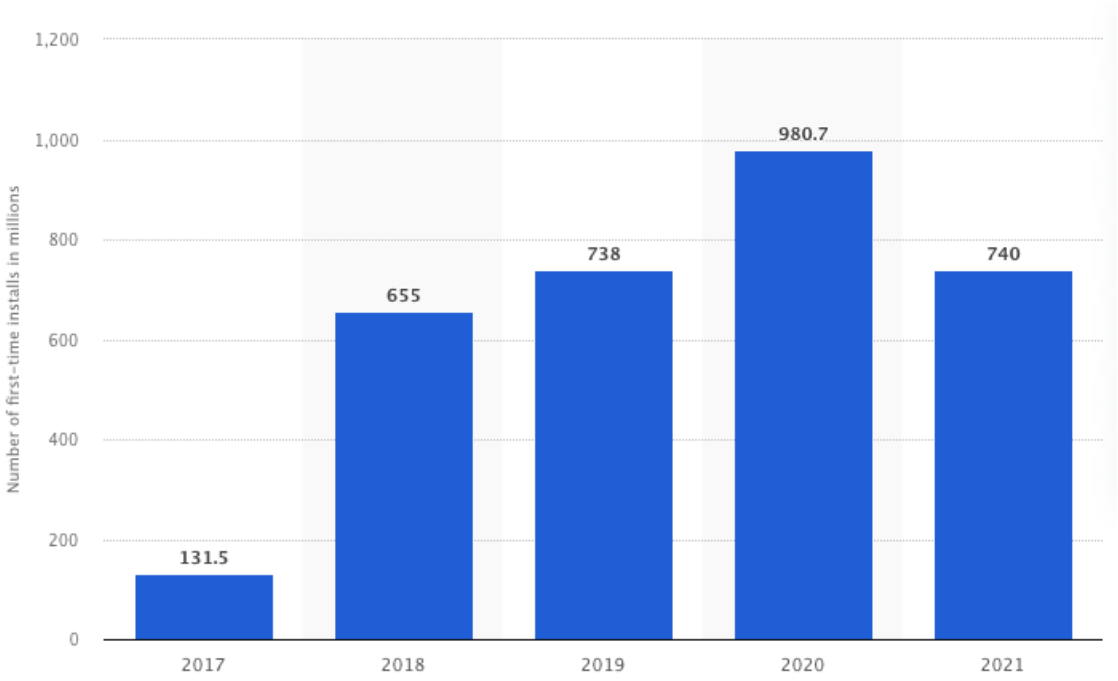

Figura 1.25 Numero di installazioni Tik Tok 2017-2021 fonte: Statista

CAPITOLO 2: I social media e gli effetti su società, educazione, business e politica

Dall'emergere di siti di social networking come Twitter e Facebook, i giornalisti e le loro organizzazioni hanno dato attenzione a queste piattaforme come strumento di veicolazione di notizie. Questi siti sono diventati una routine quotidiana per le persone, in quanto spesso vengono utilizzati durante l'arco della giornata, dalla mattina durante la colazione, alla sera prima di dormire. I social media sono stati principalmente definiti per riferirsi a "i molti strumenti elettronici relativamente poco costosi e ampiamente accessibili che facilitano a chiunque di pubblicare e accedere alle informazioni, collaborare a uno sforzo comune o costruire relazioni"

In circa un decennio, con il proprio utilizzo si sono affermati, moltiplicati e diffusi in tutti il mondo, creando abitudini, modi di comunicare e tendenze della "nuova generazione globale".

I social media hanno accelerato l'aumento delle relazioni nella vita degli individui, fenomeno già in corso da decenni. Ciò ha plasmato la dimensione comunicativa delle nostre vite. Ma questo nuovo modo di comunicare, sempre attivo, comporta stress, fatica e riduzione del tempo a disposizione. Ma il ruolo di questi media rende più tollerabile tale perdita. Al tempo stesso i social media sono qualcosa di enorme in termini di mercato, lotta politica, trasformazione delle logiche culturali e produttive. Ogni politico attuale deve fare i conti con la diffusione virali dei messaggi sui social media.

Molte tesi sui social media sottolineano la differenza tra vecchio e nuovo: il cosiddetto fenomeno del "tecno- entusiasmo" (Colombo, 2013).

Quest'ultimo risale alla rivoluzione industriale e alla prima fase delle scoperte tecnologiche moderne. Solo che l'accelerazione dell'innovazione prodotta dal digitale comporta anche una rapida perdita di interesse verso fenomeni nuovi, visto che essi tendono a diventare immediatamente vecchi.

Un effetto secondario di questa attitudine è l'alternanza fra ottimismo e pessimismo, tra entusiasmo e critica. L'approccio pessimistico richiama l'idea della net-delusion di Morozov, cioè la convinzione che il momento liberatorio, esplosivo dell'invenzione sia stato seguito da un percorso di normalizzazione.

Ciò è inevitabile perchè i social media non sono solo un fatto tecnologico ma un incrocio di dimensioni: economia, politica, organizzazione.

Secondo Castells: "una società in rete è una società in cui la struttura sociale ruota intorno alle reti attivate da tecnologie dell'informazione e della comunicazione elaborate

digitalmente", quindi ogni aspetto della società è toccato dai social media (Colombo, 2013).

2.1 Socialità: socievolezza e potere

Per un'analisi dei social media è opportuno prendere in considerazione il loro aspetto sociale, che in parte è evidente: i media sono sempre parte di una società in cui svolgono una pluralità di funzioni di intermediazione.

Molti, tra cui (Tim Berners-Lee, 1999), ritengono il web non solo un'innovazione tecnologica ma anche sociale. Infatti, le tecnologie digitali e le loro applicazioni hanno costituito oggetto privilegiato di studio per gli studiosi del Social Shaping of Technology, da sempre interessati ai processi di domesticazione, ossia integrazione di un uso tecnologico nella vita sociale degli utenti.
I media della convergenza digitale sembrano pensati per abilitare collaborazioni partecipative, cioè per innescare comunicazioni orizzontali, dal basso, anziché meccanismi organizzativi di tipo tradizionale. Anche a livello aziendali i social media appaiono come strumenti di ripensamento dei tradizionali schemi di decisione, gestione, innovazione. È evidente che la novità della rete è di essere un medium in cui "la caratteristica più evidente sono le persone".
Quando Castells parla di *self mass communication*, egli riconosce la liberazione di risorse individuali sulla scena dei media, e un'altra definizione come quella di social casting, in riferimento alla modalità di trasmissione del web sociale e partecipativo, insiste sulla stessa natura collaborativa di base. Ma bisogna fare delle precisazioni: il termine "sociale" non richiama la complessità della società, ma una delle sue dimensioni, quella della socievolezza caratterizzata dal piacere umano di stare insieme senza obiettivi determinati e funzionali, la talkative society (Dalghren, 2002).

2.2 Impatto dei social media sui giovani

Al giorno d'oggi i social media sono diventati un nuovo insieme di strumenti per coinvolgere i giovani. La vita quotidiana di molti giovani è ramificata dai social media. I giovani conversano e comunicano con i loro amici e gruppi utilizzando media e dispositivi diversi ogni giorno.

Negli anni passati i ragazzi erano in contatto solo con gli amici e i loro gruppi nelle scuole e nelle università. Ma al giorno d'oggi i giovani sono in contatto non solo con amici conosciuti ma anche con persone sconosciute attraverso siti di social network e di messaggistica istantanea. Secondo una ricerca condotta dalla BBC del 2013, viene affermato che il 67% degli utenti di Facebook è composto da giovani e studenti; quindi, questi lodano il fatto che i giovani e gli studenti abbiano più attenzione e relazione.

Secondo un sondaggio pubblicato dall' International Journal of Computer Applications Technology and Research prendendo come campione degli studenti in India, il 90% degli studenti universitari utilizza i social network (Siddiqui, Singh, 2016).
Si può utilizzare il telefono per accedere ai social network sempre e ovunque, poiché questi gadget includono computer tascabili, laptop, iPad e persino semplici telefoni cellulari (che supportano Internet).
L'articolo di Siddiqui e Singh elenca dei vantaggi e svantaggi dell'utilizzo dei social media per l'istruzione dei giovani,
Ai fini dell'educazione i social media sono stati usati come un modo innovativo.
Inoltre, potrebbero essere utilizzati per capire come sia importante il rispetto delle altre persone all'interno di altre piattaforme, evitando commenti aggressivi.
I social media hanno aumentato la qualità e il tasso di collaborazione per gli studenti. Con l'aiuto dei social media gli studenti possono facilmente comunicare o condividere informazioni rapidamente con ciascuno attraverso vari siti social come Facebook e Instagram.
Sempre secondo l'articolo "Social Media its Impact with Positive and Negative Aspects" di Siddiqui, Singh, 2016, è anche importante che gli studenti svolgano un po' di lavoro pratico. Possono anche scrivere blog per gli insegnanti così come per sé stessi per migliorare le loro capacità di conoscenza.

È chiaro che l'utilizzo di Internet da parte degli intervistati (studenti indiani) è stato per l'invio e ricezione di e-mail e la navigazione in internet rispettivamente con il 33% e il 26% (come rappresentato nella Figura 2.1).
In India, i siti di social networking stanno crescendo rapidamente per guadagnare popolarità, ma non hanno raggiunto le aspettative dello scenario globale. Sempre dalla figura 2.1, emerge che solo il 17% degli studenti ha segnalato i siti di social networking come motivo principale per l'utilizzo di Internet.
Tra i giovani indiani il 95,7% dei membri è connesso ai social media. Queste cifre aumentano di giorno in giorno. Mentre solo il 4,3% degli iscritti non è connesso ai social media.

Purpose of Internet Usage	
User	Percentage
Mail	33
Surfing	26.8
Chatting	18.7
Social Networking	17
Other	4.5
Total	100

Figura 2.1 Uso di internet fra gli studenti indiani, fonte: (Siddiqui, Singh, 2016).

Come venne evidenziato da E. H. Erikson nel 1959 lo sviluppo del senso del sé è un bisogno chiave nel momento dell'adolescenza e questo avviene soprattutto attraverso i pareri espressi da altri ragazzi della stessa età (Steinberg & Morris 2001). È importante per capire quanto peso abbia l'opinione altrui; gli individui prima di capire che immagine di sé mostrare devono capire cosa il gruppo a cui verranno esposti considerano come normativo/normale (Goffman 1959).

In "It's just a lot of Work: Adolescents' self-presentation norms and practices on Facebook and Instagram" (Yau, Reich, 2019) vengono evidenziate delle differenze nelle

interazioni effettuate nei social tra giovani adolescenti e adolescenti più grandi. I giovani prestano molta più attenzione al modo in cui i loro profili sono presentati, all'estetica, mentre gli adolescenti più adulti, al contrario, si focalizzano molto di più sulla rappresentazione delle loro amicizie. (Livingstone, 2008)

Secondo l'articolo "Social Media its Impact with Positive and Negative Aspects" di Siddiqui e Singh, i social media offrono agli studenti un modo per contattarsi efficacemente l'un l'altro in merito a iniziative di classe, compiti di gruppo o per aiuto nei compiti a casa.

Gli autori sostengono che molti degli studenti che non riescono ad intervenire durante la lezione potrebbero pensare di poter esprimere facilmente i propri pensieri sui social media, chiedendo pareri ai compagni.

Gli insegnanti possono pubblicare sui social media attività di classe, eventi scolastici, compiti a casa che saranno loro molto utili.

Dal punto di vista educativo ma anche in un secondo momento professionale, il social media marketing sta emergendo come opzione di carriera. Pertanto, può essere uno sbocco lavorativo al termine degli studi.

L'accesso ai social media offre agli insegnanti l'opportunità di insegnare come comportarsi durante l'uso degli stessi, ad esempio non lasciando commenti non educati e violenti, oltre l'uso di Internet per scopo utilitario e non solo come passatempo. (Siddiqui, Singh, 2016).

D'altro canto, un effetto negativo che viene in mente è il tipo di distrazione per gli studenti presenti in classe qualora venissero usati i social media per la lezione, poiché gli insegnanti non sono stati in grado di riconoscere chi sta prestando attenzione.

Uno dei più grandi fattori negativi dei social media nell'istruzione sono i problemi di privacy come la pubblicazione di informazioni personali sui siti online.

A causa dei social media gli studenti rischiano di perdere la capacità di impegnarsi per la comunicazione di persona, preferendo la comunicazione dietro uno schermo (Siddiqui, Singh, 2016).

2.2.1 Caso studio – Uso di sostanze, sentiment analysis: le conversazioni sui social media dei giovani che vivono senza fissa dimora

Nell'articolo *"Substance use and sentiment and topical tendencies: a study using social media conversations of youth experiencing homelessness"* prodotto da Deng, Barman-Adhikari, Lee, Dewri e Bender nel 2020, viene elaborata una *Sentiment Analysis* in merito a delle conversazioni su Facebook di giovani senza tetto e la somministrazione di un questionario. L'obiettivo di questo articolo è di indagare sulla fattibilità d'uso dei big data raccolti da Facebook e delle tecniche di *text analysis* come supplemento alle indagini condotte con il questionario, al fine di individuare e comprendere gli atteggiamenti e l'impegno dei giovani nell'uso di sostanze.

I dati raccolti risultano 92 risposte al questionario e 33,204 post e commenti su Facebook delle stesse persone che hanno somministrato il questionario.

Sui dati raccolti è stata implementata una *sentiment analysis* utilizzando il dizionario *Valence Aware Dictionary and sEntiment Reasoner* (VADER) in Python. È stato utilizzato questo dizionario in quanto include emoticons, slang words e abbreviazioni, che sono molto comuni nelle conversazioni tra giovani nei social media (Barman-Adhikari et al., 2020)

Dai risultati emerge che lo score medio per il sentiment di tutte le conversazioni di Facebook, corrisponde a 0,094 che indica un sentiment leggermente positivo.

Qui di sotto la tabella che illustra lo score medio del sentiment per sostanza tra user e non user

Substance	Non-users	Users	Difference	Groups	Non-users	Users	Difference
Alcohol	0.114 (0.447) Obs: 18,781	0.068 (0.436) Obs:14,423	0.046***	Marijuana	0.138 (0.440) Obs: 7,153	0.081 (0.431) Obs: 26,051	0.057***
Cocaine	0.097 (0.445) Obs:31,506	0.036 (0.401) Obs: 1,698	0.061***	Crack	0.094 (0.443) Obs: 33,191	0.339 (0.391) Obs: 13	−0.245*
Heroin	0.093 (0.443) Obs: 32,969	0.184 (0.407) Obs: 235	−0.091**	Meth	0.083 (0.440) Obs: 27,726	0.144 (0.451) Obs: 5,419	−0.061***
Ecstasy	0.098 (0.440) Obs: 28,314	0.065 (0.459) Obs: 4,866	0.033***				

Figura 2.2 Sentiment score per ogni sostanza tra user e non user, fonte: Deng, Barman-Adhikari, Lee, Dewri e Bender, 2020

Qui sotto invece le figure 2.3 e 2.4 confrontano il punteggio medio del sentiment nei post di Facebook tra tutti i gruppi di intervistati, in base al loro livello di consumo di una determinata sostanza.

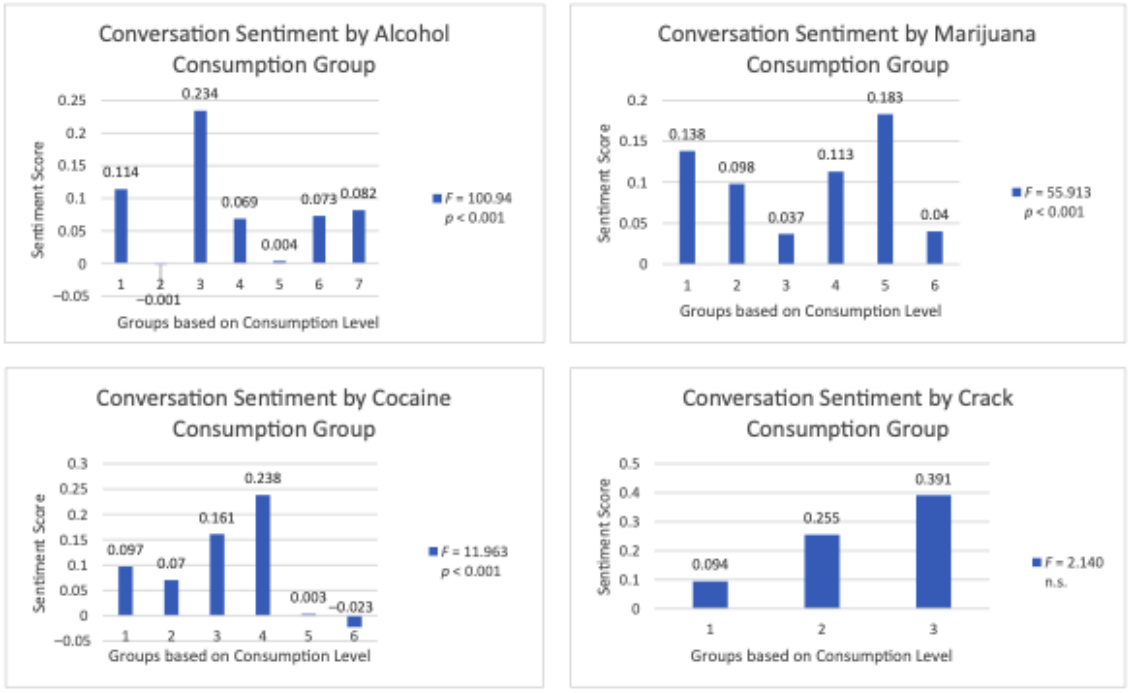

Figura 2.3 Sentiment level per sostanze in relazione al consumo, fonte: Deng, Barman-Adhikari, Lee, Dewri e Bender, 2020

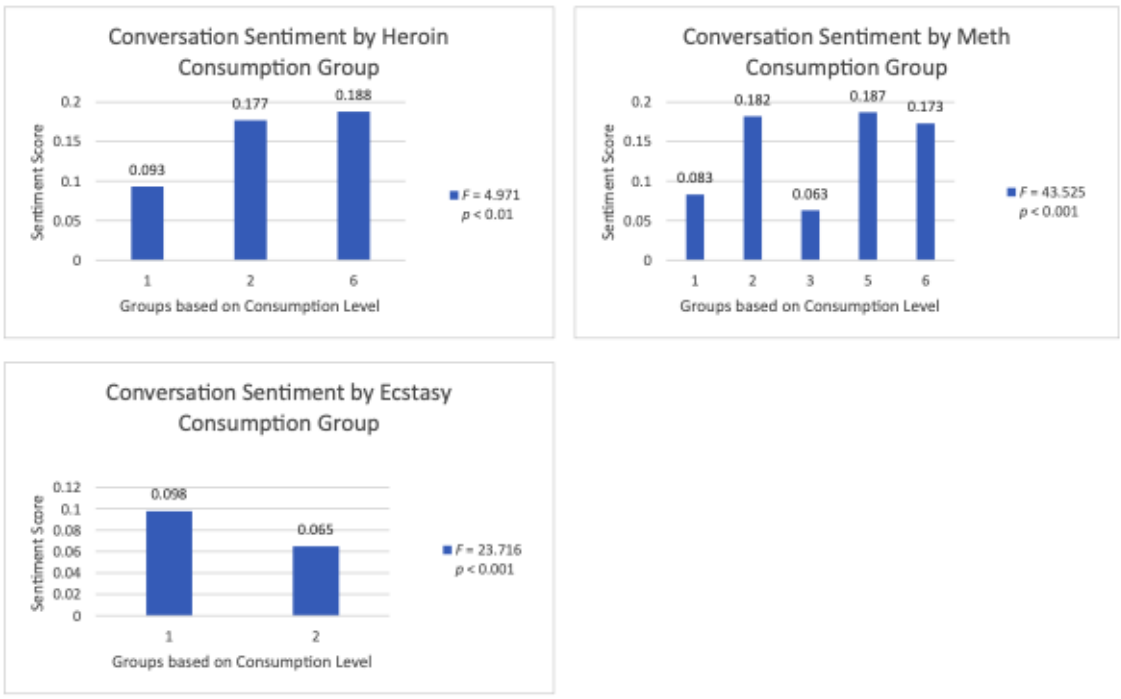

Figura 2.4 Sentiment level per sostanze in relazione al consumo, fonte: Deng, Barman-Adhikari, Lee, Dewri e Bender, 2020

Infine, è stato scoperto che più le conservazioni su Facebook sono positive da parte dei ragazzi homeless, meno è probabile che si trovi in un gruppo ad alto consumo di marijuana. In termini di argomenti citati nelle conversazioni, più i giovani menzionano

argomenti legati alla finanza, più è probabile che si trovino in un gruppo ad alto consumo di alcol e marijuana.
Per concludere, lo studio dimostra che identificare i giovani senza dimora ad alto rischio di uso di sostanze e indagare sulle loro conversazioni sui social media sono passi fondamentali per gli operatori sociali e sanitari per fornire assistenza e interventi.
(Deng, Barman-Adhikari, Lee, Dewri e Bender, 2020)

2.2.2 Impatto dei social media sull'istruzione

I social media offrono agli studenti un modo per contattarsi efficacemente l'un l'altro in merito a iniziative di classe, compiti di gruppo o per aiuto nei compiti a casa.
Molti degli studenti che non riescono ad intervenire durante la lezione potrebbero pensare di poter esprimere facilmente i propri pensieri sui social media, chiedendo pareri ai compagni.
Gli insegnanti possono pubblicare sui social media attività di classe, eventi scolastici, compiti a casa che saranno loro molto utili.
Dal punto di vista educativo ma anche in un secondo momento professionale, il social media marketing sta emergendo come opzione di carriera. Pertanto, può essere uno sbocco lavorativo al termine degli studi.
L'accesso ai social media offre agli insegnanti l'opportunità di insegnare come comportarsi durante l'uso degli stessi, ad esempio non lasciando commenti non educati e violenti, oltre l'uso di Internet per scopo utilitario e non solo come passatempo. (Siddiqui, Singh, 2016).
D'altro canto, un effetto negativo è il tipo di distrazione per gli studenti presenti in classe qualora venissero usati i social media per la lezione, poiché gli insegnanti non sono stati in grado di riconoscere chi sta prestando attenzione.
Uno dei più grandi fattori negativi dei social media nell'istruzione sono i problemi di privacy come la pubblicazione di informazioni personali sui siti online.
A causa dei social media gli studenti rischiano di perdere la capacità di impegnarsi per la comunicazione di persona, preferendo la comunicazione dietro uno schermo (Siddiqui, Singh, 2016).

2.2.3 Caso studio: *Sentiment Analysis* sulle prospettive degli studenti sulla registrazione delle lezioni

Questo articolo dal titolo *"Social Network and Sentiment Analysis: Investigation of Students' Perspectives on Lecture Recording"* scritto da Knomo, Ndukne e Daniel nel 2020, presenta le prospettive degli studenti sulle lezioni registrare come metodo di apprendimento, attraverso l'analisi delle conversazioni su Facebook.

La raccolta dei dati è avvenuta attraverso una domanda posta su una pagina Facebook dell'associazione studentesca di un'università, al quale hanno risposto 1,435 studenti con un emoji, altri 220 like e 65 commenti sono stati generati da 150 studenti unici.

La sentiment analysis è stata condotta attraverso Google natural language API per ottenere il sentiment dei 65 commenti.

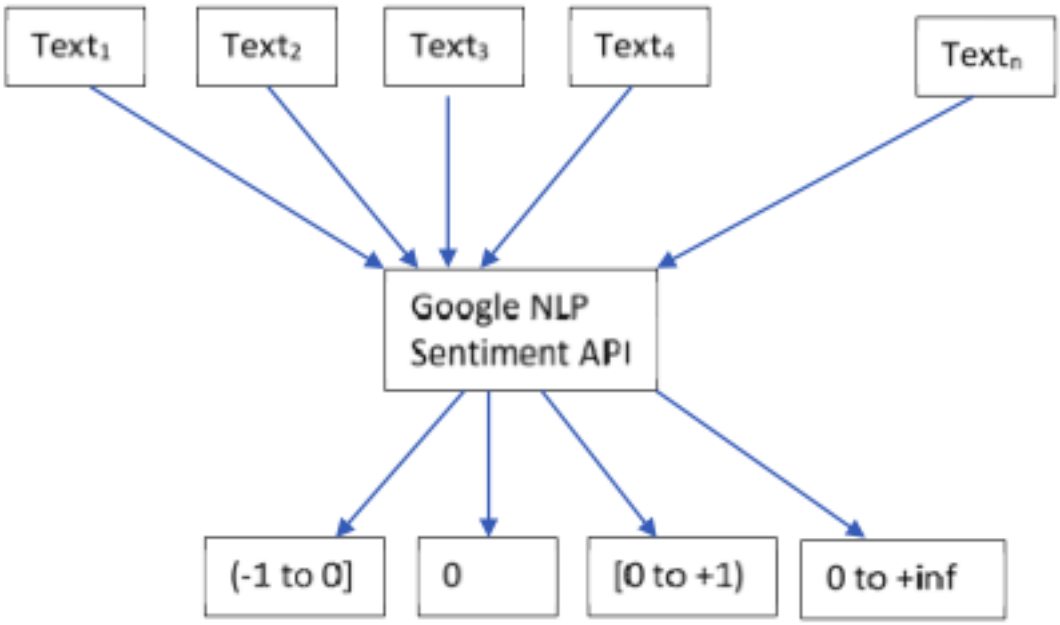

Figura 2.5 Schema generale di Google NLP per Sentiment, fonte: Knomo, Ndukne e Daniel, 2020

Sentiment	Percentage (%)
Positive	39.4
Neutral	33.3
Negative	27.3

Figura 2.6 Sentiment score per i commenti, fonte: Knomo, Ndukne e Daniel, 2020

Questo studio suggerisce che gli studenti utilizzano le lezioni registrate come materiale didattico complementare e non come sostituto delle lezioni in presenza. Gli studenti hanno sottolineato che le registrazioni sono state d'aiuto per le lezioni perse, altri hanno usato le lezioni registrate per sostituire le lezioni dal vivo a causa di circostante come la malattia ed impegni di lavoro.

2.3 Impatto dei social media sul business

I social media sono un'area di interesse nel marketing per le aziende, le organizzazioni e i marchi, che dando la possibilità di stabilire connessioni e creare comunità di follower in supporto ai prodotti.

Le aziende utilizzano i social media per migliorare le prestazioni di un'organizzazione in vari modi, ad esempio per raggiungere gli obiettivi aziendali, aumentando le vendite annuali. I social media offrono il vantaggio di essere una piattaforma di comunicazione che facilita la comunicazione bidirezionale tra un'azienda e i suoi azionisti.
Il business può essere promosso attraverso vari siti di social networking. Molte delle organizzazioni promuovono la propria attività facendo pubblicità sui social media per attirare il massimo di utenti o clienti.
I clienti possono connettersi e interagire con le aziende su un altro livello personale utilizzando i social media. Se un'organizzazione ha stabilito un marchio, i social media possono aiutare questa organizzazione a sviluppare il marchio esistente e dare voce all'azienda.
Con l'aiuto dei social media, l'organizzazione può elaborare la propria strategia per promuovere la propria organizzazione.

Qui di seguito vengono descritti come i social media possono essere utilizzati a supporto di diverse funzioni aziendali quali:

In primo luogo l'area Marketing, la quale utilizza quotidianamente i social media per creare contenuti, campagne pubblicitarie online, interagire con i propri follower.
In secondo luogo l'area delle Risorse umane: i social media sono uno strumento ottimo per identificare e coinvolgere direttamente i talenti. Le risorse umane aiutano l'azienda a mostrare i vantaggi dei dipendenti e la cultura dell'azienda al mondo esterno. La condivisione creativa consente ai team di arte, copia e design di inventare nuove idee utili all'azienda per raggiungere l'obiettivo.
Di seguito l'area di operazioni/strategia: molti dei siti come LinkedIn aiutano l'azienda collegandosi con gli esperti che possono condividere alcuni piani strategici.

Tra i diversi effetti positivi, sicuramente i social media aiutano a comprendere meglio il loro pubblico in base ai commenti lasciati dagli utenti sotto i post di un'azienda, sono canale di comunicazione per le attività promozionali. (Siddiqui, Singh, 2016).

Per un'azienda essere presente nei social media aiuta a rafforzare, per il cliente, l'esperienza del marchio che favorisce la costruzione della *brand identity*.
Un'azienda diventa più attraente per i clienti e per i dipendenti attuali e potenziali se ha un marchio ben costruito. Di conseguenza, i social media contribuiscono a costruire una buona reputazione per un'organizzazione aziendale.
La costruzione di un marchio aiuta a rafforzarlo nella mente dei consumatori, alimentano la *brand awareness*. I clienti sperimentano il privilegio del marchio durante l'utilizzo di un prodotto o di un servizio e anche quando interagiscono con un'azienda (Carraher, Parnell, McClure & Sullivan, 2006).
Le discussioni attraverso i commenti nei social media agiscono come un modo potente per comunicare il valore e gli attributi del marchio, in quanto facilitano forme aperte di comunicazione e la possibilità di confronto.

Un ulteriore aspetto positivo riguarda la promozione di una comunicazione aperta tra dipendenti e management.
I social media cconsentono ai dipendenti di condividere idee di progetto e di lavorare in team in modo efficace, il che aiuta a condividere conoscenze ed esperienze.

I social media promuovono anche contenuti migliori, come webcast e video, rispetto al semplice testo. Aiutano a comunicare in modo collaborativo tra clienti attuali e potenziali, per ricevere riscontri, definizione del prodotto, sviluppo del prodotto o qualsiasi forma di assistenza e supporto ai clienti.
I social media diventano una buona sede per le discussioni, le stesse aziende devono però assicurarsi che i dipendenti rispettino le regole e le regole d'uso dei social media.
Un altro modo per creare *brand awareness* per le aziende è quello di entrare a far parte di un forum esistente e aprire un nuovo forum per gli stakeholder (Kukulska-Hulme, 2010). Se questo viene fatto in modo corretto, si può ottenere una buona reputazione e costruire un'advocacy, il che significa che le persone tendono a parlare positivamente dell'azienda in modo pratico e sensibile (Carraher, 2011). La sfida principale per un social media è quella di essere una fonte di comunicazione affidabile, poiché non serve a limitare i danni.
Esistono diversi servizi a pagamento per il monitoraggio delle conversazioni sul web. Essi effettuano un'analisi qualitativa e quantitativa di come si svolgono le discussioni e di quanto si diffondono le informazioni su Internet.
I social media come Facebook, blog e YouTube sono diventati le principali fonti di assunzione. LinkedIn è un'altra fonte simile utilizzata dai reclutatori, e circa l'80% delle aziende lo usa per il processo di reclutamento. I social media sono diventati anche uno dei modi più semplici per migliorare la produttività e generare un interesse naturale per il lavoro svolto dai dipendenti. Grazie a questi mezzi di comunicazione, le aziende possono monitorare la creatività e l'entusiasmo dei dipendenti.

Per quanto concerne gli aspetti negativi, i social media non sono del tutto privi di rischi perché molti follower sono liberi di pubblicare la propria opinione su una particolare organizzazione, il commento negativo può influire sulla reputazione dell'azienda.
Successivamente, molte delle grandi organizzazioni sono state vittime degli hacker che sequestrano i profili delle pagine social. Inoltre, una strategia di brand online sbagliata può condannare un'azienda e comportare un enorme svantaggio sociale virale.
La maggior parte delle aziende ha difficoltà a misurare i risultati della pubblicità sui social media. (Siddiqui, Singh, 2016).

2.3.1 Caso studio – Impatto dei social media sul capitale delle aziende con un approccio di sentiment analysis

Nell'articolo "*The impact of social and conventional media on firm equity value: A sentiment analysis approach*" pubblicato da Yu, Duan, Cao nel 2012, viene messo in evidenza come l'analisi testuale può contribuire a comprendere l'impatto delle informazioni sui rendimenti azionari quando vi sono delle recensioni e discussioni nei social media. L'obiettivo dello studio è di indagare l'effetto social media e dei media convenzionali, la loro importanza relativa e la loro interazione sulle performance di mercato azionario delle imprese nel breve periodo.

In primo luogo, sono stati raccolti i dati finanziari appartenenti a 824 compagnie dal 1 luglio al 30 settembre 2011 al fine di elaborare i risultati del mercato azionario.

In contemporanea sono stati collezionati dati derivanti da Google Blog Search, forum, news e microblog come Twitter, sempre riguardanti le 824 compagnie, dal 1° luglio 2011 al 30 settembre 2011.

In secondo luogo, è stato utilizzato l'algoritmo Naïve Bayes (NB) per condurre la sentiment analysis, con parametri che misurano la polarità del sentimento, i quali si spostano da -1 (molto negativo) ad 1 (molto positivo). (Yu, Duan, Cao, 2012)

La tabella sottostante (Figura 2.7) mostra i risultati della stima su fixed effect utilizzando come variabili indipendenti il volume totale o il sentiment dei social media.

Variable	Coefficient (Std. Err.)	Coefficient (Std. Err.)	Coefficient (Std. Err.)	Coefficient (Std. Err.)
	Model (a1)	**Model (a2)**	**Model (a3)**	**Model (a4)**
Return equation: with abnormal return AR_{it} as dependent variable				
Constant	– .0002 (.0002)	– .0003 (.0002)	– .0002 (.0002)	– .0003 (.0002)
$NEWS\_NUM_{i,t-1}$	.0002 (.0005)	.001 (.001)		
$MEDIA\_NUM_{i,t-1}$	9.43e – 06 (.00002)	.00003 (.00002)		
$NEWS\_SENTI_{i,t-1}$			.00003 (.0006)	.001 (.001)
$MEDIA\_SENTI_{i,t-1}$			.00002 (.00002)	.00003 (.00002)
$NEWS\_NUM_{i,t-1}$* $MEDIA\_NUM_{i,t-1}$		– .00004 (.00002)*		
$NEWS\_SENTI_{i,t-1}$* $MEDIA\_SENTI_{i,t-1}$				

Figura 2.7 Effetti dei social media su il ritorno dell'investimento, fonte: Yu, Duan, Cao, 2012

Risk equation: with idiosyncratic risk IR_{it} as dependent variable

Constant	.03 (9.97e – 06)	.03 (.00001)	.03 (9.75e – 06)	.03 (9.88e – 06)
$NEWS\_NUM_{i,t-1}$	.00004 (.00003)**	.0001 (.00004)***		
$MEDIA\_NUM_{i,t-1}$	8.36e – 06 (9.4e – 07)***	1.00e – 05 (1.06e – 06)***		
$NEWS\_SENTI_{i,t-1}$			.00003 (.00003)	.00005 (.00004)
$MEDIA\_SENTI_{i,t-1}$			7.92e – 06 (1.35e – 06)***	8.50e – 06 (1.44e – 06)***
$NEWS\_NUM_{i,t-1}$* $MEDIA\_NUM_{i,t-1}$		– 4.46e – 06 (1.31e – 06)***		
$NEWS\_SENTI_{i,t-1}$* $MEDIA\_SENTI_{i,t-1}$				– 2.55e – 06 (2.21e – 06)
N = 49,807 Group = 824				

Figura 2.8 Effetto Sentiment su rischio di investimento fonte: Yu, Duan, Cao, 2012

Dalle due tabelle troviamo che il sentiment dei social media ha una forte relazione positiva con il rischio azionario, indicando che il sentiment complessivo dei canali social può aumentare la fluttuazione del mercato azionario.

Questi risultati evidenziano l'importanza dei social media e dei media convenzionali (in misura minore) sulla performance azionaria delle imprese.
Successivamente, vengono esaminati gli effetti dei social media sulla performance azionarie delle imprese a seconda del tipo di social media (ad esempio, blog, Twitter e forum). In particolare, utilizzando misure di sentiment per ogni singolo media, viene scoperto che il sentiment dei blog ha un impatto positivo, mentre quello dei forum ha un impatto negativo sul rendimento. Inoltre, sia il sentiment dei blog che quello di Twitter hanno un effetto positivo sul rischio.
In aggiunta, è emerso che l'effetto di interazione tra il sentiment di Twitter e quello delle news ha un effetto negativo significativo sui rendimenti, ma non un effetto significativo sul rischio.

Infine, viene esaminato il conteggio dei messaggi positivi e negativi dei media sulla performance azionaria per raccogliere informazioni sull'effetto a un livello più dettagliato.
Gli autori documentano che i messaggi positivi sui blog hanno un forte impatto positivo sul rendimento, mentre i messaggi negativi sui forum hanno un forte impatto negativo sul rendimento. Si ipotizza che i messaggi dei blog contengano più contenuti positivi, mentre i messaggi dei forum siano più orientati in senso negativo. Pertanto, una migliore ricerca sui social media può essere associata alla qualità della disponibilità e dell'elaborazione delle informazioni. (Yu, Duan, Cao, 2012)

2.4 Impatto dei social media sulla politica

Nel valutare il reale impatto dei social media sulla politica dobbiamo tener presente che la dimensione partecipativa è solo uno degli aspetti della socialità, infatti c'è da considerare anche il concetto di socievolezza simmeliano ossia il piacere estetico della conversazione disinteressata, del pettegolezzo.

Ci sono 3 rischi molto concreti dei social media come contesti discorsivi:

1. Polarizzazione delle opinioni: nei social media è presente una forte tendenza al *blaming*, ossia all'insulto delle persone con opinioni diverse dalle proprie. In tal caso il ruolo dei social media si discosta molto poco da quello delle classiche testate informative;

2. Infiltrabilità: dipendenza dei social media dai grandi media tradizionali e di mainstream per l'approvvigionamento di notizie. Es: Twitter che è quasi sempre usato per far circolare articoli, per commentare trasmissioni e di rado per testimoniare live certi avvenimenti (manifestazioni).

3. Fragilità: i rapporti con il potere sono piuttosto ambigui e complessi. Di solito i regimi dittatoriali tendono a far tacere l'intera rete bloccando i server o stipulando accordi specifici con le aziende titolari delle piattaforme, come una conferma a

contrario del potenziale democratico dei social network. Inoltre, i regimi dittatoriali tendono a sfruttare la viralità della rete per produrre disinformazione, facendo circolare notizie false. Infine, questi regimi tendono a usare le informazioni sugli utenti messe a disposizione dai social media per individuarli, seguirne le attività. Quindi internet consente una forte tracciabilità dei dati relativi alla privacy che nei paesi democratici è tutelata da un'apposita legislazione e nei paesi totalitari non lo è affatto.

A fronte dei 3 rischi appena menzionati è, comunque, possibile interpretare i social media con fattori positivi nelle democrazie, in quanto ognuno ha la possibilità di esprimere il proprio parere a patto di essere consapevoli dei loro limiti. Riepiloghiamo i fattori di innovazione che essi hanno comportato nel dibattito e nell'azione pubblica:

1. Lo sviluppo di blog, social network e Web 2.0 ha ridotto il ruolo di "colli di bottiglia" che svolgevano i media tradizionali per due ragioni:

- L'informazione sui social media raggiunge soprattutto i cittadini che sfuggono dall'informazione mainstream, i quali sono i più attivi e capaci di mobilitare.
- Le notizie e i commenti presenti sui social media sono fonti che i media tradizionali non possono rifiutare per concorrenza: una notizia che potrebbe rivelarsi fondata non si può tacere con il rischio che sia un'altra testata o un social media pubblicarla.

2. I social media in quanto mezzi orizzontali garantiscono miglior integrazione tra circolazione delle idee e riuscita organizzativa di vere e proprie azioni politiche (manifestazioni, boicottaggi elettorali).

3. in quanto propensi alla costruzione di identità e di appartenenza ideale, i social media sono strumenti assai forti in mano a movimenti fondati su qualche tipo di identità e di progetto esplicito.

Quindi i social media possono essere eccellenti strumenti di democrazia dove vi siano le condizioni per una buona politica, la quale si fonda si fonda sulla qualità delle argomentazioni e informazioni circolanti ma non in modo esclusivo: senza senso delle

istituzioni, coscienza civile e disponibilità partecipativa la democrazia tace o muore e nemmeno i social media la potrebbero salvare.

Anche i social media come strumenti di comunicazione alternativi partecipano al dibattito e all'azione politica come contesti discorsivi dentro spazi più ampi: quelli dell'intera società. Vi è, comunque, una loro specificità che dà un'idea di come esse abbiano contribuito a smuovere le acque della crisi politica.

2.4.1 Il caso studio – Text analysis dell'utilizzo dei social media durante le elezioni federali tedesche del 2013

Nell'articolo "Election Campaigning on Social Media: Politicians, Audiences, and the Mediation of Political Communication on Facebook and Twitter**"** pubblicato nel 2018 da Stier, Bleier, Lietz e Strohmaier, viene messo in evidenza come i social media vengono utilizzati dai politici durante le campagne delle elezioni. L'obiettivo dell'articolo indaga se i candidati alle elezioni affrontano gli argomenti più importanti per il pubblico di massa e in che misura la loro comunicazione è modellata dalle caratteristiche di Facebook e Twitter.

Oggetto di questo studio sono le elezioni federali tedesche del 2013.

I dati sono stati raccolti attraverso l'indagine "German Longitudinal Election Study" (GLES), condotta dall'8 luglio 2013 al 3 novembre 2013, intervistando 7.882 partecipanti due volte, in un'occasione pre-elettorale e in una post-elettorale, per un totale di 23604 osservazioni;

In aggiunta ai dati dell'indagine, si sono considerati anche i post pubblicati nei social media Facebook e Twitter dai candidati e i relativi commenti, per un totale di 49573 post di Facebook e 134462 tweets.

Il modello implementato è a metà strada tra la classificazione e il clustering: sono state utilizzate le risposte al sondaggio etichettate come dati di addestramento e divise in

categorie, poi sono stati raggruppati i messaggi dei social media che utilizzano parole o combinazioni di parole diverse dalle risposte al sondaggio in nuovi argomenti che vanno a formare ulteriori categorie. Per la text analysis è stato utilizzata la procedura di campionamento Gibbs ogni 100 interazioni.

Dai risultati della text analysis rappresentati nella figura 2.9, emerge che i politici utilizzano Facebook e Twitter per scopi differenti; infatti, nel primo social i messaggi riguardanti temi legati alla campagna elettorale sono il 42.3% mentre su Twitter sono il 26.1%. D'altro canto, i dibattiti politici degli elettori sono molto più presenti su Twitter rispetto a Facebook.

		Politicians		Audience	
	Survey	Facebook	Twitter	Facebook	Twitter
Known topics from the survey					
Labor Market	19.1	4.9	6.3	8.3	6.3
General Social Policy	12.9	1.1	1.5	1.4	1.4
Currency & Euro	12.5	1.6	2.6	1.9	2.2
Education	7.2	1.1	1.4	0.5	1.1
Economy	7.0	0.5	0.6	0.5	0.6
Infrastructure	6.8	3.0	5.7	1.5	5.1
Health Care & Pensions	5.7	0.9	1.0	0.5	0.7
Migration & Integration	4.0	0.2	0.2	0.1	0.3
Polity I	4.0	0.4	0.3	0.3	0.2
Family Policy	3.3	2.4	2.4	2.3	3.0
Law & Order	3.3	3.9	7.5	3.9	9.4
Foreign Policy (Defense)	2.9	2.4	3.4	2.3	3.2
Budget & Debt	2.7	0.2	0.2	0.1	0.1
Taxes	2.6	0.2	0.3	0.1	0.2
Foreign Policy (Europe)	2.2	0.1	0.1	0.2	0.1
General Fiscal Policy	1.7	0.1	0.1	0.0	0.1
Environment	1.5	0.1	0.1	0.0	0.1
Politics	0.6	3.8	0.8	1.3	0.7
New topics found on social media					
Campaigning (Local)		21.2	13.7	5.6	7.4
Campaigning (Events)		21.1	12.4	1.9	4.6
Political Debates		8.5	13.6	21.1	18.8
Polity II		5.7	7.2	22.0	13.2
Coalition Formation		5.6	6.7	12.0	8.1
Post Election		3.8	3.8	6.9	4.8
Parliamentary Procedures		3.3	3.2	0.5	1.4
Demonstrations		2.3	2.8	0.5	2.0
NSA Surveillance		0.4	1.0	0.4	3.1
Misconduct		0.3	0.5	0.3	1.4

Figura 2.9 Topic salience per Corpus %, fonte: Stier, Bleier, Lietz e Strohmaier, 2018

Sulla base del modello di text analysis sviluppato per questo studio, emerge che i politici ed il loro pubblico discutono di argomenti diversi rispetto al pubblico di massa. Inoltre, i politici usano Facebook e Twitter per scopi diversi, che in questo studio sono stati messi in relazione con i diversi gruppi target che i candidati incontrano. Nel complesso, i risultati suggeriscono che le strategie di campagna elettorale e la comunicazione politica in generale sono mediate dalle diverse possibilità sociotecniche delle piattaforme dei social media (Stier, Bleier, Lietz e Strohmaier, 2018).

2.5 Gli effetti dei social media nelle diverse tematiche sociali, politiche ed economiche e lo studio di esse attraverso la text analysis

In questo capitolo è stato illustrato come innanzitutto come i social media sono influenti nella sfera sociale, in particolare sui giovani e sull'istruzione.
In merito a queste tematiche, sono stati presentati due articoli in merito ai giovani senza tetto dipendenti dalle droghe che utilizzano i social, gli studenti che spiegano su Facebook la loro opinione in merito alle lezioni registrate.
Successivamente si è discusso di come i social media siano importante per un'azienda, sia per comunicare il proprio prodotto, i propri valori, sia per raccogliere i commenti degli utenti che vanno a determinare la reputazione aziendale. Legato a questo argomento vi è l'articolo in merito a come i social media possano influenzare i risultati delle azioni di capitale nel breve periodo.
Infine, si è parlato di come l'approccio dei politici sia cambiato con l'avvento dei social network, quanto quest'ultimi possono essere decisivi durante una campagna elettorale, è stato visto attraverso un articolo come i politici cambino linguaggio e temi in relazione al social network che utilizzano.
Nel capitolo successivo vengono presentati dei metodi di text analysis che possono essere utili per analizzare alcuni dati raccolti dai social media in merito ad alcune tematiche come quelle citate sopra.

CAPITOLO 3: ANALISI TESTUALE E SENTIMENT ANALYSIS

L'analisi testuale, comunemente chiamata *text analysis*, è un campo di ricerca che si occupa di estrarre informazioni da fonti di dati non strutturati. La text analysis si può applicare a testi, documenti, recensioni su siti internet, o post dei social media, al fine di trovare informazioni utili per produrre nuove conoscenze.
Ad esempio, la text analysis può essere utilizzata per analizzare i post dei social media, ad esempio per un'organizzazione di assistenza clienti, per avere una panoramica generale della reputazione del suo brand, attraverso l'analisi dei commenti sui social network. Può essere utilizzata nelle risorse umane per vari scopi, come la comprensione della percezione dell'organizzazione da parte dei candidati o la corrispondenza tra le descrizioni delle mansioni e i curriculum. Il text mining ha implicazioni di marketing per misurare la salienza delle campagne.
La text analysis è un ramo del Data Mining che analizza database di piccole o grandi dimensioni, attraverso sistemi automatici, con lo scopo di trovare dei "pattern", ovvero una rappresentazione sintetica e ricca di semantica che deve essere valida, comprensibile, precedentemente sconosciuta e potenzialmente utile. A supporto del Data Mining vi è il Machine learning, in quanto è necessario per attuare tecniche di analisi dei dati.

Le attività tipiche del Machine learning possono essere raggruppate principalmente in due categorie: tecniche di apprendimento supervisionato che utilizzano alcune variabili per predire il valore non conosciuto o futuro di altre variabili obiettivo, e le tecniche di apprendimento non supervisionato che trovano dei pattern che siano interpretabili e che descrivano i dati.

Apprendimento automatico supervisionato

L'approccio di apprendimento automatico supervisionato si riferisce a tutte le classi di tecniche in cui un algoritmo apprende schemi da un insieme di dati di addestramento (training set). L'idea intuitiva è che questi algoritmi possono imparare a codificare i testi se diamo loro esempi sufficienti di come dovrebbero essere codificati. Un esempio semplice è l'analisi del sentiment, che utilizza una serie di testi codificati manualmente come positivo, neutro, o negativo, in base al quale l'algoritmo può apprendere quali caratteristiche (parole o combinazioni di parole) hanno maggiori probabilità di verificarsi nei testi positivi o negativi. Dato un *unseen text* (un testo nuovo non utilizzato nella fase

di training, da cui l'algoritmo non è stato addestrato), il sentimento del testo può quindi essere stimato in base alla presenza di queste caratteristiche. La parte deduttiva è che i ricercatori forniscono i dati di addestramento, che contengono buoni esempi che rappresentano le categorie che i ricercatori stanno tentando di prevedere o misurare. Tuttavia, i ricercatori non forniscono regole esplicite su come cercare questi codici. La parte induttiva è che l'algoritmo di apprendimento automatico supervisionato apprende queste regole dai dati di addestramento.

Apprendimento automatico non supervisionato

Negli approcci di apprendimento automatico non supervisionato, non vengono specificate regole di codifica e non vengono forniti dati di addestramento. Invece, un algoritmo crea un modello identificando determinati schemi nel testo. L'unica influenza del ricercatore è la specifica di alcuni parametri, come il numero di categorie in cui sono classificati i documenti. Esempi popolari sono il topic modeling per classificare automaticamente i documenti sulla base di una struttura topica sottostante (Blei et al., 2003; Roberts et al., 2014) e il modello fattoriale parametrico "Wordfish" (Proksch & Slapin, 2009) per ridimensionare i documenti su una singola dimensione sottostante, come l'ideologia sinistra-destra.
Grimmer e Stewart (2013) sostengono che l'apprendimento automatico supervisionato e non supervisionato non siano metodi concorrenti, ma soddisfano scopi diversi e possono benissimo essere utilizzati per completarsi a vicenda. I metodi supervisionati sono l'approccio più adatto se i documenti devono essere collocati in categorie predeterminate, perché è improbabile che un metodo non supervisionato produca una categorizzazione che rifletta queste categorie e il modo in cui il ricercatore le interpreta. Il vantaggio dei metodi non supervisionati è che possono elaborare categorie che i ricercatori non avevano considerato. Anche se ciò può presentare problemi per l'interpretazione post-hoc quando i risultati non sono chiari.

Al fine di elaborare un'analisi del testo, è necessario scegliere lo strumento per poterla eseguire.
In questo elaborato viene usato il software R (www.r-project.org) ed alcune librerie dedicate all'analisi testuale.

Che cos'è R?

R è un ambiente di programmazione gratuito, open source e multipiattaforma. A differenza della maggior parte dei linguaggi di programmazione, R è stato specificamente progettato per l'analisi statistica, il che lo rende particolarmente adatto per le applicazioni di statistiche e data science.
Gli strumenti disponibili per eseguire l'analisi del testo in R si identificano per il loro obiettivo, partendo dall'acquisizione dei dati, il pre-processing fino allo sviluppo di approcci di apprendimento supervisionato o non supervisionato come precedentemente prodotto.

3.1 Tecniche di acquisizione dei dati

3.1.1 Preparazione dei dati

La preparazione dei dati è il punto di partenza per qualsiasi analisi dei dati.
Inoltre, la preparazione di testi per l'analisi richiede scelte che possono influenzare l'accuratezza, la validità e i risultati di uno studio di analisi del testo tanto quanto le tecniche utilizzate per l'analisi (Crone, Lessmann e Stahlbock, 2006; Günther & Quandt, 2016; Leopold & Kindermann, 2002). Qui distinguiamo cinque passaggi generali: importazione di testo, operazioni sulle stringhe, pre-elaborazione, normalizzazione e rimozione delle *stopword*.

Importazione di testo

I dati testuali possono essere memorizzati in un'ampia varietà di formati di file. R supporta nativamente la lettura di normali file di testo flat come CSV e TXT. Lavorare con questi diversi pacchetti e le loro diverse interfacce e output può essere impegnativo, soprattutto se diversi formati di file vengono utilizzati insieme nello stesso progetto. Una soluzione conveniente per questo problema è il r*eadtextpackage* (Benoit & Obeng, 2017), che raggruppa vari pacchetti di importazione insieme per offrire un'unica funzione catch-all per l'importazione di molti tipi di dati in un formato uniforme.

Operazioni sulle stringhe

Uno dei requisiti fondamentali di un framework per l'analisi computazionale del testo è la capacità di manipolare testi digitali. Il testo digitale è rappresentato come una sequenza di caratteri, chiamata stringa. In R, le stringhe sono rappresentate come oggetti chiamati tipi "carattere", che sono vettori di stringhe. Il gruppo di operazioni sulle stringhe si riferisce alle operazioni di basso livello per lavorare con dati testuali. Le operazioni di stringa più comuni sono: unire, dividere, e l'estrazione di parti di stringhe (indicate collettivamente come analisi) e l'uso di espressioni regolari per trovare o sostituire modelli.

Preelaborazione

Per la maggior parte dei metodi di analisi computazionale del testo, i testi completi devono essere trasformati in token, ovvero il testo viene scomposto in unità più piccole, quindi elementi di testo più piccoli e più specifici, come parole o combinazioni di parole. Inoltre, le prestazioni computazionali e l'accuratezza di molte tecniche di analisi del testo possono essere migliorate normalizzando le caratteristiche o rimuovendo le "parole non significative": parole designate in anticipo come prive di interesse e che vengono quindi scartate prima dell'analisi. Nel loro insieme, queste fasi preparatorie sono comunemente denominate "pre-elaborazione". Nel seguito verranno introdotte alcune delle più comuni tecniche di pre-elaborazione e verrà mostrato come eseguire ciascuna tecnica con la libreria “sentiment analysis”.

Tokenizzazione.

La tokenizzazione è il processo di suddivisione di un testo in token. Questo è fondamentale per l'analisi computazionale del testo, perché i testi completi sono troppo specifici per eseguirne uno qualsiasi. Molto spesso i token sono parole, perché questi sono le componenti base semanticamente significativi più comuni dei testi.

Normalizzazione: lettere minuscole e stemming.

Il processo di normalizzazione si riferisce in generale alla trasformazione delle parole in una forma più uniforme. Questo può essere importante se per una certa analisi un computer deve riconoscere quando due parole hanno (approssimativamente) lo stesso significato, anche se sono scritte in modo leggermente diverso. Un altro vantaggio è che riduce la dimensione del vocabolario (cioè l'intera gamma di caratteristiche utilizzate

nell'analisi). Una tecnica di normalizzazione semplice ma importante consiste nel rendere tutto il testo minuscolo. Se non eseguiamo questa trasformazione, un computer non riconoscerà che due parole sono identiche se una di esse è stata scritta in maiuscolo perché si trovava all'inizio di una frase.

Rimozione delle stopword.

Parole comuni come articoli e congiunzioni, ad esempio "the" in lingua inglese, sono molto frequenti in un testo, ma non sono informative perchè il loro scopo è puramente grammaticale. Filtrare queste parole ha il vantaggio di ridurre la dimensione dei dati, ridurre il carico computazionale e, in alcuni casi, anche migliorare la precisione. Per rimuovere queste parole in anticipo, vengono abbinate a elenchi predefiniti di "parole non significative". Diversi pacchetti di analisi del testo forniscono elenchi di stopword per varie lingue, che possono essere utilizzati per filtrare manualmente le stopword.

3.1.2 Scoring

La tecnica di scoring prende in considerazione un corpus di testi come un insieme di punti rappresentati su una retta, a ciascuno dei quali viene assegnato un punteggio. Con questa tecnica si crea un ordinamento dei testi per arrivare all'interpretazione del problema, come rappresentato in modo semplice in Figura 3.1. Si osserva una retta con al centro il valore 0, tutti i punti (rappresentati come rombi nella Figura) che si trovano alla sua sinistra hanno un valore negativo, d'altraparte tutti quelli che si trovano a destra hanno valore positivo, per quanto concerne i punti che più si avvicinano allo zero vengono considerati neutri. Rappresentazione di Scoring in Figura 3.1

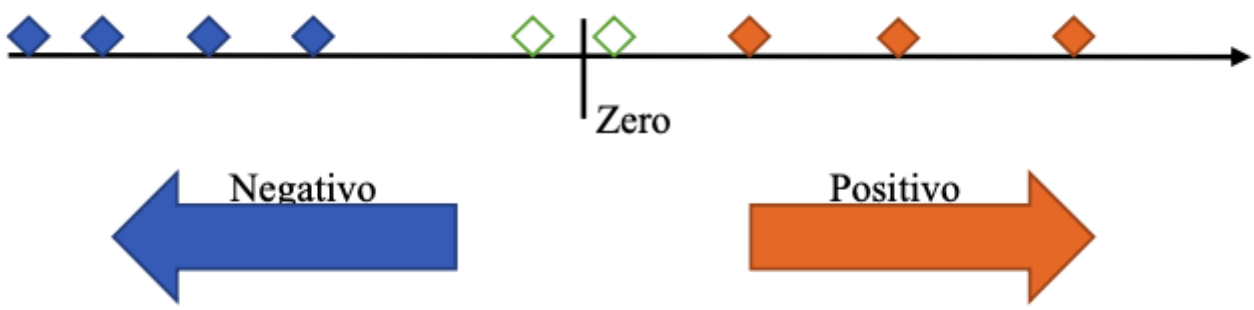

Figura 3.1 Esempio di scoring con categoria positiva, negativa e neutra

Uno degli approcci di scoring è l'algoritmo Wordfish (Slapin e Proksch, 2008), finalizzato all'analisi di testi non strutturati come ad esempio articoli di giornale, libri e post sui Social Network, che non presentano dunque strutture predefinite nel ritrovare un certo dato sempre nella stessa posizione nel documento. Questo algoritmo ordina i testi in base alla frequenza con cui i termini compaiono, posizionandoli lungo un asse ideologico identificato a posteriori dal ricercatore (ad esempio, giudizio positivo-negativo, asse sinistra-destra ecc.). Durante la fase di pre-processing, i testi vengono trasformati in vettori contenti il numero di volte che le parole compaiono, indipendentemente dalla posizione che hanno nel testo, perciò la posizione che una parola assume nel testo non influenza la posizione delle altre parole. Ogni vettore rappresenta dunque un documento del corpus, l'insieme dei vettori che corrispondono ai documenti del corpus, formano una matrice, detta "term-matrix" che diventa l'oggetto di analisi della ricerca. L'interpretazione dei testi viene fatta a posteriori in base al contenuto, individuando così la dimensione latente dei testi stessi.

3.1.3 Topic model e LDA

Tra le tecniche di analisi non supervisionate, la più frequentemente utilizzata è la Topic Model.

Le tecniche di Topic Modeling (Blei et al., 2003) sono una serie di metodi di analisi testuale che ricercano le strutture tematiche nascoste all'interno di un corpus di documenti e raggruppano i documenti in base al loro argomento principale. Una delle principali metodologie di topic modeling è Latent Dirichlet Allocation (LDA), dove si suppone esista un modello statistico basato sulla distribuzione di Dirichlet attraverso il quale ogni testo risulta essere composto da una mistura di argomenti latenti (topic) a cui viene associata una sequenza di parole. Ognuno di questi topic è una distribuzione multinomiale sulle parole, quest'ultime raggruppate in un vocabolario definito in precedenza sulla base dei testi analizzati: le parole con probabilità più alta forniscono un'idea dei temi trattati nel corpus di documenti. In altre parole, si assume che vi sia una distribuzione di topic nel corpus di documenti e ad ogni topic è associata una sequenza di parole. L'analisi LDA perciò, presuppone un processo di generazione del testo in due stadi: prima si sceglie il topic e poi si sceglie un gruppo di parole per discutere quel certo topic.

Nella figura 3.2 viene illustrato come si costruisce l'analisi LDA.

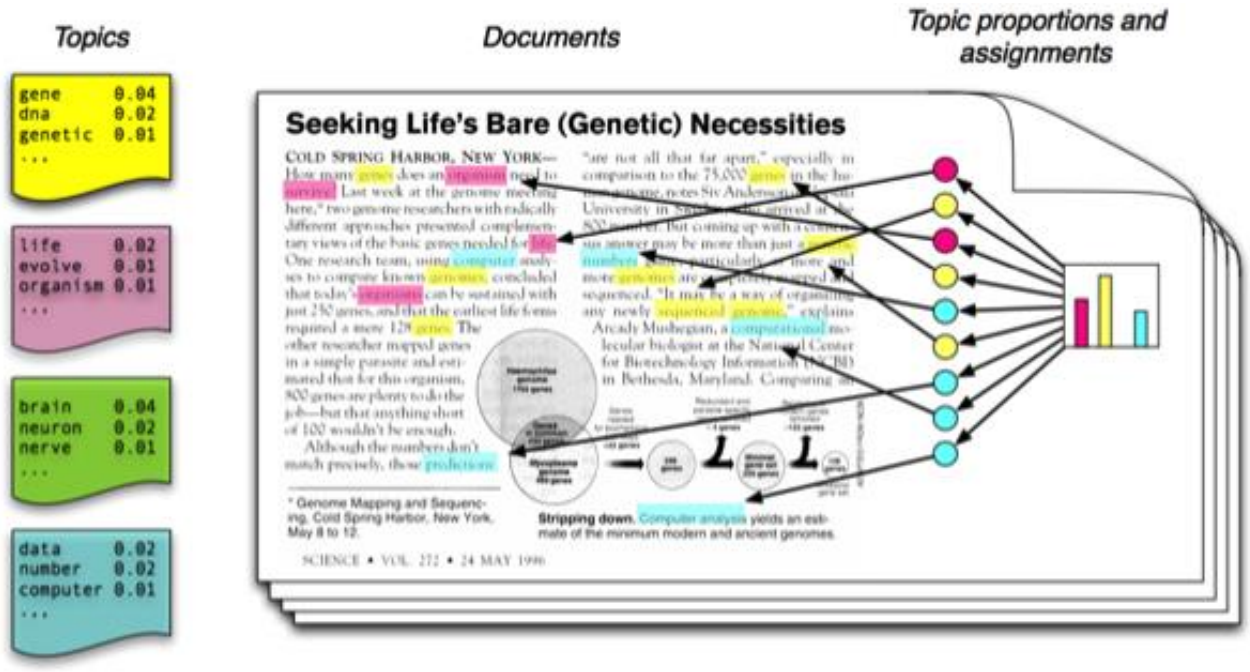

Figura 3.2 Esempio di topic model, fonte: Biel, 2012

3.2 Sentiment Analysis

Tra i vari campi di applicazione dell'analisi testuale, troviamo la Sentiment Analysis, una delle tecniche di analisi supervisionata più frequentemente utilizzata.

Inizialmente, l'analisi del sentiment può sembrare facile: si tratta di analizzare l'intento emotivo di un autore in classi distinte come felice, frustrato o sorpreso. Essa prende spunto da discipline come la linguistica, la psicologia e, naturalmente, l'elaborazione del linguaggio naturale.

L'analisi del sentiment è il processo di estrazione dell'intento emotivo di un autore da un testo.

Le sfide della sentiment analysis derivano non solo dalla sua fondazione interdisciplinare, ma anche dalle differenze culturali e demografiche tra gli autori.

È difficile quantificare la differenza tra felice, euforico o lo spettro tra annoiato, disinteressato e interessato. In effetti, senza la cattura del tono emotivo esplicito dell'autore al momento della scrittura, tutta la sentiment analysis può essere compromessa da pregiudizi dell'analista o della modellazione.

A complicare ulteriormente le difficoltà dell'analisi del sentiment può esserci un sentiment specifico per ogni caratteristica. Ciò si verifica quando l'argomento di cui si

scrive può avere più di un sentiment per caratteristica all'interno dell'argomento complessivo. Ad esempio, una recensione di un ristorante su Yelp può affermare che i prezzi sono ottimi ma il cibo è nella media. Nel complesso, quindi, la recensione può essere decente, ma la recensione stessa contiene due stati emotivi distinti (ottimo e medio) applicati a una specifica caratteristica del ristorante. (Kwartler, 2017)

Di norma, l'analisi dei sentimenti tenta di determinare la disposizione di un oratore, saggista o altri soggetti in termini di tema attraverso risposte in brevi commenti, testi o pensieri scritti.
La disposizione potrebbe essere un giudizio o una valutazione, piena di emozione (in altre parole, la condizione appassionata del creatore o dell'oratore) o un'aspettativa di risposte entusiaste (in altre parole, l'impatto voluto dal creatore o dall'acquirente).

La sentiment analysis è un'attività che sfrutta gli approcci del *Natural Language Processing* (PNL) e dell'estrazione delle informazioni (IE) per analizzare un ampio numero di archivi al fine di raccogliere i sentimenti dei commenti posti da diversi autori. Questo processo incorpora varie strategie, tra cui l'etimologia computazionale e il recupero delle informazioni (IR).
L'idea di base dell'indagine sui sentimenti è rilevare la polarità dei documenti di testo o delle brevi frasi e classificarli in base a questa premessa. La polarità del sentimento è classificata come "positiva", "negativa" o "imparziale" (neutrale). È importante evidenziare il fatto che il sentiment mining può essere eseguito su tre livelli come segue:
- Classificazione del sentiment a livello di documento: a questo livello, un documento può essere classificato interamente come "positivo", "negativo" o "neutro".
- Classificazione del sentimento a livello di frase: a questo livello, ogni frase è classificata come "positiva", "negativa" o imparziale.
- Classificazione del sentimento a livello di aspetto e caratteristica: a questo livello, le frasi/i documenti possono essere classificati come "positivi", "negativi" o "apartitici" alla luce di alcuni aspetti delle frasi/archivi e comunemente noti come "raggruppamento di valutazione a livello di prospettiva".

3.2.1 Sentiment Analysis in Twitter

Come presentato nel capitolo 1, il social media Twitter, è una piattaforma gratuita di Social Networking e micro-blogging che fornisce ai propri utenti la possibilità di creare un profilo personale aggiornabile tramite messaggi di testo e link della lunghezza massima di 140 caratteri. La missione di Twitter è dare a tutti la possibilità di creare e condividere idee e informazione istantaneamente, abbattendo qualsiasi barriera.

Twitter è caratterizzato da alcune funzionalità specifiche che elenchiamo di seguito:

- Twitta: Un tweet è un singolo messaggio pubblicato su Twitter. Il contenuto di un tweet, che può essere, al massimo, di 140 caratteri, può variare da informazioni personali o opinioni personali su prodotti o eventi ad altri come link, notizie, foto o video.

- Utente/Nome utente: Un utente deve essere registrato con la piattaforma per pubblicare tweet. L'utente seleziona uno pseudonimo (nome utente) durante la registrazione, che verrà successivamente utilizzato per inviare messaggi.
- Citare: le menzioni in un tweet indicano che il post menziona un altro utente. Per fare riferimento a un nome utente, gli utenti utilizzano il simbolo @ seguito dal nome utente specifico a cui si riferiscono (@nomeutente). Le menzioni sono posizionate ovunque nel corpo del tweet.

- Risposte: le risposte in un tweet vengono utilizzate per indicare che il post è una risposta a un altro tweet e vengono solitamente utilizzate per creare conversazioni. Analogamente alle menzioni, vengono create utilizzando il simbolo @ seguito dal nome utente a cui si riferiscono. Le risposte vengono posizionate accanto al nome utente che crea la risposta.

- Seguace: I follower si riferiscono agli utenti che seguono i tweet e l'attività di un utente. Seguire altri utenti è il modo principale per connettersi ad altri utenti su Twitter. Gli utenti su Twitter ricevono aggiornamenti da coloro che seguono e inviano i loro aggiornamenti a coloro che li seguono.

- Retwitta: I retweet si riferiscono ai tweet che vengono ridistribuiti. Quando un utente trova interessante un tweet, può ripubblicarlo utilizzando la funzionalità di retweet. Il retweet è considerato un potente strumento di diffusione delle informazioni. Il tweet che viene condiviso rimane invariato e viene solitamente contrassegnato con la sigla RT seguita dal nome utente dell'autore (RT@username). Il retweet può contenere anche un breve commento.

- Hashtag: Gli hashtag vengono utilizzati per indicare la rilevanza di un tweet rispetto a un determinato argomento. Gli hashtag che vengono creati utilizzando il carattere # seguito dal nome dell'argomento (#topic) sono nati dalla necessità di etichettare le informazioni sui messaggi che sono stati postati. I tag vengono generati spontaneamente dagli utenti e possono essere utilizzati per ottenere tutti i tweet con lo stesso hashtag. Gli hashtag che compaiono in un numero elevato di tweet sono caratterizzati come argomenti di tendenza.

- Riservatezza: Twitter dà la possibilità a un utente di decidere se i suoi tweet saranno visibili a tutti o solo ai suoi follower approvati su Twitter.

3.2.2 Il pacchetto "Sentiment Analysis" in R

La libreria utilizzata nel presente elaborato si chiama "*sentiment analysis*".
Questo pacchetto è stato implementato da Pröllochs N, Feuerriegel S, Neumann D (2018). Il metodo con cui viene svolta la sentiment analysis riguarda l'approccio della regolarizzazione della funzione LASSO, al fine di estrarre le parole che sono statisticamente decisive in base ad un risultato di una variabile.

L'approccio LASSO è un metodo di regolarizzazione, che di solito viene utilizzato nei modelli di regressione; infatti, in questo caso gli autori hanno ipotizzato che ogni testo sia un modello di regressione, dove la variabile dipendente può assumere valore negativo o positivo, mentre ogni parola del testo pre processata viene considerata come una covariata.

Con il metodo LASSO, il coefficiente di alcune covariate viene stimato pari a zero, concentrandosi sulle parole effetivamente rilevanti per l'output finale.
Come si può vedere in questa funzione dei minimi quadrati, dove viene minimizzata la somma tra la differenza del valore osservato al netto delle stime, viene aggiunta la sommatoria delle ßt, che ha un vincolo implicito che deve essere minore o uguale al parametro "λ" (lambda).
Se lambda si avvicina allo zero, si riduce il numero di covariate e quindi il numero di parole non rilevanti nel testo.

$$\boldsymbol{\beta}_{\text{LASSO}} = \underset{\beta_0,\ldots,\beta_n}{\arg\min} \sum_{i=1}^{|D|} \left[y_i - \beta_0 - \sum_{t=1}^{n} \beta_t \hat{x}_{d,t} \right]^2 \text{ s. t. } \sum_{t=1}^{n} |\beta_t| \leq \lambda$$

Successivamente, dopo aver implementato la regolarizzazione LASSO, rimangono le parole utili per la sentiment analysis, che vengono classificate positive, neutre, negative, attraverso l'utilizzo di tre dizionari: Harvard-IV, Henry's Financial dictionary e McDonald Financial dictionary.

La regolarizzazione LASSO produce tipicamente stime in cui alcuni dei coefficienti sono impostati esattamente a zero. Quindi esegue una selezione implicita delle caratteristiche.
In pratica il parametro "λ" viene selezionato utilizzando la *cross-validation* per trovare un valore che minimizzi l'errore sul set di dati. Successivamente, il modello viene riadattato con quella specifica "λ" utilizzando tutte le osservazioni per determinare i coefficienti.
Di conseguenza, la procedura identifica le parole statisticamente rilevanti, mentre i coefficienti corrispondenti ne misurano la polarità.
In questo modo, non si corre il rischio di etichettare parole per motivi soggettivi o sulla base di conoscenze errate, poiché tutti i risultati misurano l'influenza delle parole sulla variabile dipendente con una validazione statistica (Pröllochs N, Feuerriegel S, Neumann D, 2018)

Capitolo 4: Il caso studio del Pride, la sua percezione in Italia ed all'estero

Nel presente elaborato, si è voluto analizzare il sentiment del Pride in Italia e all'estero, in quanto le opinioni in merito alla comunità LGBTQIA+ sono molte e spesso divergenti; quindi, con questa tesi è stata data l'opportunità di condurre un'indagine per capire il pensiero di alcuni utenti in merito al Pride Month.
Il 28 giugno 1970, in occasione dell'anniversario della rivolta di Stonewall, si sono tenute le prime marce del Pride a New York, Los Angeles e Chicago. Migliaia di persone LGBT+ si riunirono per commemorare Stonewall e manifestare per la parità di diritti. Gli eventi di Stonewall e i movimenti di liberazione che ne sono seguiti sono stati il risultato diretto di decenni precedenti di attivismo e organizzazione LGBT+. In particolare, le tradizioni del Pride sono state adattate dai "Picchetti del Giorno del Ricordo" che si tenevano annualmente (1965-1969) il 4 luglio presso l'Independence Hall di Philadelphia, in Pennsylvania (Metcalf, 2020).

I picchetti annuali del Giorno del Ricordo sono stati organizzati dalla Eastern Regional Conference of Homophile Organizations (E.R.C.H.O). L'E.R.C.H.O. (inizialmente chiamata E.C.H.O.) si è formata nel 1962 come organizzazione di gruppi omofili della costa orientale che comprendevano il Capitolo di New York delle Figlie di Bilitis, la Janus Society di Filadelfia e la Mattachine Society di Washington e New York, e che sarebbero cresciuti fino ad includerne altri.
In figura 4.1 viene rappresentata tre uomini con dei cartelli di slogan durante il Reminder del Picket Day del 1968.

Figura 4.1 Immagine Reminder Picket Day 1968, fonte: web

Dopo la rivolta di Stonewall (giugno 1969), gli organizzatori del Annual Reminder Picket Day (Conferenza regionale orientale delle organizzazioni omofile) suggerirono di spostare l'attenzione dalla pianificazione del picchetto del Giorno del Ricordo all'organizzazione di una manifestazione annuale in commemorazione di Stonewall.

Alla Conferenza dell'E.R.C.H.O del novembre 1969, le 13 organizzazioni votanti presenti hanno adottato la seguente risoluzione:
"Proponiamo che ogni anno, l'ultimo sabato di giugno a New York, si tenga una manifestazione per commemorare le dimostrazioni spontanee del 1969 in Christopher Street e che questa manifestazione venga chiamata CHRISTOPHER STREET LIBERATION DAY". Lo Stonewall Inn si trova in Christopher Street ed è stato il punto di partenza della Rivolta (Metcalf, 2020).

Fin dall'inizio, gli organizzatori hanno pensato a una celebrazione nazionale: "Proponiamo anche di contattare le organizzazioni omofile di tutto il Paese e di suggerire loro di organizzare manifestazioni parallele in quel giorno. Proponiamo una dimostrazione di sostegno a livello nazionale".

Per avviare la pianificazione, hanno formato il Christopher Street Liberation Day Umbrella Committee. Il comitato ha definito l'obiettivo di organizzare una marcia di massa al culmine della settimana dell'orgoglio gay (22-28 giugno). (Metcalf, 2020)

Il primo Christopher Street Liberation Day fu un successo clamoroso, con migliaia di partecipanti che superarono le aspettative degli organizzatori. New York, Los Angeles e Chicago iniziarono subito a pianificare il 1971, e presto altre città, stati e paesi avrebbero iniziato a stabilire le proprie tradizioni annuali del Pride. Nella figura 4.2 viene fotografata la marcia per il Christopher Street Day del 1970.

Figura 4.2 Immagine marcia Christofer Day 1970, fonte: web

Dal giugno 1970, le persone LGBTQ+ hanno continuato a riunirsi a giugno per marciare con il Pride.

Al giorno d'oggi, il PRIDE viene celebrato in diversi paesi del mondo, nel mese di giugno.

4.1 Analisi dei dati estratti da Twitter

L'obiettivo di questo elaborato è quello di raccogliere, analizzare ed interpretare i dati estratti sottoforma di tweet dal social network Twitter, in merito alla tema del Pride in Italia e all'estero, ovvero la percezione degli utenti in relazione al Pride, tutto questo attraverso la Sentiment Analysis con R.

Innanzitutto, è necessario porre alcune assunzioni preliminari: sono stati divisi i testi in lingua italiana ed in lingua inglese, si dà per assodato che non vi è geolocalizzazione per i tweet scaricati perchè non sempre disponibile.

4.1.1 Download dei dati

Per scaricare i dati è stato utilizzato il pacchetto "Rtweet" che permette di scaricare i dati ed altre informazioni in base ad una specifica parola: in questo caso tutti i tweet sono stati scaricati con l'hashtag *pride*.
È bene sottolineare che l'accesso ai dati di Twitter è regolato dalle API, nello specifico sono parti di codice che danno la possibilità di scaricare, in modo limitato, i dati di Twitter.
Sono stati scaricati i dati nel periodo temporale dal 18 giugno al 2 luglio 2022 ed i tweet sono stati scaricati sia in italiano che in inglese e salvati in due diversi dataset. Il periodo temporale è rilevante in quanto si tratta del cosiddetto "Pride Month", il mese dedicato alla manifestazioni per la sensibilizzazione sul tema oggetto dell'indagine.
I dati scaricati sono contenuti in un dataset di 90 colonne, le colonne più rilevanti sono: nome utente, momento temporale in cui è stato pubblicato il tweet, il testo del tweet, il numero di likes, il numero di retwitt.

4.1.2 Processamento dei dati

Una volta scaricati i dati, è stato usato il pacchetto "Sentiment Analysis", che si occupa della pulizia dei dati, trasformazione del testo, rimozione della punteggiatura, eliminazione delle stopwords o parole vuote.
Una volta pulito e preparato il dataset, è stata svolta la Sentiment Analysis utilizzando lo stesso pacchetto. Per i risultati dell'analisi, sono state utilizzate 3 categorie: 0 corrisponde ad un sentiment negativo, 0,5 se il sentiment è neutro, 1 se il sentiment è positivo.

Per la creazione dei wordcloud, è stato usato il pacchetto "tm" (Feinerer, K. Hornik, and D. Meye, 2008) dove i testi dei tweet vengono importati e organizzati in corpus (cioè una collezione di testi importati e organizzati ai fini dell'analisi). Dopo aver creato il corpus, sono stati implementati ulteriori metodi di pulizia dei dati, i quali hanno riguardato:

- Rimozione degli URL;
- Trasformazione del testo in caratteri minuscoli;
- Rimozione della punteggiatura e numeri;
- Eliminazione delle stopwords o parole vuote;
- Riduzione dei testi alle radice tramite le procedure di stemming.
- Rimozione delle emoji

Dopo questa fase di pulizia, si ottiene il dataset pronto per essere analizzato.

Per la procedura dettagliata dell'analisi, si rimanda all'appendice A, dove sono riportati gli script R usati nelle varie fasi dell'analisi.

4.2 Analisi dei tweet in italiano

La ricerca è stata svolta selezionando gli utenti Twitter che utilizzano la lingua italiana. La parola chiave utilizzata per scaricare i dati è stata l'hashtag pride. È stato scelto questo

hashtag per ottenere tweet in relazione al pride month di Giugno 2022 al fine di avere un'analisi più precisa rispetto al tema scelto.

I dati ottenuti si riferiscono al periodo 14 giugno - 2 luglio, per un totale di 98720 tweet.

Tra questi, il numero di retweet è 44513, quindi il 45% del totale dei tweet scaricati.

Il numero di utenti che hanno scritto almeno un tweet corrisponde a 15032.

Dalla figura 4.3, si può notare che i giorni in cui sono stati scritti più tweet sono il 25 giugno (14491) e il 26 giugno (9976).

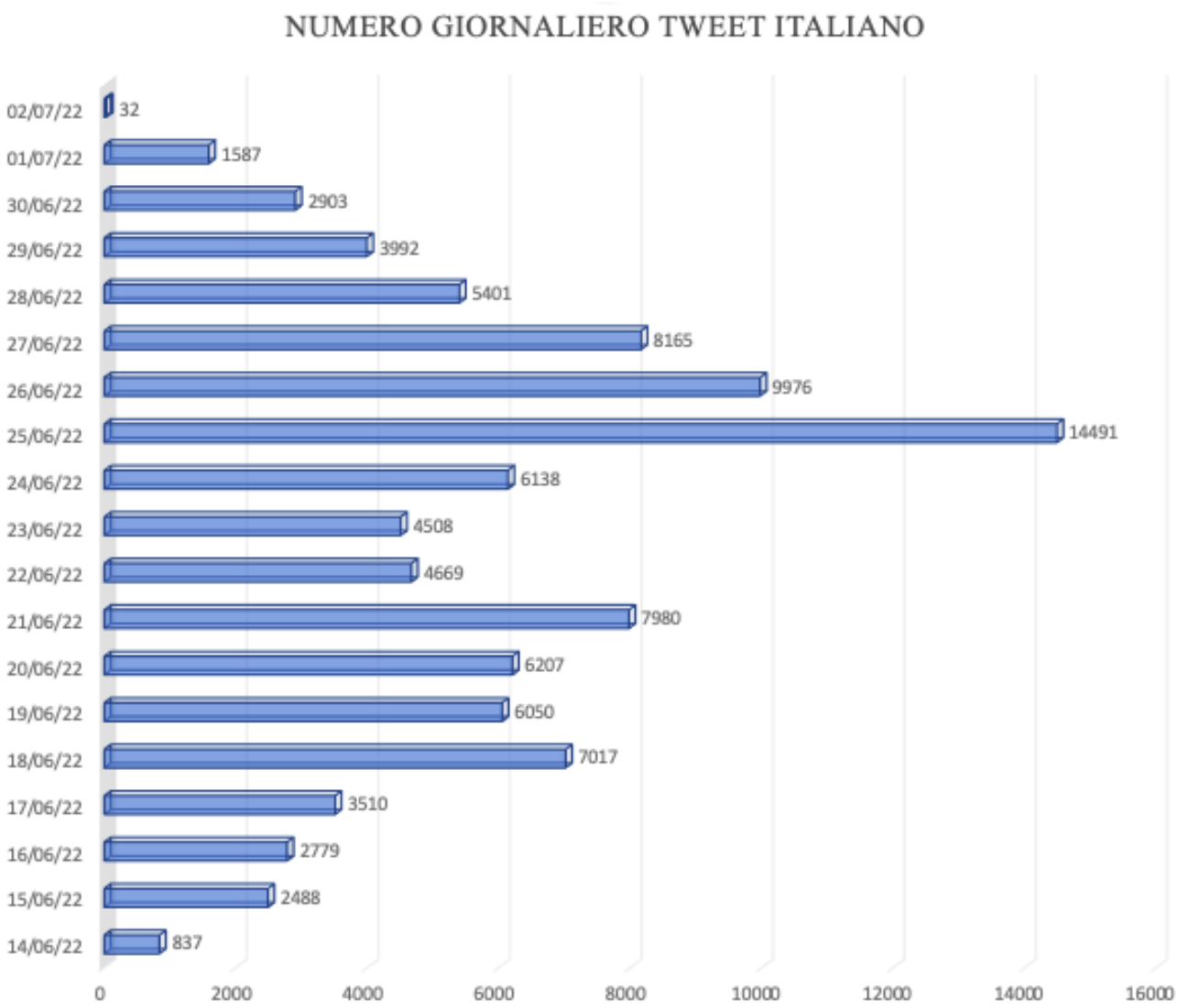

Figura 4.3 Distribuzioni giornaliere tweet in italiano

Tra i tweet in italiano alcuni esempi sono riportati nel seguito:

"Orgogliosi di supportare il Pride Lo Juventus Store di Milano è pronto
https://t.co/OQo6bxfYX9 Il ricavato sarà raddoppiato e destinato al Rainbow Social Fund fo ndo di solidarietà nato con i proventi del @MilanoPride #MoreColorfulTogether #DifferencesMakeTheDifference https://t.co/LbyEhkYHbj"

"Le persone #LGBT non si nascondono più perché non hanno nessun motivo per doversi nasco ndere, fatevene una ragione. A nascondersi è tempo che siano gli omofobi #Pride https://t.co/K YzrXDLt4B"

"La transfobia è un problema reale e urgente, la transfobia distrugge i sogni, ti fa perdere il lavor o, ti nega un affitto, ti urla per strada, ti tira uno schiaffo. La transfobia ti isola dalla società, per l a quale non esisti. La #transfobia esiste, ti uccide. #Pride"

Il tweet con il più alto numero di retweet (14346) è stato:

"Roma, Italia… https://t.co/qOaHPcfqaN" riguardante la fontana di Trevi e alcuni ragazzi con la bandiera arcobaleno.

Ora è interessante vedere quelle che sono le parole più frequenti, attraverso l'introduzione della Document Term Matrix. La DTM è una matrice le cui righe corrispondono ai tweet e le colonne a ciascun termine presente nei testi. Nella matrice sono perciò presenti i conteggi di ogni termine per ogni tweet. Per costruire la matrice ci appoggiamo alla libreria "tm" (Feinerer, K. Hornik, and D. Meye, 2008). A questo punto si può utilizzare il comando "DocumentTermMatrix" che richiede un argomento di tipo vector:

Possiamo ora vedere i termini maggiormente frequenti, ad esempio i primi 10:

pride	**gay**	**torino**	**lgbt**	**month**
8429	1924	1115	645	625
diritti	**persone**	**mese**	**arcobaleno**	**transfobia**
608	592	571	550	437

Le parole più frequenti sono riportate in un worldcloud, come riportato in figura 4.4

Figura 4.4 Wordcloud con parole più frequenti in italiano

Dal wordcloud, si può notare che le parole più ripetute sono “pride”, “gay”, “Torino” (dove si è tenuta la parata il 18 giugno 2022), “month”, “lgbt” e “mese”.

4.2.1 Risultati sentiment analysis per i tweet in lingua italiana

Per quanto concerne la sentiment analysis, sono stati selezionati 10000 tweet sul totale di quelli scaricati, per una questione di limite della capacità di memoria del software R nel pc.

Utilizzando il pacchetto “*SentimentAnalysis*”, si può vedere, dalla figura 4.4, che per quanto concerne i tweet in italiano, il sentimento è negativo; infatti, le barre sono più spostate nella prima metà rispetto allo 0,5 che rappresenta la neutralità del sentimento.

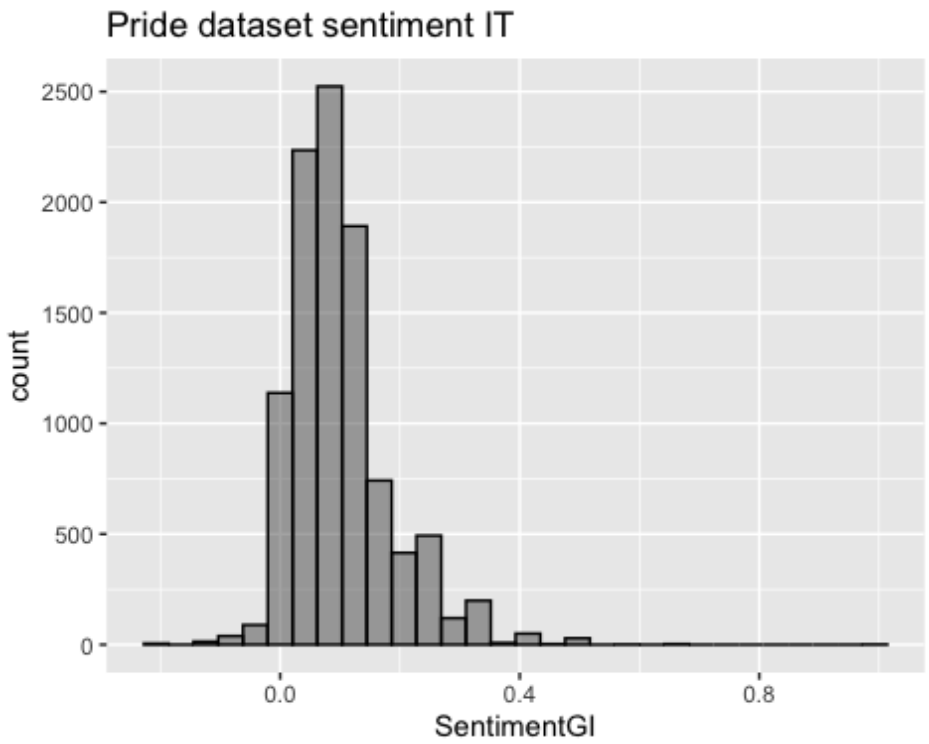

Figura 4.5 Sentiment analysis per tweet in italiano

Le 10 parole positive (parola e relativa frequenza assoluta):

pride (8429); diritti (608); arcobaleno (550); essere (425); vita (393); colori (307); grazie (272), libertà (271); sostengono (177); incoraggino (174).

Come si può osservare tra le 10 parole associate ai tweet con sentiment positivo ci sono la parola *pride*, che rappresenta l'orgoglio di appartenere alla comunità LGBTQIA+. La parola *diritti* è la seconda più ripetuta ed evidenzia una voglia di ottenere alcuni diritti secondo la comunità che ad oggi non sono concessi.
L'*arcobaleno* è la terza parola più ripetuta ed è simbolo della comunità.
Essere è un verbo, ma se lo contestualizziamo al tema del Pride month, riguarda la rappresentazione libera di sé stessi, che a volte è soffocata dalla paura del giudizio delle persone che ci circondano.
Vita è una parola molto ripetuta, il significato è positivo perchè la vita è un dono e ci permette di vivere esperienze che lasciano sempre una lezione.
Colori è un'altra parola ripetuta molto e fa riferimento alla parola arcobaleno.
Grazie è una parola di riconoscenza che nel contesto del Pride viene usata dopo aver partecipato alla parata ed aver condiviso dei momenti di allegria con altre persone.

Libertà è una parola che viene ripetuta spesso in quanto nella comunità viene rivendicata la libertà di essere sé stessi.
Sostengono ed *incoraggino* sono parole che fanno riferimento all'appello della comunità verso alcuni personaggi pubblici e politici in merito a temi oggetto di discussione.

Esempi di tweet positivi:

"□□ *Sempre una gioia unirsi al fiume rainbow del #TorinoPride*
La strada per i diritti di tutte e tutti è ancora lunga.
Ma siamo sempre di più.
Grazie a chi c'era □
#pride #loveislove https://t.co/b3Oio2IzpM"
"

"Il popolo del Pride invade Parma, l'urlo per i diritti: "Non siete soli: siamo noi la vostra famiglia": Musica, colori, danze. Centinaia di persone sfilano per le vie del centro per i diritti Lgbtqi+. https://t.co/SbEbPJCcw0"

Elodie conquista il Pride di Roma: "Abbiamo tutti gli stessi diritti, viviamo con amore non con odio": https://t.co/wN0ILen79A ... https://t.co/81lOF7NjuT

Le 10 parole negative (parola e relativa frequenza assoluta):

transfobia (437); problema (244); bisogno (244); omofobi (200); sequestrato (175); mai (173); pedofilia (164); condanna (156); insostenibile (150); perdere (149).

Al contrario, è possibile identificare quali siano le parole associate ai tweet con sentiment negativo, *Transfobia* è la parola più ripetuta, sotto il campo semantico della paura può essere affiancata ad *omofobi*. Seguono *problema* e *bisogno,* che riguardano la presenza di qualche nodo da sciogliere e qualche azione da intraprendere per colmare un desiderio o un diritto.

Sequestrato è una parola molto ripetuta che può riguardare il sequestro di una persona facente parte della comunità; può esservi affiancata la parola *condanna* che esprime un giudizio negativo verso una situazione o una persona.
Mai è una negazione del tempo che può esprimere la mancanza di un'azione o di un evento nel tempo.

Pedofilia è una parola che esprime un problema verso l'abuso dei bambini e che se riportato nel Pride può essere utilizzato per comparare la comunità LGBTQIA+ con scandali emersi in altre comunità, come ad esempio quella ecclesiastica.
Insostenibile è una parola di valore negativo che rappresenta la stanchezza di sopportare una situazione.
Perdere è una parola negativa che riguarda lo smarrimento di qualcosa, se contestualizzato nel Pride potrebbe riguardare la sottrazione di un diritto.

Esempi di tweet negativi:

"*Parata del Gay Pride in Germania aggredita da «uomini di origine meridionale - RENOVATIO 21: Una parata del gay pride in Germania è stata attaccata da un gruppo di giovani uomini di «origine meridionale». Così è stato riportato su vari ... https://t.co/g4TFIok0pa*"

"*La transfobia è un problema reale e urgente, la transfobia distrugge i sogni, ti fa perdere il lavoro, ti nega un affitto, ti urla per strada, ti tira uno schiaffo. La transfobia ti isola dalla società, per la*
Quale non esisti
La #transfobia esiste, ti uccide. #Pride"

"Offese e prese a sputi mentre tornano a casa dopo il Pride https://t.co/VKha78iIyd via @LaStampa"

4.3 Analisi dei tweet in inglese

Per valutare se esiste o meno una percezione simile sul tema anche al di fuori dei confini italiani, la stessa analisi è stata condotta su tweet riportati in lingua inglese.
Come precedentemente osservato, non è stato possibile rilevare anche la geolocalizzazione di tutti i tweet, quindi l'analisi si basa sull'assunzione che un tweet espresso in lingua inglese si riferisca ad un utente straniero.
È possibile che anche soggetti italiani usino la lingua inglese, ma per convenienza assusiamo questa generalizzazione.
La ricerca è stata svolta selezionando gli utenti Twitter che utilizzano la lingua inglese.
La parola chiave utilizzata per scaricare i dati è stata ancora l'hashtag pride.
I dati ottenuti si riferiscono al periodo 20 giugno - 2 luglio, per un totale di 97610 tweet.
Il numero di retweet corrisponde a 71717, ovvero il 73% dei tweet scaricati.
Il numero di utenti che hanno scritto almeno un tweet è 86580.
Dalla figura 4.6, si può notare che i giorni in cui sono stati scritti più tweet sono il 26 giugno (17902) e il 28 giugno (17452).

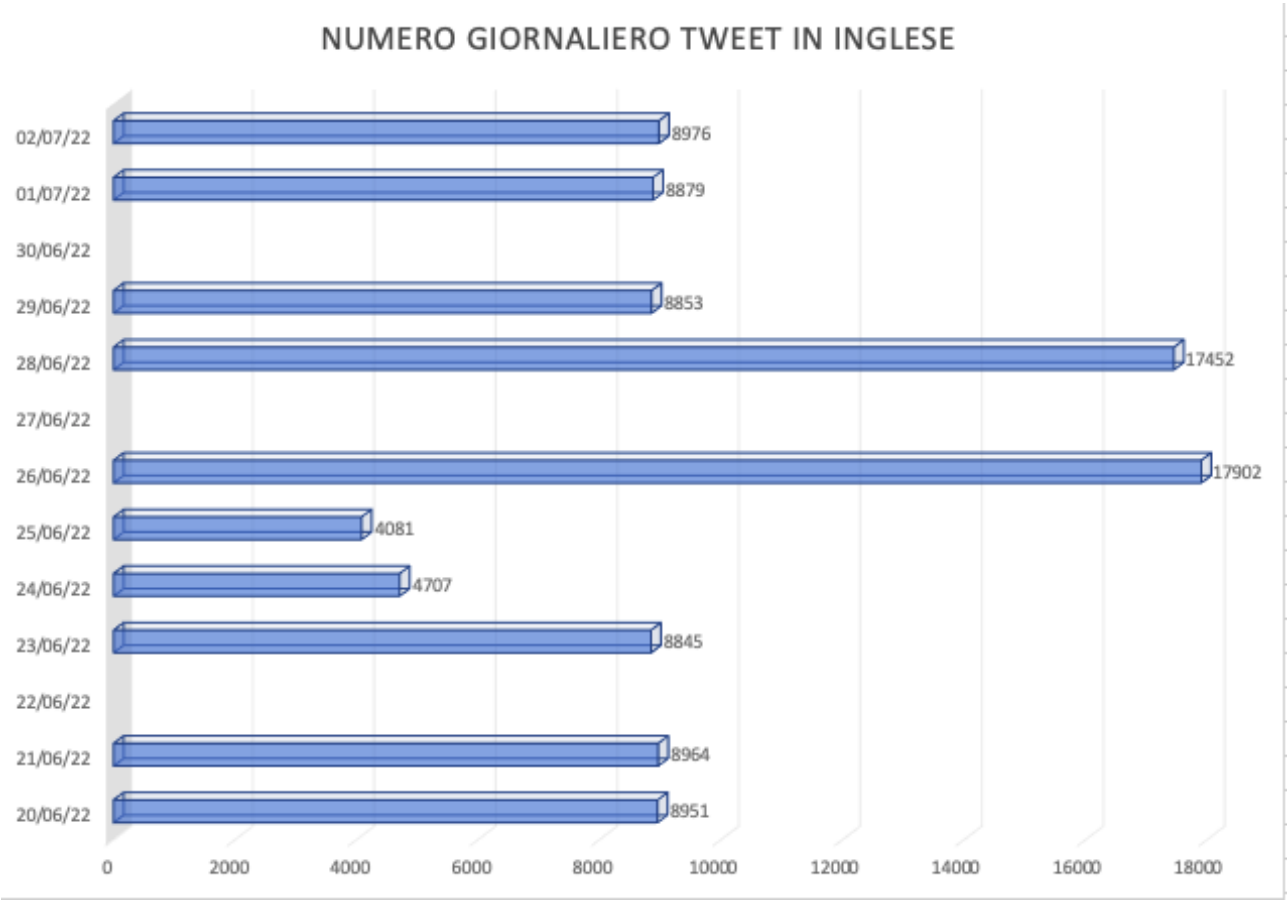

Figura 4.6 Distribuzioni giornaliere tweet in inglese

Tra i tweet in inglese alcuni esempi sono riportati nel seguito:

“Welcome to the United States. Where guns have more rights than women”

“Why #Pride is important #Pride2022 #PrideMonth https://t.co/UGqiSjHdm4”

“Stop being the one who makes all the effort. Sit back and let the ship sink.”

Il tweet con il più alto numero di retweet (165907) è stato:
“Welcome to the United States. Where guns have more rights than women”

Anche in questo caso, attraverso la matrice DTM, si è riuscito a scoprire le parole più ripetute, di seguito le prime 10. Nel seguito, a ciascuna parola viene associata la frequenza assoluta rilevata nella matrice.

pride	**month**	**happy**	**people**	**gay**
9482	2413	2326	994	625
like	**flag**	**Love**	**proud**	**Lgbtq**
543	464	460	401	375

Le parole più frequenti sono rappresentate per nel wordcloud in figura 4.7:

Figura 4.7 Wordcloud parole più frequenti in inglese

Dal wordcloud, si può notare che le parole più ripetute sono “pride”, “month”, “happy”, “people”, “gay”, “people”.

4.3.1 Risultati sentiment analysis dei tweet in inglese

Una volta analizzato le parole più frequenti, per quanto concerne la sentiment analysis, sono stati selezionati 10000 tweet sul totale di quelli scaricati, come nel caso dei tweet in italiano per una questione di limite di capacità di memoria del software R nel pc.
Utilizzando il pacchetto #SentimentAnalysis, si può vedere, nella figura 4.6, che per quanto concerne i tweet in inglese, il sentimento è negativo; infatti, le barre sono più spostate nella prima metà rispetto allo 0,5 che rappresenta la neutralità del sentimento.

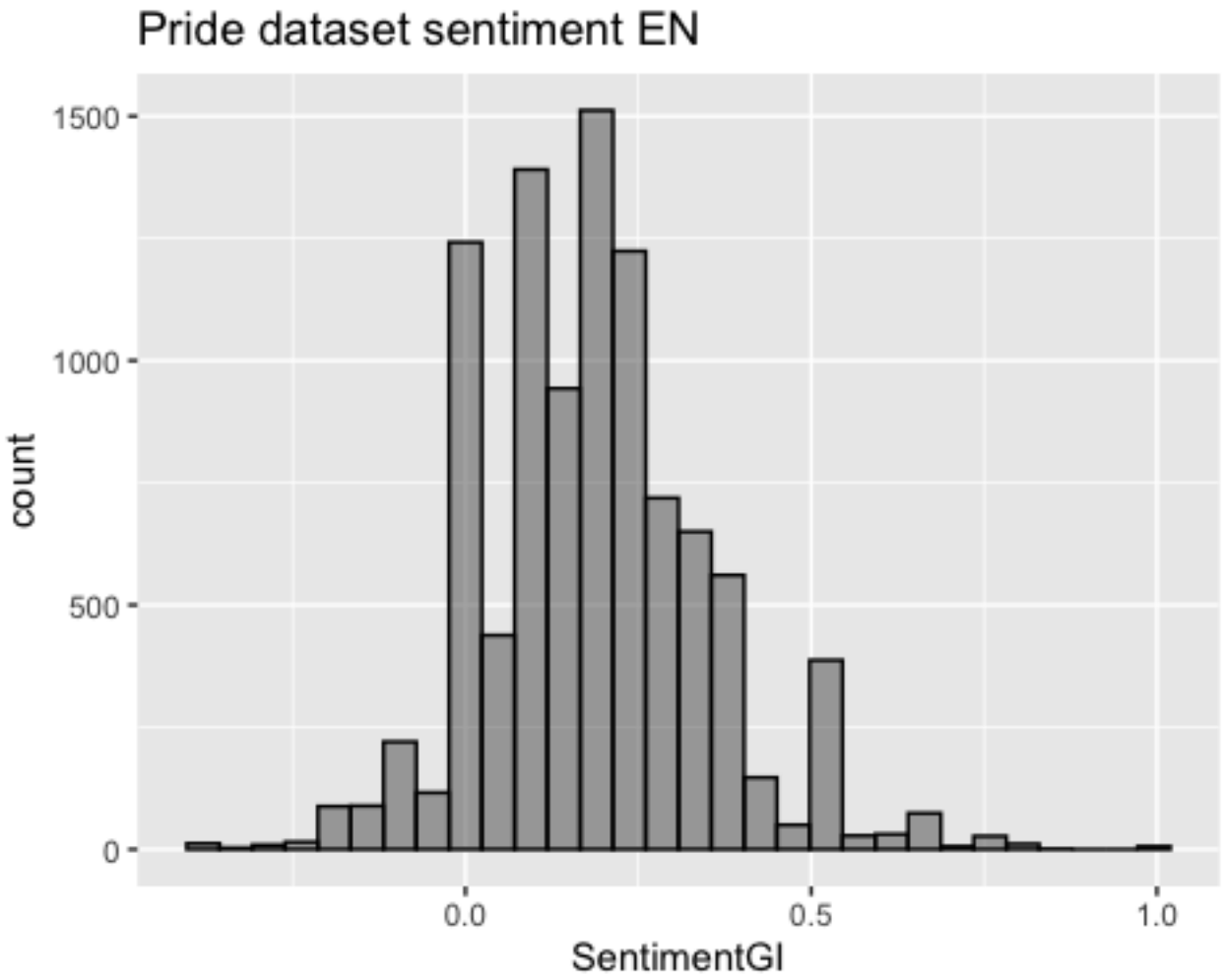

Figura 4.8 Analisi del sentiment tweet in inglese

Le 10 parole positive (parola e frequenza relativa assoluta):

Pride (9482); happy (2326); like (543); love (460); proud (401); anniversary (367); feel (358);
Amazing (312); community (297); support (243).

All'interno dei tweet con sentiment positivo, la parola ripetuta più volte è *pride* che supporta il movimento LGBTQIA+.
Successivamente vi è la parola *happy* che significa felice, ma può essere anche usata come augurio.
Like è la terza parola più ripetuta, testimonia qualcosa che piace.
Love è un'altra parola molto ripetuta che in questo contesto può significare l'amore per il prossimo, per gli altri. *Proud* è un'altra parola che sta per significare essere fiero di qualcosa o qualcuno.
Anniversary è un'altra parola che viene utilizzata in occasioni di memoria, spesso positive come un matrimonio, un fidanzamento, la celebrazione di un evento come può essere il Pride month.
Feel è una parola che sta per sentimento, provare un'emozione.
Amazing descrive qualcosa di bello, di straordinario, come può essere un gesto o una sensazione.
Community è una parola che significa un insieme di persone che si sono accomunate da qualcosa.
Support è una parola che significa l'approvazione nella scelta di qualcuno, se contestualizzata indica il supporto di qualche persona, comune o conosciuta verso la communità.

Esempi tweet positivi:

"Happy pride month! □⚧□□

A tribute to the community, and most of all, a tribute and celebration for trans and gender nonconforming individuals who started this movement. https://t.co/3wXjF1yOZ9"

"I think my favorite Pride song is Goodbye Earl"

“The Samuel Group of Companies standard of excellence extends to hiring and working with the best, brightest and most commitment professionals. Pride Month gives us a chance to engage the LGBTQ+ community to be proud of their courage, integrity, and person…https://t.co/c9WkMfgceW”

Le 10 parole negative (parola e relativa frequenza assoluta):

late (389); sorry (378); lose (291); forces (272); never (157); thrill (129); violence (102); banned (96); stop (91); discriminate (88).

All’interno dei tweet classificati con sentiment negativo, *Late* è la parola più ripetuta, sta a significare qualcosa in ritardo, in questo contesto potrebbe essere un’azione che la comunità si aspetta da tempo.
Sorry è una parola che viene usata per scusarsi da qualche azione sbagliata.
Lose è una parola che riguarda la perdita di qualcosa o qualcuno.
Forces è intesa come forza militare, che quindi porta scontri.
Never è una parola che significa il mancato accadimento di un evento o di un’azione.
Thrill è una parola che significa brivido, è una reazione ad un evento negativo.
Violence è la violenza, in merito ad un evento che ha portato scontri o attacchi verbali.
Banned significa espulso, qualcuno che si comporta in maniera negativa e non viene più ammesso.
Stop indica la richiesta di porre fine a qualcosa.
Discriminate è una parola che, se applicata in un contesto sociale, rivela una disparità di giudizio e di trattamento. Contestualizzandola al Pride month, vi è un problema di discriminazione di alcune persone della comunità da parte di individui non appartenenti ad essa.

Esempi tweet negativi:

“They are attacking pride events. They are protesting queer bars. They are storming Drag shows. The are calling for us to be rounded up. They are calling for us to be killed. Every bit of progress we’ve made is at risk, and so are we.”

“Wow! The Texas GOP just banned Gay Republicans from their convention — in the middle of Pride Month. The Log Cabin Republicans were shut out by members of their own party. I encourage the Log Cabin Republicans to switch their party affiliation. The Democrats don’t discriminate.”

“In this month of Pride, a warning to corporate washers.

Beware

Like us all, you have failed to read the small print. In your case it is the small print after LGB, the TQ+.

This small print is full of corporate hazard and you ignore it at your own risk.

A /1 https://t.co/QrEJGILurf”

Confronto tra tweet in italiano e tweet in inglese:

Dalle sentiment analysis condotte sui campioni dei tweet in italiano ed in inglese, si può notare una similarità nei risultati, infatti entrambe hanno un sentiment negativo.
Le parole positive in comune tra i due campioni sono *pride* e *sostegno.* Se accomunate hanno un significato molto completo in quanto queste persone danno sostegno al Pride.
Le parole negative in comune sono *mai* e *perdere, condanna* e *violenza.*

4.4 Caso studio: Sentiments comparison on Twitter about LGBT in US

Al fine di comprendere se l’analisi condotta in questo caso studio, possa trovare riscontro nelle evidenze empiriche prodotte dalla letteratura scientifica, riportiamo nel seguito la presentazione di un articolo recente che si occupa delle stesse tematiche.
L’articolo “*Sentiment comparison on Twitter about LGBT*” pubblicato nel 2022 da Aldinata, Soesanto, Chandra, Suhartono, ha come obiettivo capire il sentiment degli utenti americani rispetto alla comunità LGBT. La ricerca è stata condotta su persone che abitano nei 50 Stati americani. I tweet sono stati filtrati in merito alla parola chiave “LGBT” e raggruppati in base alla geolocalizzazione del tweet. L’analisi dei tweet è stata eseguita su un Training set composto da 27481 testi e su un Test set di 3584 testi.

Sono stati usati 5 differenti algoritmi per la sentiment analysis, ovvero TextBlob PatternAnalyzer, Naive Bayes, Linear Support Vendor Machine, regressione logistica e XGBoost.

Dalla figura 4.7 si può notare che la regressione logistica ha dato il tasso più alto di accuratezza con il 70.87%.

Algorithm	Precision	Recall	F1-Score
TextBlob PatternAnalyzer	0.61	0.60	0.59
Naive Bayes	0.7121	0.5929	0.6047
Linear Support Vector Machine	0.7041	0.6762	0.6835
Logistic Regression	**0.7233**	**0.7006**	**0.7087**
XGBoost	0.7385	0.6877	0.7009

Figura 4.9 Risultati prima del pre-processamento dei testi, fonte: Aldinata et al, 2022

Qui sotto nella figura 4.8, si hanno i risultati della sentiment analysis usando la regressione logistica. Emerge che l'84,72% dei tweet sono neutri, significa che la maggiorparte delle persone del campione hanno un giudizio neutro rispetto alla comunità LGBT. I tweet positivi sono al di poco sopra dei tweet negativi, rispettivamente 8,1% e 7,18%.

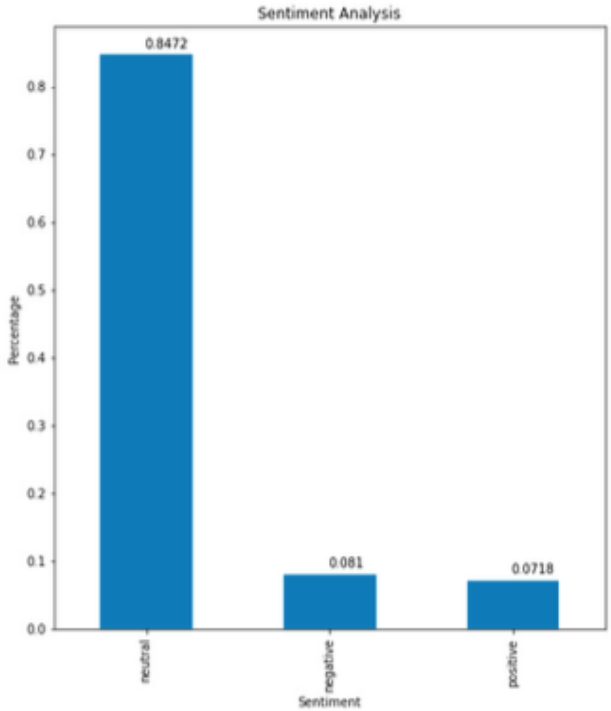

Figura 4.10 Grafico a barre della Sentiment Analysis, fonte: Aldinata et al, 2022

Rispetto alla sentiment analysis condotta per questa tesi, ci sono stati risultati diversi, infatti nell'articolo menzionato prevalgono i tweet neutri, mentre nella tesi sono maggiori i tweet con negativi. Va sottolineato che è stato utilizzato un algoritmo diverso nelle diverse analisi e anche i campioni sono differenti.

Conclusioni

L'obiettivo della presente tesi è l'analisi di testi, nello specifico di tweet, con l'utilizzo di tecniche di Text Mining che permettono di studiare la Sentiment Analysis. Il metodo utilizzato è stato: LASSO regularization, utilizzato dalla libreria in R attraverso il pacchetto "SentimentAnalysis".

Durante l'analisi si sono riscontrate alcune difficoltà, in particolare con il numero di tweet da poter utilizzare. Sebbene siano stati scaricati più di 90 mila tweet in lingua italiana ed altrettanti in lingua inglese, vi è mancata la possibilità di analizzarli tutti per troppo sforzo di memoria richiesto dal pc.

Le analisi pre-processing sono state complicate perchè se da un lato è stato relativamente semplice utilizzare il pacchetto "SentimentAnalysis" con il metodo LASSO regularization, dall'altro non è facile comprendere come vengono processati i testi in fase di pulizia.

Va sottolineato che sebbene l'indagine fosse di capire la percezione del Pride in Italia ed all'estero, non vi è la certezza che i tweet in lingua inglese fossero scritti solo da persone straniere, in quanto non è stata implementata la geolocalizzazione al momento del download dei tweet, per esempio potrebbe essere che una persona italiana abbia scritto un tweet in inglese.

Un altro punto da sottolineare è che il campione utilizzato non è rappresentativo della popolazione italiana, i fattori sesso e fascia d'età non sono stati presi in considerazione al momento di scaricamento dei tweet, dunque i risultati ottenuti, che dimostrano avere sentiment negativo, non è sicuro che rispecchino il pensiero della totalità della popolazione Questa considerazione va attribuita anche ai tweet in lingua inglese, a maggior ragione per il campione che è più ampio, dal momento che vengono considerati più Stati.

Vi è una necessità di conoscenza del problema, per poter interpretare i risultati forniti, in questo elaborato sono state presentate le origini del Pride Month, al fine di contestualizzare il tema, per essere di supporto nell'analisi dei dati.

La fase di analisi testuale ha poi prodotto alcune difficoltà in merito alla selezione finale di parole che si è verificata, nonostante sia stata implementata la pulizia dei testi, vi sono state alcune parole che risultavano molto ripetute ma non significative per il contesto.

I risultati ottenuti dall'applicazione della metodologia per la sentiment analysis, ha prodotto risultati interessanti per l'indagine oggetto di studio.
Tra le parole positive sia in italiano che in inglese, le frequenze sono più alte rispetto alle parole negative italiane ed inglesi. Infatti, le parole come pride, happy, love per quanto concerne l'inglese e pride, diritti, arcobaleno per quanto concerne l'italiano, hanno frequenze più alte.
Per quanto riguarda i grafici proiettati dalla sentiment analysis, sia per i tweet in italiano che per i tweet in inglese, i sentiment risultano negativi, verso la neutralità, ma sicuramente non positivi.
Ciò contrasta con quanto risulta dalle frequenze di parole positive rispetto alle parole negative.

Si consideri che si è analizzato solo il testo dei tweet, il periodo di download dei dati è stato circoscritto ad uno specifico periodo dell'anno, in questo caso giugno perché è il mese dove viene festeggiato il Pride ed avvengono le sfilate delle comunità LGBTQIA+.
Le opinioni possono essere influenzate dagli avvenimenti che circondano le persone. Il web è spesso condizionato da eventi esogeni che condizionano gli utenti a schierarsi in una posizione rispetto ad un'altra.

Si precisa che in questo elaborato si è voluto evidenziare come delle informazioni possono essere estratte dai social network, senza costi e senza la necessità di saper utilizzare degli strumenti in modo approfondito.
Infatti, i dati sono facilmente reperibili, anche se non necessariamente rappresentativi.
È quindi possibile selezionare uno specifico argomento e ottenere i testi pubblicati spontaneamente dagli utenti.
Questa caratteristica vi è meno nei questionari tradizionali, in quanto il soggetto deve rispondere a domande già stabilite precedentemente, quindi la scelta della metodologia di raccolta dei dati va ad influenzare i risultati.
La scelta di combinare differenti dati ottenute da diverse fonti differenti potrebbe essere efficiente per osservare il fenomeno oggetto di studio sotto diversi punti di vista.

APPENDICE:

A1. Il download dei dati

Per poter scaricare i dati da Twitter è necessario seguire alcuni passaggi che permettono di accedere alle informazioni di cui abbiamo bisogno. Innanzitutto, per la fase di autenticazione abbiamo utilizzato un codice R creato appositamente per scaricare dati da Twitter.
Quindi è stato installato il pacchetto RTweet ed inserito le funzioni che permettono l'autenticazione e l'ottenimento delle autorizzazioni per accedere alla raccolta dei tweet. Il comando "search_tweets" è quello che viene utilizzato per scaricare i tweet. In questa indagine è stato scelto il tweet con l'hashtag Pride e le lingue italiano ed inglese. Ogni giorno venivano scaricati 9000 tweet per la lingua italiana e per la lingua inglese, che è il massimo consentito per scaricarli gratuitamente.

```
install.packages("rtweet")
api_key <- "SUEx2gj50xiMf1MlKjawcClt9"
api_secret_key <- "BzrEv3f7QKs20w0Ss0BjIWPbFu2A7FiUYRS2gXs1AyPiMKKLK8"
access_token <- "3013069619-pWd0LwYmIEtY60QcQ4cA1eloRgZ99y1cJi7bpgg"
access_token_secret <- "cGpoNLZoibYN9T4XvI4ltmnIA97qhBfC0sGT2QbgJykva"
token <- create_token(app = "AlbertoBresca", consumer_key = api_key, consumer_secret = api_secret_key, access_token = access_token, access_secret = access_token_secret )

#inserire gli hashtag che si vogliono scaricare

Pride_search_df1 <- search_tweets(q="pride", n=9000, lang="it")
Pride_search_df2 <- search_tweets(q="pride", n=9000, lang="en")
```

Il dataframe dei dati è composto da 90 variabili, ma le più importanti sono:

- Testo del tweet
- Se il tweet è stato inserito tra i favoriti
- N° di preferiti ricevuti

- Il nome dell'utente che eventualmente riceve il twet
- Data e ora di invio
- Se il tweet è stato troncato o meno
- L'ID dell'utente
- L'ID dell'utente che eventualmente riceve il tweet
- Nome utente che lo ha mandato
- N° di retweet ricevuti
- Se il tweet è un retweet di un altro utente

A2. Il pre processamento per la text analysis

È necessario rimuovere le stopwords.

```
# Remove stopwords

Tweet_ita$text <- as.character(Tweet_ita$text)
Tweet_ita$text <- Tweet_ita$text%>%str_remove_all(" ?(f|ht)(tp)(s?)(://)(.*)[.|/](.*)")
Tweet_ita$text <- Tweet_ita$text%>%str_remove_all("@[[:alnum:]_]{4,}")
Tweet_ita$text <- Tweet_ita$text%>%str_remove_all("#[[:alnum:]_]+")
Tweet_ita$text <- Tweet_ita$text%>%str_remove_all("^RT:? ")
Tweet_ita$text <- Tweet_ita$text%>%str_replace_all("\\\n", " ")
Tweet_ita$text <- Tweet_ita$text%>%str_to_lower()
Tweet_ita$text <- Tweet_ita$text%>%str_trim("both")
Tweet_ita$text <- Tweet_ita$text%>%str_replace_all("&amp;", "e")
Tweet_ita$text <- gsub("[[:punct:]]", " ", Tweet_ita$text)
Tweet_ita$text <- gsub("[^\x01-\x7F]", " ", Tweet_ita$text) #rimuove le emoji
Tweet_ita$text <- gsub("[[:digit:]]", "", Tweet_ita$text)
```

Per creare il wordlcoud

```
#word cloud
docs <- Corpus(VectorSource(Tweet_ita$text[1:10000]))
a<-tm_map(docs, removeWords, stopwords("ita"))
dtm <- TermDocumentMatrix(a)
matrix <- as.matrix(dtm)
words <- sort(rowSums(matrix),decreasing=TRUE)
df <- data.frame(word = names(words),freq=words)
```

```
set.seed(1234)
wordcloud(words = df$word, freq = df$freq, min.freq = 50,
        max.words=80, random.order=FALSE, rot.per=0.35,
        colors=brewer.pal(4, "Dark2"))
```

#per trovare le parole più frequenti, richiamare “words”.

Stesso procedimento per le parole i tweet in inglese:

```
#inglese
Tweet_en$text <- as.character(Tweet_en$text)
Tweet_en$text <- Tweet_en$text%>%str_remove_all(" ?(f|ht)(tp)(s?)(://)(.*)[.|/](.*)")
Tweet_en$text <- Tweet_en$text%>%str_remove_all("@[[:alnum:]_]{4,}")
Tweet_en$text <- Tweet_en$text%>%str_remove_all("#[[:alnum:]_]+")
Tweet_en$text <- Tweet_en$text%>%str_remove_all("^RT:? ")
Tweet_en$text <- Tweet_en$text%>%str_replace_all("\\\n", " ")
Tweet_en$text <- Tweet_en$text%>%str_to_lower()
Tweet_en$text <- Tweet_en$text%>%str_trim("both")
Tweet_en$text <- Tweet_en$text%>%str_replace_all("&amp;", "and")
Tweet_en$text <- gsub("[[:punct:]]", " ", Tweet_en$text)
Tweet_en$text <- gsub("[^\x01-\x7F]", " ", Tweet_en$text) #rimuove le emoji
Tweet_en$text <- gsub("[[:digit:]]", "", Tweet_en$text)

#word cloud
docs_en <- Corpus(VectorSource(Tweet_en$text[1:10000]))
b <-tm_map(docs_en, removeWords, stopwords("en"))
dtm_en <- TermDocumentMatrix(b)
matrix_en <- as.matrix(dtm_en)
words_en <- sort(rowSums(matrix_en),decreasing=TRUE)
df_en <- data.frame(word = names(words_en),freq=words_en)

set.seed(1234)
wordcloud(words = df_en$word, freq = df_en$freq, min.freq = 50,
        max.words=80, random.order=FALSE, rot.per=0.35,
        colors=brewer.pal(4, "Dark2"))
```

#parole più frequenti: richiamare “words_en”

A3. Sentiment analysis:

Una volta scaricati tutti i tweet per il periodo di interesse, vanno aggregati in un unico grande dataset:

```
#unisco i dataset per lingua
dataset_en <- do.call("rbind", list(Pride_search_df2, Pride_search_df4, Pride_search_df6,
                                    Pride_search_df9, Pride_search_df11, Pride_search_df13,
                                    Pride_search_df15, Pride_search_df17, Pride_search_df19,
                                    Pride_search_df21, Pride_search_df23))


dataset_it <- do.call("rbind", list(Pride_search_df, Pride_search_df3, Pride_search_df5, Pride_search_df10,
                                    Pride_search_df12, Pride_search_df14, Pride_search_df16, Pride_search_df18,
                                    Pride_search_df20, Pride_search_df22, Pride_search_df24))
```

Successivamente attuo la sentiment analysis, con la funzione “analyzeSentiment”

```
#analisi del sentiment globale en
Pride_sentiment_inglese <- analyzeSentiment(Tweet_en$text[1:10000])

Pride_sentiment_italiano <- analyzeSentiment(Tweet_ita$text[1:10000])
```

Per utilizzare dei grafici che mostrano i risultati dell’analisi, è stato utilizzata la libreria “ggplot”

```
ggplot(Pride_sentiment_italiano, aes(SentimentGI)) + ggtitle("Pride dataset sentiment IT") +
  geom_histogram( bindwith=30, color="#000000", alpha=0.5)
```

```
ggplot(Pride_sentiment_inglese, aes(SentimentGI)) + ggtitle("Pride dataset sentiment EN") +
  geom_histogram( bindwith=0.05, color="#000000", alpha=0.5)
```

A.4 Analisi descrittiva

Al fine di trovare il numero di tweet al giorno, numero di utenti, i tweet con più likes, quelli con più retweet:

```
#numero giornaliero di tweet
giorno <- as.Date(Dataset_italiano$created_at)
dayen <- as.Date(dataset_en$created_at)

serieit <- table(giorno)
serieit
serieen <- table(dayen)
serieen

#fav count e retwitt in inglese
A <- dataset_en[, c(5,13)]
Fav_count <- sort(A$favorite_count, decreasing = T)
Most_Q <- A[order(A$favorite_count,decreasing= T),]
View(A)

B <- dataset_en[, c(5,14)]
Fav_retwitt <- sort(B$retweet_count, decreasing = T)
Most_R <- B[order(B$retweet_count,decreasing= T),]
Most_R

#fav count e retwitt in italiano

C <- Dataset_italiano[, c(5,13)]
Fav_count <- sort(C$favorite_count, decreasing = T)
```

```
Most_c <- C[order(C$favorite_count,decreasing= T),]
View(C)

#numero di retweet
table(dataset_en$is_retweet)
table(Dataset_italiano$is_retweet)
D <- Dataset_italiano[, c(5,14)]
Fav_ret <- sort(D$retweet_count, decreasing = T)
Most_d <- D[order(D$retweet_count,decreasing= T),]
pl

#numero utenti
n_distinct(dataset_en$user_id)
n_distinct(Dataset_italiano$user_id)
```

Bibliografia:

Aldinata, Soesanto, Chandra, Suhartono (2022), *"Sentiments comparison on Twitter about LGBT"*

Blei et al., (2003), "*Latent Dirichlet Allocation"*

Boyd (2007), "*Social Network Sites: Definition, History, and Scholarship"*

Carraher, Parnell, McClure & Sullivan, (2006), *"Entrepreneurial Service Performance and Technology Management: A Study of China and Japan"*

Censis (2022), *"I MEDIA DELLE CRISI"*

Colombo (2013), *"Potere socievole"*

Cohen (2009), *"Is There A Difference Between Social Media And Social Networking?"*

Crone, Lessmann e Stahlbock, (2006), "*The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing"*

Dalghren, (2002), *"In Search of the Talkative Public: Media, Deliberative Democracy and Civic Culture"*

Deng, Barman-Adhikari, Lee, Dewri, Bender (2020), "*Substance use and sentiment and topical tendencies: a study using social media conversations of youth experiencing homelessness"*

Edosomwan, Kalangot Prakasan, Kouame, Watson, Seymour, (2011) "*The History of Social Media and its Impact on Business"*

Goffman (1959), *"The presentation of self in everyday life"*

Grimmer, Stewart (2013), "*Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts"*

Günther & Quandt, (2016), "*Word Counts and Topic Models Automated text analysis methods for digital journalism research"*

Hartshorn (2010), *"5 Differences Between Social Media and Social Networking"*

Junco, Heibergert, & Loken, (2011), "*The Effect of Twitter on College Student Engagement and Grades. Journal of Computer Assisted Learning, 27, 119-132."*

Knomo, Ndukne e Daniel (2020), *"Social Network and Sentiment Analysis: Investigation of Students' Perspectives on Lecture Recording"*

Kukulska-Hulme, Agnes (2010), *"Learning cultures on the move: where are we heading?"* Journal of Educational Technology and Society, 13(4) pp. 4–14.

Kwartler, (2017), "*Text Mining in Practice with R"*

Leopold & Kindermann, (2002), "*Text Categorization with Support Vector Machines. How to Represent Texts in Input Space?"*

Livingstone, S. (2008), *"Taking Risky Opportunities in Youthful Content Creation: Teenagers' Use of Social Networking Sites for Intimacy, Privacy and Self-Expression. New Media & Society", 10, 393-411*.

Metcalf, (2020), "*The History of PrideHow Activists Fought to Create LGBTQ+ Pride"*

Nations, D. (2010), "*Web Trends*"

Pröllochs N, Feuerriegel S, Neumann D, (2018), *"Statistical inferences for polarity identification in natural language"*

Proksch & Slapin, (2009), *"Ideological Clarity in Multi-Party Competition: A New Measure and Test Using Election Manifestos"*, *British Journal of Political Science*.

Ritholz, (2010), *"The history of Social Media"*

Roberts et al., (2014), "*Structural Topic Models for Open-Ended Survey Responses"*

Siddiqui, Singh, (2016), *"Social Media its Impact with Positive and Negative Aspects"*

Stier, Bleier, Lietz e Strohmaier (2018), *"Election Campaigning on Social Media: Politicians, Audiences, and the Mediation of Political Communication on Facebook and Twitter"*

Tim Berners-Lee, (1999), *"Weaving the Web"*

Tomlinson, (1999), *"Globalization and Culture"*

Yau, Reich, (2019), *"It's just a lot of Work: Adolescents' self-presentation norms and practices on Facebook and Instagram"*

Yu, Duan, Cao (2012), *"The impact of social and conventional media on firm equity value: A sentiment analysis approach*"

Sitografia:

https://www.merriam-webster.com

https://www.statista.com

www.r-project.org