POLITECNICO DI TORINO

Corso di Laurea Magistrale
in Ingegneria Gestionale

Tesi di Laurea Magistrale

La politica e l'informazione nei Social Network: analisi della diffusione delle Fake News legate al Covid-19

Relatore
Prof. Carlo Cambini

Correlatori
Prof. Luca Cagliero
Prof. Luca Vassio

Candidata
Marinella Di Pierro

Anno accademico 2020/2021

SOMMARIO

INDICE DELLE FIGURE

INDICE DELLE TABELLE

Riassunto

L'avvento di Internet e dei Social Network ha mutato radicalmente la fruizione dell'informazione: i media tradizionali hanno lasciato spazio alle piattaforme algoritmiche dove flussi di notizie ufficiali e non, si mescolano nell'homepage di ciascun utente. Il diffondersi dell'epidemia da Coronavirus e l'incremento di teorie complottiste ha estremizzato gli effetti negativi di tale scenario informativo, dando vita alla circolazione di una quantità eccessiva di notizie false, prevalentemente sui Social Network.

In un contesto in cui la diffusione di informazione e la creazione di contenuti affidabili da parte delle personalità più seguite sui social sembra essere l'unica arma contro la disinformazione, le personalità politiche ricoprono un ruolo di estrema responsabilità.

La domanda da cui è scaturito questo lavoro di tesi magistrale è stata proprio questa: in questo scenario, qual è stato il ruolo delle personalità politiche in Italia?

L'obiettivo di tale lavoro, pertanto, è stato inizialmente quello di evidenziare gli influencer politici italiani e le coalizioni maggiormente attive e seguite sui Social Network durante il periodo della prima ondata di Covid-19. Sono stati individuati gli influencer politici che hanno creato più contenuti sulle tematiche legate al Coronavirus e che hanno ottenuto un elevato livello di engagement. Successivamente, è stata effettuata una analisi di caratterizzazione della presenza di tematiche riferite alle Fake News legate al Coronavirus nei testi dei post pubblicati dai profili politici e nei commenti associati a tali post. L'analisi si è soffermata sulle Fake News dichiarate dall'Autorità per le Garanzie nelle Comunicazioni (AGCOM). L'obiettivo ultimo è stato quello di osservare le similarità tra i topic dei contenuti testuali creati dai profili politici e dai commentatori estratti utilizzando il modello probabilistico Latent Dirichlet Allocation (LDA).

Tutto ciò è stato svolto confrontando due dei Social Network più popolari: Instagram e Facebook. Sono stati infatti analizzati i dati raccolti dal team SmartData@Polito su tali piattaforme, scegliendo di focalizzare l'attenzione sul periodo di riferimento (Gennaio-Giugno 2020) e sui contenuti creati da personalità politiche, caratterizzandone i post e i commenti e like ricevuti.

Abstract

Internet and Social Networks have radically changed the use of information: traditional media have given way to algorithmic platforms where official and non-official news flows are mixed on the homepage of each user. The spread of the Coronavirus epidemic and the increase in conspiracy theories has exacerbated the negative effects of the information scenario, fostering the circulation of an excessive amount of false news, mainly on Social Networks.

The dissemination of information and the creation of reliable content by the most followed personalities on Social Networks seems to be the only weapon against disinformation, so political figures play a role of extreme responsibility. The starting research question that motivates this Master Thesis is: what was the role of political personalities in the Italian context?

Firstly, the goal of this work is to highlight the Italian political influencers and the most active and followed coalitions on Social Networks during the first wave of the Covid-19. The political influencers who have created more contents on the issues related to the Coronavirus and who have obtained a high level of engagement have been identified.

Subsequently, a characterization analysis of the presence of fake news related to Coronavirus was carried out in the texts of the posts published by political profiles and in the comments associated with these posts. The analysis focuses on the Fake News declared by the Authority for the Guarantee of Telecommunication (Agcom). Lastly, I observed the similarities between the topics extracted from the textual content created by political profiles and commentators, using the probabilistic model Latent Dirichlet Allocation (LDA). Two of the most popular Social Networks are compared: Instagram and Facebook. I analysed data collected by the SmartData@PoliTO team on these platforms. The attention is focused on the period between January 2020 and June 2020 and on the content created by political personalities, characterizing the posts, comments and likes they received.

Capitolo 1

Il fenomeno delle Fake News durante il periodo Covid-19

Questo capitolo ha l'intento di introdurre il contesto di informazione e disinformazione attivo in Italia e discutere di come si è evoluto il fenomeno "infodemico" nel periodo relativo alla prima ondata della pandemia da Covid-19. Verrà anche introdotto l'argomento relativo al ruolo che hanno i Social Network nella diffusione dell'informazione. I Social Network saranno al centro dell'analisi presentata nei capitoli successivi.

1.1. L'informazione in Italia

1.1.1. Il consumo di informazione in Italia

Con l'avvento di Internet e dei mezzi di comunicazione online, l'informazione ha visto completamente mutare il suo modello di fruizione. L'incremento delle applicazioni digitali in circolazione ha fortemente aumentato in maniera parallela sia la domanda di informazione, sia la sua offerta. Dal punto di vista della domanda, infatti, Internet e le piattaforme online hanno dato la possibilità agli utenti di accedere facilmente e gratuitamente agli spazi informativi. Dal punto di vista dell'offerta, sono ad oggi presenti molte più fonti informative della sola televisione, radio e giornali degli anni pre-Web.
Nell'ultimo "Rapporto sul consumo di informazione" dell'Autorità per le Garanzie nelle Comunicazioni (1) è stato riscontrato che la principale fonte informativa in Italia è la televisione con il 48,2%, ma il dato in forte crescita è la percentuale dei cittadini italiani che scelgono di informarsi su Internet, il 26,3% nel 2017. Il Rapporto afferma che il 55% degli italiani accedono almeno ad una piattaforma online per raggiungere l'informazione, e se da un lato la consultazione delle fonti editoriali online si ferma al 39%, sia i social network che i motori di ricerca come Google vengono consultati dal 37% della popolazione. Come mostrano i dati quindi, ad oggi ci si informa prevalentemente attraverso fonti cosiddette "algoritmiche", come i social network e i motori di ricerca,

mentre le fonti tradizionali e i siti web resistono raggiungendo una minor percentuale degli italiani.

Si può certamente affermare che informarsi, oggi, risulta molto più facile e immediato. Tramite un solo click si possono raggiungere fonti informative di diversa provenienza e qualità. Si parla a questo proposito all'interno del "Rapporto sul consumo di informazione" del cosiddetto concetto di "**cross-medialità",** che consiste nell'uso congiunto di mezzi di informazione tradizionali e delle modalità offerte dal web. Se da un lato tale fenomeno aumenta l'esposizione all'informazione dando la possibilità all'utente di confrontare varie fonti a disposizione e quindi poter creare una opinione autonoma sulle notizie, dall'altro l'abitudine di simultaneità negli usi dei media e il cosiddetto '"**information overload**" (2) generato dalle troppe informazioni a disposizione, può provocare un consumo superficiale e poco accurato delle notizie e accrescere quindi il rischio di disinformazione.

1.1.2. La produzione di informazione in Italia

Nel Report "News vs. Fake nel sistema dell'informazione" redatto dall'AGCOM nel 2018 (3), viene introdotta la definizione della produzione di informazione, la quale "identifica il processo di realizzazione e offerta al pubblico di contenuti informativi aventi ad oggetto fatti, accadimenti, fenomeni, in altre parole notizie di qualsiasi genere. È il processo da cui dipendono la quantità, la varietà e la qualità dell'informazione che raggiunge i cittadini, sulla base della quale gli stessi formano le proprie opinioni e punti di vista."

I protagonisti della produzione di informazione sono gli editori, i quali ricoprono il ruolo di proprietari delle fonti informative e di finanziatori dell'attività di produzione e diffusione dell'informazione e le figure professionali occupate nelle strutture redazionali come i giornalisti, che costituiscono il fattore produttivo principale impiegato in questo processo. Un giornalista ha il compito di reperire, analizzare e approfondire le notizie e ne compone successivamente il contenuto informativo. L'output della loro attività coincide con articoli di giornale e riviste, servizi televisivi o radiofonici e, oggi, anche con post pubblicati sui social network.

Da un punto di vista quantitativo dell'informazione prodotta in Italia, è necessario effettuare una distinzione tra le fonti informative citate. Sul (3) sono stati riportati i risultati di un'analisi effettuata sui differenti mezzi di comunicazione. La fonte che offre un maggior contributo quantitativo sono i quotidiani (nazionali e locali), la seconda è Internet e segue per ultimo il contributo derivante dai canali televisivi e radiofonici.

Questi risultati sono stati rapportati ad un aspetto che risulta essere necessario: il numero di unità impiegate di fattore produttivo. Nella figura 1 è rappresentata una stima della funzione di produzione dell'informazione, intesa come la curva che, a parità di altre condizioni, esprime in ogni punto la relazione tra la forza giornalistica impiegata e il numero di output informativi prodotti.

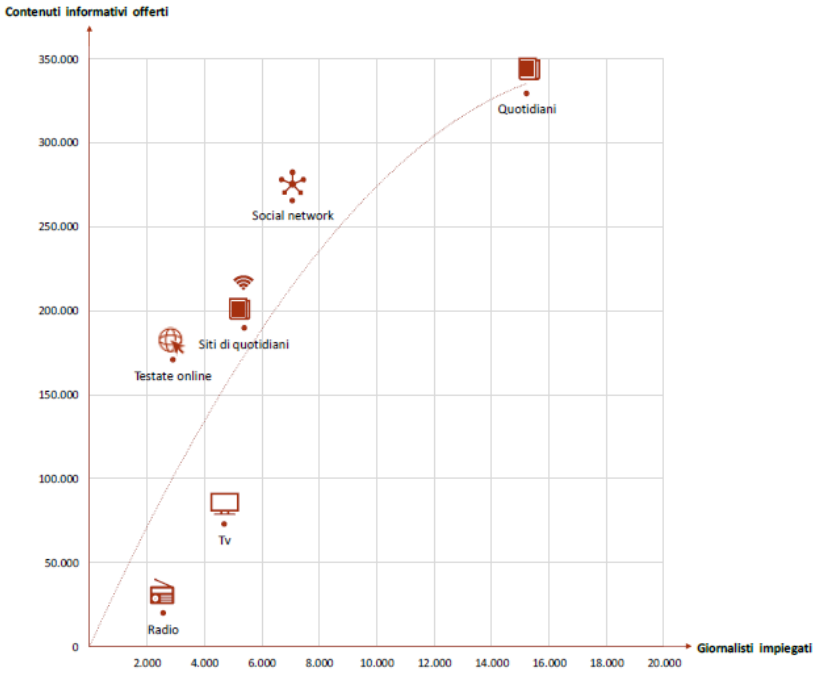

Figura 1 - Funzione di produzione dell'informazione (stima, valori medi mensili)
Fonte: elaborazioni Agcom su dati Volocom e aziendali (per i contenuti informativi offerti) e Osservatorio Agcom sul giornalismo – II edizione (per i giornalisti impiegati)

Tale figura suggerisce la possibilità di individuare tre distinte tipologie di mezzi di comunicazione che si trovano rispettivamente nelle tre posizioni in linea, al di sopra o al di sotto della curva. In particolare, l'Autorità ha riscontrato che le fonti informative online (testate online, siti quotidiani e social network) offrono una quantità di informazione maggiore a parità di risorse impiegate. Al contrario, i canali televisivi e radiofonici risultano produrre una quantità inferiore di informazione a parità di risorse. La terza categoria di mezzi di comunicazione, quella che si trova in corrispondenza della curva di produzione dell'informazione, è rappresentata dai quotidiani. È stato quindi interpretato che, per quanto riguarda i mezzi di comunicazione online, vi è un sovra utilizzo della forza giornalistica impiegata che può, quindi, compromettere la qualità del prodotto creato in termini di accuratezza e approfondimento. Nel caso delle emittenti televisive e

radiofoniche invece si evince una minore intensità produttiva che può essere sintomo di una maggiore accuratezza e approfondimento dell'informazione offerta. I quotidiani risultano i più vicini al valore medio in termini di intensità produttiva dei giornalisti i quali, dedicandosi esclusivamente alle mansioni più tipiche della professione, possono dedicarsi con maggiore cura alla qualità del prodotto che diffondono. Tale indice di intensità produttiva è stato messo inoltre in relazione con la reputazione del mezzo di comunicazione, cioè con l'affidabilità riconosciuta dai cittadini nei confronti di ciò che consumano. A questo proposito, nel Rapporto si riscontra che all'aumentare dell'intensità produttiva del giornalista, si percepisce meno qualità dell'informazione offerta e, quindi, una minore reputazione del mezzo.

1.1.3. Le tematiche dell'informazione

L'offerta di informazione in Italia è caratterizzata da una varietà di generi e tematiche. L'Agenzia per le Garanzie nelle Comunicazioni ha analizzato milioni di contenuti informativi italiani prodotti fino al 2017 e ha classificato cinque categorie riferite a diverse tematiche: le "hard news", termine che indica le notizie di cronaca politica e fatti di rilevanza nazionale, "cultura e spettacolo", "economia", "scienza e tecnologia" e "sport". Secondo i dati risalenti al 2017, in un mese medio il 40% dell'offerta informativa riguarda le "hard news", seguono notizie riguardanti la cultura e lo spettacolo e lo sport, che costituisce il 17% dell'informazione totale prodotta. Le categorie meno presenti nel sistema informativo italiano riguardano proprio quelle più specializzate che, come detto in precedenza, necessitano di figure aventi competenze specifiche: il cosiddetto "input produttivo" tale per cui si può ottenere un output di qualità.

È utile notare, a tal proposito, che l'informazione specializzata viene offerta in Italia in una percentuale minore rispetto al resto delle categorie e soprattutto viene prodotta da figure non specializzate nelle materie di riferimento. Ne consegue che, se per le categorie come le cosiddette "hard news", cultura e spettacolo e sport vi è un eccesso di offerta, per le tematiche più specialistiche come scienza e tecnologia ed economia la domanda non risulta sufficientemente soddisfatta sia da un punto di vista quantitativo, sia da un punto di vista qualitativo. Nella figura 2 è visualizzabile il livello di conoscenza specialistica dei giornalisti rispetto alla domanda potenziale dei cittadini, per categoria.

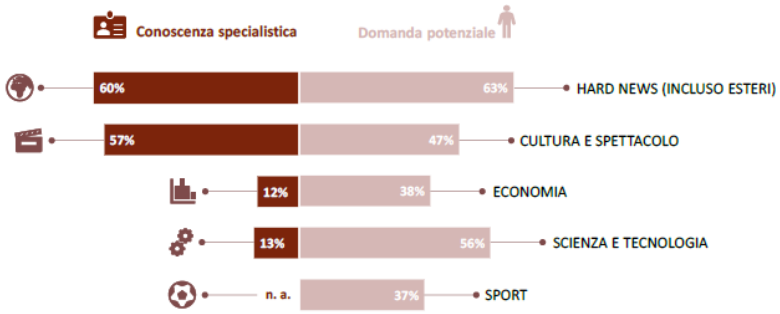

Figura 2 – Livello di conoscenza specialistica dei giornalisti e domanda potenziale dei cittadini per categoria. Fonte: Osservatorio Agcom sul giornalismo – II edizione ed elaborazioni Agcom su dati Reuters Institute for the Study of Journalism, Digital News Report 2017

1.2. Fake news e disinformazione

1.2.1. Definizioni

Il termine "**fake news**" è un'espressione che in lingua italiana è tradotta come "notizie false" e va principalmente ad indicare la diffusione di contenuti "fake" in Internet. Esso è genericamente utilizzato per indicare una vasta gamma di disturbi dell'informazione che possono essere distinti in macrocategorie per poter cogliere le sfumature dei problemi dell'informazione online.

In particolare, si parla di *misinformation* quando ci si riferisce a contenuti informativi non veritieri o inaccurati non creati con un intento doloso ma comunque atti ad essere recepiti come notizie su fatti reali come, per esempio, la satira/parodia, i contenuti fuorvianti e le false connessioni. Si parla invece di *malinformation* quando i contenuti informativi sono fondati su fatti reali ma sono contestualizzati in modo da poter essere anche virali e divulgati per arrecare danno ad una persona, un'organizzazione o un Paese o per affermare/screditare una tesi, come a titolo di esempio le fughe di notizie che favoriscono incitamento all'odio (hate speech) e molestie (online harassment) o l'amplificazione di notizie.

Per *disinformation* si intende la diffusione consapevole di contenuti informativi falsi, infondati, manipolati o riportati in maniera non veritiera, creati nel modo più opportuno affinché risultino verosimili nel contesto mediatico. Esempi di fenomeni disinformativi sono le "false contestualizzazioni (contenuti veritieri condivisi con false informazioni di contesto), contenuti veicolati da false fonti (che impersonificano fonti autentiche), contenuti creati in maniera artificiosa (totalmente falsi e infondati per ingannare e/o

danneggiare) e le notizie manipolate (informazioni veritiere manipolate in modo volutamente ingannevole)". Nello specifico, nel rapporto "Le strategie di disinformazione online e la filiera di contenuti fake" (4), vengono individuati gli elementi distintivi della disinformazione online che sono di seguito elencati:

- La falsità dei contenuti;
- Contagiosità, cioè l'attitudine a trasferire stati emotivi e percezioni tra gli utenti;
- L'intento doloso sottostante alla loro creazione;
- La motivazione politico/ideologica o economica di chi li crea per poi diffonderli;
- La diffusione degli stessi in maniera massiva;
- L'attitudine a produrre un impatto per il pluralismo informativo, cioè a generare effetti sulla formazione dell'opinione dei cittadini.

1.2.2. La filiera dei contenuti fake

L'immissione nel sistema informativo dei contenuti fake è concretizzata mediante un processo composto da quattro step che può essere definito una vera e propria "filiera dei contenuti fake" e prevede: la creazione del messaggio, la fase di produzione in cui il messaggio viene trasformato in un contenuto informativo online, la distribuzione mediante la quale il contenuto viene pubblicato e diffuso tra gli utenti e la valorizzazione del contenuto, fase che porta alla produzione o meno di un guadagno o dello scopo desiderato.

È possibile osservare i contenuti disinformativi sotto due punti di vista diversi: una componente soggettiva e una componente oggettiva del fenomeno. La componente soggettiva rappresenta la fonte da cui proviene il messaggio diffuso e i destinatari dello stesso, i cosiddetti soggetti coinvolti; la componente oggettiva coincide con l'oggetto del messaggio, cioè il contenuto di ciò che viene divulgato.

Dal punto di vista soggettivo, i soggetti coinvolti nelle quattro fasi della filiera dei contenuti fake possono essere numerosi e di vario genere. Come cita (4), "gli ideatori e gli esecutori del messaggio possono essere singoli individui, imprese editoriali e non, organizzazioni con finalità svariate, servizi di intelligence, governi o Stati". È importante sottolineare che invece nella fase di distribuzione possono contribuire meccanismi automatici come i "**bot"**, account falsi creati appositamente per ampliare il raggio di diffusione del contenuto fake. Nella stessa fase è considerata anche l'azione degli stessi destinatari che, anche inconsapevolmente, rilanciano i contenuti fake favorendone la diffusione.

La creazione e la diffusione di messaggi disinformativi possono derivare da motivazioni che hanno radici di natura economica come la massimizzazione dei profitti attraverso la raccolta pubblicitaria, politica, come ad esempio screditare una parte politica e favorirne un'altra, ma anche psicologiche e ludico-satiriche.

Dal punto di vista oggettivo, un messaggio informativo si contraddistingue per il formato, che può essere testo, audio, video ecc., per la durata e per il grado di falsità, scorrettezza e manipolazione. A questo proposito, come affermato dall'Autorità per le Garanzie nelle Comunicazioni, gli strumenti tecnologici utilizzati per la creazione di contenuti fake giocano un ruolo fondamentale. Durante la fase di ideazione dei messaggi di disinformazione, sono particolarmente adoperati i sistemi attuali di tracciamento delle attività di navigazione degli utenti in rete, utili per conoscere le preferenze, i gusti ma anche e soprattutto gli aspetti psicologici degli individui. Ciò comporta la possibilità, per gli ideatori e gli esecutori della filiera dei contenuti fake, di progettare messaggi ad hoc che stimolino gli interessi e le emozioni dei destinatari della disinformazione, in modo da confermare quanto citato da Hannah Arendt nel "La menzogna in politica. Riflessioni sui Pentagon Papers" (5): "il bugiardo ha il grande vantaggio di sapere in anticipo cosa l'ascoltatore desidera o si aspetta di sentire".

Durante la fase di produzione del contenuto sono molto utilizzati i software che permettono la manipolazione di contenuti e la generazione automatica degli stessi. Per quanto riguarda la fase di distribuzione e diffusione dei contenuti fake, gli strumenti tecnologici sempre più all'avanguardia sono i sistemi di posting sui social network, i software per la creazione e gestione dei bot, i server che permettono la gestione contemporanea di una molteplicità di device. In particolare, è noto che, nel processo di diffusione dei contenuti di disinformazione online, giocano un ruolo chiave gli algoritmi delle piattaforme online che definiscono il ranking dei contenuti mostrati nei risultati di ricerca (tra i più famosi e discussi, l'algoritmo di Google, PageRank (6)), e quelli utilizzati nelle piattaforme social che consentono la personalizzazione automatica dei contenuti visualizzati dagli utenti, un fenomeno detto "profilazione", che sarà trattato nei paragrafi successivi.

Per concludere la trattazione della filiera dei contenuti fake è necessario citare anche l'aspetto economico che la caratterizza. Infatti, all'interno del Rapporto tecnico riguardante le strategie di disinformazione (4), si afferma, nonostante non ci siano ad oggi dati quantitativi affidabili, che i costi di produzione di contenuti fake siano piuttosto bassi e che quelli di distribuzione tendono a zero. Ciò comporta un incentivo non indifferente

per gli ideatori di messaggi di disinformazione che si ritrovano protagonisti di un mercato con bassissime barriere all'entrata.

1.2.3. Le tematiche di disinformazione

Le tematiche maggiormente colpite dal fenomeno di disinformazione in Italia riguardano principalmente le "hard news" e le notizie di carattere scientifico e tecnologico. Per queste ultime viene prodotta una gran quantità di disinformazione sia perché idonee a produrre effetti sulla sfera ideologica e psicologica dei cittadini sia perché il livello quantitativo e qualitativo della produzione di tale categoria è tale per cui può facilmente essere screditato da informazioni non veritiere. Nella figura 3 vengono rappresentate le percentuali di distribuzione dell'offerta di contenuti "fake" in Italia per categoria. (3)

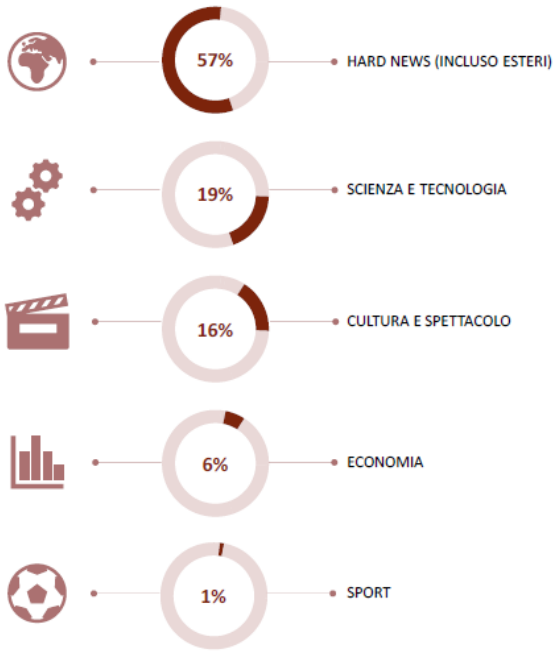

Figura 3 - Distribuzione dell'offerta di contenuti fake per categoria
Fonte: elaborazioni Agcom su dati Volocom (3)

1.3. Il ruolo dei social network nella diffusione dei contenuti fake

1.3.1. Il ruolo dei social network

Come più volte affermato, l'avvento di Internet e delle piattaforme online ha cambiato radicalmente il consumo dell'informazione. Il modello di integrazione verticale del processo produttivo dell'informazione in cui l'editore esercitava il controllo

dell'informazione prodotta e le notizie venivano fornite da un insieme di fonti ufficiali si è fortemente ridimensionato per dare spazio all'intermediazione delle piattaforme algoritmiche dove flussi di notizie ufficiali e non si mescolano nell'homepage di ciascun utente. È inoltre attivo uno scenario in cui le persone partecipano attivamente sia alla diffusione che alla creazione dei contenuti. È proprio in un contesto così confusionario e in cui il sistema informativo appare avere numerose falle, che entrano in gioco la sfiducia e la diffidenza nel mezzo di comunicazione e anche le notizie prodotte da fonti ufficiali si scontrano con una resistenza sempre più crescente. Gli individui, così, sono sempre più propensi ad informarsi affidandosi alla propria rete di contatti, ad attribuire credibilità ai contenuti e alle fonti che confermano i propri pregiudizi: tutti fenomeni che prendono piede e si rafforzano sui social network.

1.3.2. La profilazione algoritmica

Come citato nel sottoparagrafo 1.2.2., nella fase della diffusione dei contenuti fake giocano un ruolo chiave gli algoritmi nelle piattaforme online e social che danno vita al concetto di profilazione degli utenti. Il termine "profilazione", infatti, va ad indicare "l'insieme di attività di raccolta ed elaborazione dei dati inerenti agli utenti" (7), al fine di suddividerli in gruppi omogenei in base a gusti, comportamenti e interessi. Le piattaforme social come Facebook o Instagram, infatti, raccogliendo i nostri dati e i nostri click insieme a quelli dei nostri amici, sono in grado di suggerirci ciò che è probabile possa interessarci e ciò che possa catturare la nostra attenzione. I social network utilizzano tale meccanismo per selezionare e ordinare quanto viene mostrato nell'homepage, ad esempio il newsfeed di Facebook è un algoritmo di raccomandazione delle notizie. La conseguenza di tale fenomeno è che, ciascuno può costruirsi online il proprio "**daily me**" (2) quotidiano, una sorta di palinsesto creato sulla base dei nostri interessi. Ciò da un lato può essere considerato un passo in avanti nella consultazione di più fonti informative, dall'altro può tradursi in un meccanismo di esposizione selettiva alle informazioni, che accresce il rischio di disinformazione.

A questo proposito è interessante lo studio citato nel rapporto dell'Agcom in cui è stato valutato con quante fonti informative diverse interagisce generalmente l'utente su una piattaforma social. A seguito di una osservazione di 376 milioni di utenti Facebook in un arco temporale di 6 anni è stato notato che il numero di pagine di fonti informative con cui un individuo interagisce in un anno varia al variare di due fattori: la durata dell'intervallo temporale compreso tra la prima e l'ultima interazione dell'utente (per

esempio con un post delle fonti informative) e il livello di coinvolgimento dell'utente (in termini di quantità di like espressi per un post) (3). Come mostrato nella figura 4, un utente tende ad interagire con un insieme di fonti informative limitato e, all'aumentare della durata del periodo di interazione e del livello di coinvolgimento, il numero di fonti informative con cui l'utente interagisce si riduce. Addirittura, tra i risultati di questo studio si è riscontrato che gli utenti più attivi interagiscono solo con dieci pagine informative in un anno. Questo è un chiaro segno di esposizione selettiva all'informazione che appare ancor più accentuato con l'aumentare dell'attività social degli utenti.

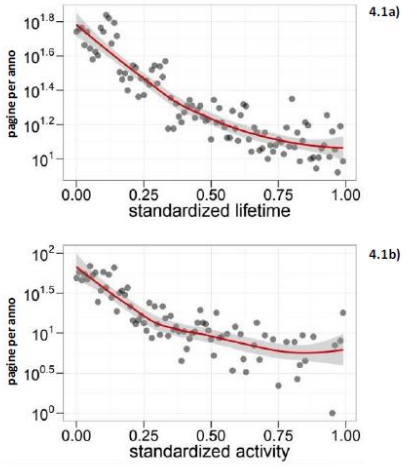

Figura 4 - Eterogeneità del consumo informativo,
per durata di interazione (4.1a) e livello di coinvolgimento degli utenti (4.1b) (3)

1.3.3. Confirmation bias e polarizzazione

Si è riscontrato, quindi, che l'informazione sulle piattaforme social, affinché sia completa e esaustiva, dipende dalla qualità dell'attività di ricerca dei singoli utenti. Le informazioni che vengono "selezionate" dagli algoritmi di profilazione spesso non sono sufficienti per creare un "daily me" con una qualità elevata. Succede spesso che le notizie offerte sono addirittura troppe e ciò richiede una fase di filtraggio da parte dell'utente, cioè tempo in più richiesto per la ricerca delle informazioni. Spesso quindi è l'utente stesso a decidere quando fermarsi nella ricerca, ed è qui che nasce il fenomeno denominato "confirmation bias" nella fruizione dell'informazione. Tale concetto, in psicologia, indica un fenomeno cognitivo umano per il quale gli individui tendono a muoversi entro un ambito delimitato dalle loro convinzioni acquisite. Ed è proprio ciò che accade all'utente nella sua attività

di ricerca e di filtraggio dell'informazione che gli viene proposta nelle piattaforme social nel momento in cui decide di fermarsi attratto dalle notizie che meglio soddisfano le sue convinzioni di partenza, cioè quando l'informazione proposta conferma la sua visione del mondo. Tale atteggiamento non è assolutamente nuovo: già nel 1922 Walter Lippman nel suo saggio magistrale "L'opinione pubblica" (8) affermava che "il modo in cui vediamo le cose è una combinazione di quello che c'è e di quello che ci aspettavamo di trovare".
Tale fenomeno è rafforzato dal cosiddetto "paradigma della post-verità" per cui "i fatti obiettivi sono meno rilevanti nel formare l'opinione pubblica rispetto al richiamo a emozioni e convinzioni personali". Ciascun utente, cioè, è attratto dalle notizie che confermano il proprio modo di vedere il mondo e anche da ciò che maggiormente stimola le proprie emozioni.
La conseguenza più pericolosa di questi fenomeni sta nel fatto che profili che hanno preferenze e atteggiamenti simili sulla piattaforma social tendono a concentrarsi su narrazioni specifiche e a riunirsi in determinati gruppi per rafforzare la propria visione del mondo, è ciò che viene comunemente identificato dalla parola "polarizzazione". Nella figura 5 vi è una chiara rappresentazione della struttura delle comunità di pagine per tipologia di azione informativa dove i nodi (lungo il cerchio) rappresentano le pagine e due pagine sono legate se un utente esprime apprezzamento su almeno un post di entrambe. La dimensione dell'arco è quindi determinata dal numero di utenti che le due pagine hanno in comune e i colori dell'arco identificano l'appartenenza di una pagina a una specifica comunità. Da questa analisi condotta tramite l'algoritmo Fast Greedy (algoritmo di rilevazione delle comunità), si è riscontrato che la maggior parte degli utenti rimane confinata all'interno di specifiche comunità. (3)

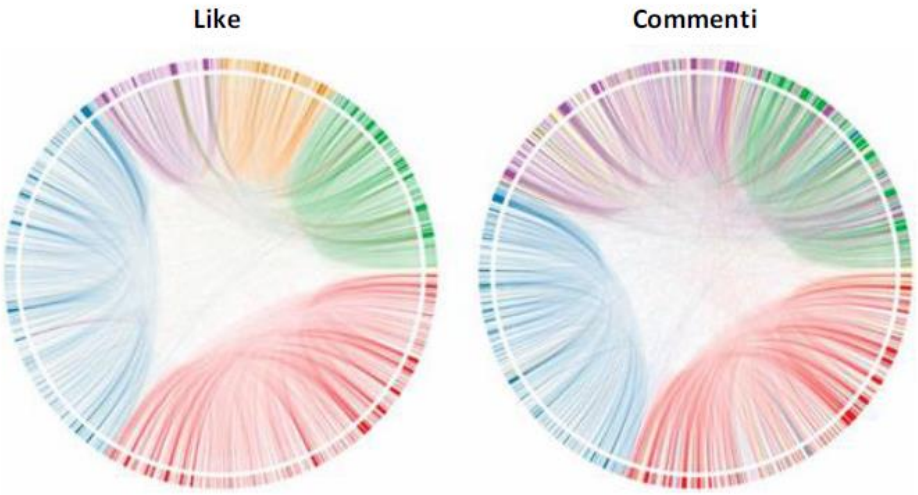

Figura 5 - Struttura delle comunità di pagine, per tipologia di azione informativa (3)

Il fenomeno della polarizzazione, dunque, porta gli utenti a chiudersi nella propria comunità, non dando modo al flusso di notizie di confutare le opinioni di ciascuno. Sembra essere una vera e propria contraddizione con il pluralismo informativo, insito nel sistema democratico dato dall'esistenza di numerose fonti confrontabili, che avrebbe dovuto rafforzarsi con l'avvento di Internet. Secondo il ricercatore del Laboratory of Computational Social Science, Networks Department, IMT Alti Studi Lucca, Walter Quattrociocchi, "la polarizzazione è la principale causa di disinformazione" (9).

1.4. Covid 19 e "infodemia"

Dopo aver introdotto l'attuale contesto informativo e disinformativo attivo in Italia, è ora necessario riassumere i fatti che hanno caratterizzato il periodo della prima ondata della pandemia da Coronavirus e che hanno generato un incremento della diffusione di disinformazione, che è oggetto dell'analisi di tale lavoro di tesi.

1.4.1. La pandemia Covid-19 in Italia

Il 31 dicembre 2019 le autorità sanitarie cinesi informano l'Organizzazione Mondiale della Sanità (Oms) un focolaio di una anomala polmonite a Wuhan, una città della Cina centrale. In pochi giorni si contano 41 casi, il resto del mondo però considera la notizia di poca importanza. Nel mese di gennaio le autorità cinesi identificano il nuovo virus, chiamato 2019-nCoV, della stessa famiglia dei coronavirus come quelli responsabili della

SARS e viene confermata la trasmissione del nuovo virus da uomo a uomo. Il 23 gennaio a Wuhan scatta il lockdown comprendente l'obbligo di non uscire di casa e di indossare la mascherina come protezione, vengono anche cancellati tutti i festeggiamenti del Capodanno cinese. Mentre il 30 gennaio l'Oms dichiara l'epidemia un'emergenza sanitaria pubblica internazionale, In Italia, il presidente del Consiglio, Giuseppe Conte, conferma i primi due casi di contagio riscontrati in Italia: due turisti cinesi in isolamento a Roma. Il 21 febbraio il nuovo coronavirus, la cui malattia viene denominata Covid-19, arriva ufficialmente anche in Italia: viene identificato il paziente 1 a Codogno. Da lì a pochi giorni sarebbero state registrati decine di nuovi casi, insieme alla prima vittima italiana. Una volta scoperto il primo focolaio interno, vengono adottate le prime misure di contenimento del contagio, la chiusura totale di 11 comuni dell'Italia settentrionale e la sospensione di manifestazioni ed eventi negli stessi comuni.

Nonostante l'invito proveniente dai media a non creare panico e a non fermarsi (esempio l'hashtag rilanciato, tra gli altri, dal sindaco di Milano, Giuseppe Sala, #milanononsiferma) e le numerose affermazioni provenienti da numerosi esponenti che ritenevano la nuova malattia Covid-19 simile ad un'influenza, la situazione precipita: aumentano rapidamente i casi positivi e i decessi.

Nel mese di marzo vengono sospese tutte le attività scolastiche in tutto il territorio italiano e la Lombardia diventa zona rossa: è il preludio di una chiusura che viene presto estesa nell'intero Paese dal 9 marzo, con il "**decreto #iorestoacasa**". Ha inizio la cosiddetta "fase 1" del periodo di contrasto alla diffusione del contagio da Coronavirus. L'11 Marzo 2020 l'Oms, considerato il numero di decessi e di paesi colpiti dalla diffusione del virus, dichiara lo stato di pandemia. Mentre il governo italiano introduce il "**decreto Cura Italia**" come prima misura di sostegno economico in seguito all'emergenza sanitaria, l'immagine di colonne di mezzi militari che trasportano bare di decine di vittime del Covid-19 verso i cimiteri di altre città diventa virale. Le misure del Governo sono sempre più stringenti: vengono chiusi i parchi e vietato lo sport se non nei pressi della propria abitazione, pochi giorni dopo vengono sospese gran parte delle attività produttive e viene vietato ai cittadini di spostarsi al di fuori del proprio comune di residenza.

Il 4 maggio, dopo la registrazione di un periodo di diminuzione del numero dei positivi, il Presidente del Consiglio Giuseppe Conte annuncia l'inizio della "fase 2" caratterizzata da un graduale allentamento delle misure di contenimento del contagio, che dà il via al ritorno al lavoro a 4 milioni di italiani e la possibilità di incontrare i "congiunti"; resta

l'obbligo di indossare la mascherina e di mantenere la distanza interpersonale di 1 metro (10).

1.4.2. Informazione e disinformazione durante il Coronavirus in Italia

Come è facilmente immaginabile, il diffondersi dell'epidemia nel territorio italiano è stato accompagnato da una parallela crescita dello spazio dedicato dai media all'argomento "Coronavirus". Secondo l'Autorità per la Garanzia nelle Comunicazioni, la quale ha dedicato all'analisi della produzione di informazione e disinformazione sul tema del Covid-19 tre "Osservatori sulla disinformazione online – Speciale Coronavirus" (11), se dal 1° al 29 febbraio il tempo dedicato al tema Coronavirus corrispondeva al 27.6% dal totale spazio dedicato all'informazione nei principali canali televisivi, solo dal 1° al 10 marzo tale valore è raddoppiato fino a raggiungere il 63,4%. È il momento in cui c'è stato il picco della curva di incidenza giornaliera delle notizie riguardanti il Coronavirus. Allo stesso tempo, il verificarsi della crisi sanitaria ed economica dovuta alla pandemia ha visto anche un incremento sostanziale della circolazione di fake news e di campagne di disinformazione sul Covid-19.

Proprio a causa dell'entità della diffusione di tale fenomeno, è stato coniato dall'Organizzazione mondiale della sanità, il termine "**infodemia**", composto dalle parole "informazione" ed "epidemia", per indicare la circolazione di una quantità eccessiva di informazioni, talvolta non vagliate con accuratezza, che rendono difficile orientarsi su un determinato argomento per la difficoltà di individuare fonti affidabili.

Nella figura 6 si può visualizzare l'incidenza giornaliera delle notizie riguardanti il Coronavirus sul totale della disinformazione online a confronto con l'informazione nei periodi corrispondenti alla prima ondata della pandemia.

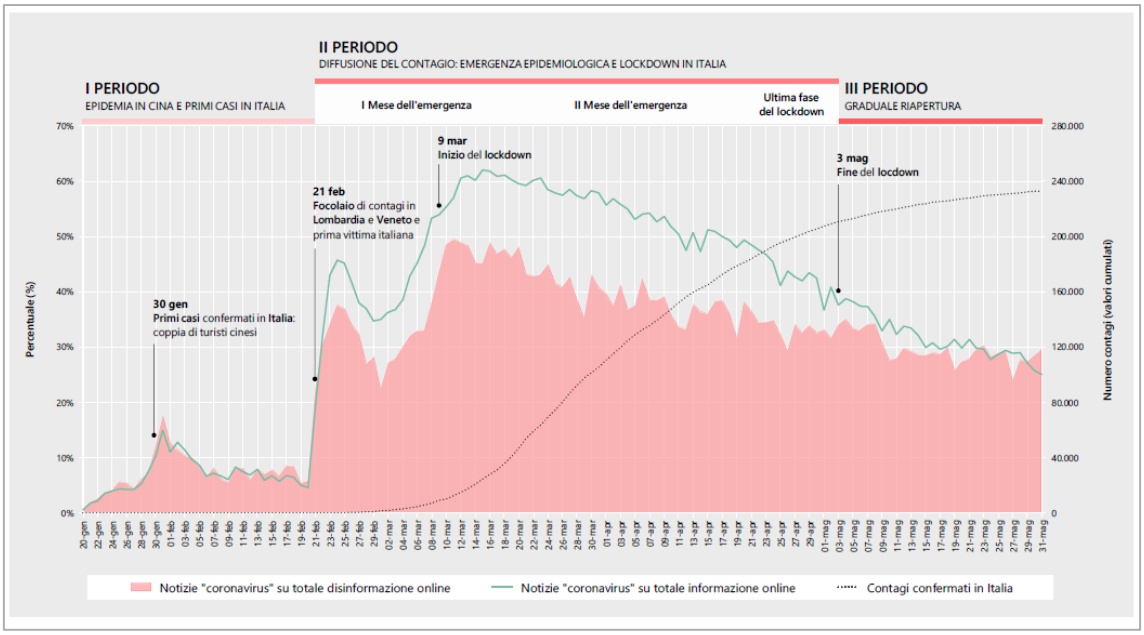

Figura 6 - Incidenza giornaliera delle notizie riguardanti il coronavirus sul totale disinformazione: confronto con l'informazione online [9]

Si può notare come nel I periodo dell'epidemia lo spazio dedicato al coronavirus è stato mediamente maggiore per le fonti di disinformazione rispetto a quelle informative. Nel II periodo si è invertita tale tendenza riscontrando una maggiore crescita per l'informazione. Dopo il I mese dell'emergenza, che è coincisa con il picco della produzione di informazione e disinformazione, si assiste ad un trend di riduzione per entrambi gli spazi, anche se con un tasso di decremento meno accentuato nel caso della disinformazione.

L'Agcom, che, nel novembre 2020, ha dato il via a un'indagine sul mondo dell'informazione, ha stimato che il 73% della popolazione giornalistica ha affermato di essersi imbattuta in casi di disinformazione e il 23% di questi afferma di aver riscontrato un caso di disinformazione al giorno nei mesi della pandemia. (12)

Da uno studio dell'Ong Avaaz "How Facebook can flatten the curve of the Coronavirus Infodemica" nell'aprile 2020 (13), si evince che addirittura le fake news sul Coronavirus sono state più viste di contenuti verificati e scientificamente accurati riguardanti l'emergenza sanitaria in corso. Infatti, nonostante alcuni tentativi di frenare tale fenomeno da parte dell'azienda di Mark Zuckemberg, nell'aprile del 2020, mentre la maggior parte dei Paesi si trovava nel picco dell'emergenza sanitaria, la disinformazione a tema salute raggiungeva i suoi livelli massimi.

Dalla ricerca "La pandemia immateriale. Gli effetti del Covid-19 tra social asintomatici e comunicazione istituzionale", condotta dal 1° febbraio al 10 aprile 2020 da parte di Luigi Giungato, ricercatore della Società Italiana di Intelligence (SOCINT) (14), il

fenomeno più diffuso con il corso della pandemia è stato la psicosi subìta da parte della popolazione italiana. Secondo tale studio che è stato focalizzato sui trend di interesse del pubblico rispetto alla tematica del "Coronavirus", è stato riscontrato che l'Italia ha avuto un ruolo trainante nella narrazione dell'emergenza globale. In merito alla situazione italiana, infatti, Giungato ha affermato che "La percezione nazionale è stata prevalentemente determinata dalle decisioni e dalle dichiarazioni delle istituzioni pubbliche, non sempre coerenti fra di loro, che hanno influito non solo sulla percezione del rischio, quanto sulla narrazione della paura. (…) Si è rilevato che il periodo di distanziamento sociale ha reso la popolazione più dipendente che mai da computer, smartphone e tv, e la percezione della realtà si è basata esclusivamente sui mezzi di informazione di massa e interpersonali, tra i quali emerge il ruolo di WhatsApp. (…)".
L'esempio più lampante riferito a tale fenomeno è stato raccontato dal servizio "Il virus nero" di Report, andato in onda nella serata del 27 aprile 2020 (15). Il video del Tgr Leonardo del 16 novembre 2015 in cui si parlava di un esperimento condotto da ricercatori cinesi in un laboratorio di Wuhan su un "Coronavirus", era rimasto per 5 anni privo di visualizzazioni nell'archivio della Rai. Una persona intervistata durante la trasmissione televisiva lo ha ritrovato nel marzo 2020, il mese in cui il Coronavirus è entrato a far parte delle vite degli italiani, e lo ha condiviso su WhatsApp alla sua cerchia ristretta di persone. In pochi giorni, il video è diventato virale su tutti i social di tutto il mondo, presentato come la dimostrazione che il Covid-19 sia stato costruito in laboratorio. È l'esempio di una "falsa contestualizzazione", cioè di una notizia vera di cui è stato cambiato il contesto, distorcendone così la percezione dei destinatari. È il motivo per cui, nonostante il tentativo di numerosi siti ufficiali di evidenziare la manipolazione di senso del video, sui Social tale video è arrivato ad avere milioni di visualizzazioni, rafforzando le teorie complottistiche sulla generazione del virus.

1.4.3. Le principali fake news sul Coronavirus

Le fake news maggiormente diffuse correlate all'argomento "Coronavirus" hanno visto come protagoniste false cure, trattamenti non comprovati, teorie del complotto sulle origini del virus e hanno preso di mira personalità come Bill Gates. L'analisi di monitoraggio e segnalazione dei siti che hanno pubblicato un maggior numero di contenuti fake sul Coronavirus condotta dall'Autorità per le Garanzie nelle Comunicazioni, ha portato alla elaborazione delle seguenti 10 bufale più diffuse sul Covid (11):

1. “Il virus del Covid-19 è stato sottratto da un laboratorio canadese da spie cinesi”;
2. “Il virus del Covid-19 contiene ‘sequenze simili all’Hiv’, lasciando intendere che si tratti di un virus costruito artificialmente”;
3. “La pandemia di Covid-19 era stata prevista in una simulazione”;
4. “Un gruppo finanziato da Bill Gates ha brevettato il virus del Covid-19”;
5. “Il virus del Covid-19 è un’arma biologica creata dall’uomo”;
6. “La tecnologia dei telefoni cellulari 5G è collegata alla pandemia di coronavirus”;
7. “L’argento colloidale può curare il Covid-19”;
8. “La Miracle Mineral Solution può curare il Covid-19”;
9. “L’aglio può curare il Covid-19”;
10. “È stato dimostrato che dosi massicce di vitamina C siano un trattamento efficace per il Covid-19.

È necessario sottolineare che tali fake news hanno avuto numerose conseguenze sugli effetti comportamentali dei destinatari, specie quelle riguardanti i rimedi preventivi e terapeutici contro il Covid. Da uno studio internazionale coordinato da esperti presso la *University of New South Wales* (16) in Australia e pubblicato sull'American Journal of Tropical Medicine and Hygiene è stato riscontrato che circa 800 decessi sono stati collegati a disinformazione, nonché 5.876 ricoveri e infortuni gravi. Si pensi che 60 persone hanno perso la vista dopo aver bevuto metanolo come cura per il coronavirus. Ancora, il mito secondo cui ad alte concentrazioni l'alcol uccide il virus che hanno circolato quasi ovunque possono portare a comportamenti a rischio. In India almeno 12 persone (tra cui 5 bambini) si sono ammalate gravemente dopo aver consumato un liquore a base di semi di datura, una pianta molto pericolosa, dopo aver visto un video online che lo suggeriva come rimedio contro il coronavirus.

1.4.4. L’infodemia e la politica

Dopo aver analizzato i fatti accaduti durante la prima ondata di pandemia e di “infodemia” del 2020, è arrivato il momento di restringere il campo della trattazione per dar spazio all’oggetto della ricerca di tale lavoro di tesi. In un contesto in cui la diffusione di informazione e la creazione di contenuti affidabili da parte delle personalità più seguite sui social sembra essere l’unica arma contro la disinformazione, le personalità politiche ricoprono un ruolo di estrema responsabilità. Secondo la teoria degli atti linguistici di John L. Austin (17), le parole, che sembrano entità volatili, sono in realtà meccanismi complessi e potenti che diventano veri e propri “atti” che influiscono sul mondo

circostante. L'uso delle parole, ovunque esse siano condivise, implica responsabilità e crea la necessità di fronteggiare le conseguenze degli atti che possono derivarne.

Non è sicuramente da sottovalutare il fatto che la propaganda politica sia una pratica da sempre caratterizzata da inganni e distorsioni informative costruite per screditare le fazioni e le tesi opposte. Come cita (18), "l'incrocio tra l'arena della politica e la tecnologia, però, ha portato a una crescita esponenziale nella creazione di "fake" e nella loro circolazione." I Social Network si ritrovano ad essere un'arma a doppio taglio: le personalità maggiormente seguite, i cosiddetti "influencer", hanno la possibilità di raggiungere un numero molto elevato di destinatari per condividere le proprie idee tramite un solo post, ma ciò risulta essere molto pericoloso quando i contenuti provocano disinformazione.

In riferimento al caso in esame, la Bbc ha stilato una lista dei politici che hanno diffuso più fake news sull'argomento "Coronavirus" (19): per quanto riguarda l'Italia, secondo l'emittente britannica, il maggior diffusore di contenuti "fake" è il leader della Lega, Matteo Salvini. Le valutazioni dell'emittente televisiva britannica sono emerse durante la trasmissione *Reality Check*, specializzata nella verifica di notizie che circolano in rete. A finire sotto accusa sono stati i post condivisi dell'ex ministro dell'Interno sulla possibilità che il coronavirus fosse frutto di una manipolazione umana, e nato in un laboratorio di Wuhan.

Come citato in un articolo curato da Jaime D'Alessandro (La Repubblica, 19 maggio 2020) (20), l'azienda americana NewsGuard, ha rintracciato sedici account che funzionano da "super diffusori" di fake news sul coronavirus su Twitter, cinque dei quali sono italiani, tra cui compaiono Alessandro Meluzzi, ex parlamentare di Forza Italia, Byoblu, sito del blogger Claudio Messora, Patrizia Rametta, coordinatrice regionale della Lega in Sicilia, la quale ha diffuso messaggi che accusavano Bill Gates di aver brevettato il Coronavirus, Cesare Sacchetti e Elio Lannutti, senatore del Movimento 5 Stelle, che affermava che "la vitamina C per via endovenosa può aiutare a curare la polmonite e prevenire la replicazione virale".

Capitolo 2

Background e lavori correlati

Prima di passare alla trattazione della ricerca sperimentale oggetto di questo lavoro di tesi, si è reso opportuno indagare sullo stato dell'arte relativo alle metodologie utilizzate nell'analisi. Nell'ultimo paragrafo saranno inoltre presentati alcuni lavori che per contesto, strumenti e risultati risultano affini all'analisi che sarà presentata successivamente.

2.1. Analisi delle Fake News: background

Il fenomeno delle Fake News ha spinto il mondo della ricerca verso argomentazioni sempre più complesse riguardanti la tematica della loro creazione e diffusione. In particolare, grazie alle informazioni sull'analisi delle Fake News fornite dalle teorie del comportamento umano sviluppate in discipline come la psicologia, la filosofia, le scienze sociali e l'economia, sono state introdotte numerose opportunità per studi qualitativi e quantitativi sui dati. Tale contributo è stato inoltre fondamentale per la costruzione di modelli utili per il rilevamento e l'intervento nel contesto delle Fake News.
Secondo gli autori della raccolta di ricerche relative all'argomento (21), le prospettive da cui è possibile studiare il fenomeno delle notizie false sono quattro: la conoscenza, ponendo l'attenzione sul falso contenuto di una Fake News; lo stile, cioè come le Fake News vengono scritte; la propagazione, come esse vengono diffuse; e la credibilità degli utenti che si occupano della creazione e diffusione. Gli studi basati sulla conoscenza e sullo stile possono essere condotti nel momento in cui una Fake News viene creata, mentre gli studi sulla propagazione e la credibilità mirano a sfruttare le informazioni relative ai social network, dopo che tali notizie vengono pubblicate.

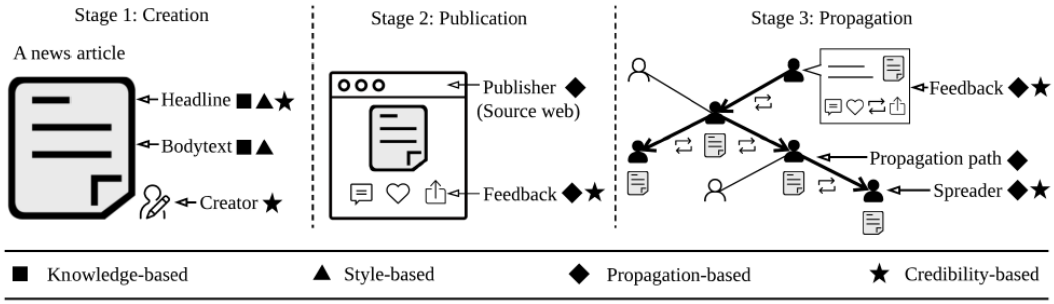

Figura 7 - Le quattro prospettive dei modelli di analisi delle Fake News (21)

Un esempio di studio basato sulla conoscenza è la rilevazione delle notizie false utilizzando un processo noto come "Fact-Checking", che mira a valutare l'autenticità delle notizie confrontandole con il contenuto di notizie verificate e i fatti noti. Il Fact-Checking può essere manuale, per esempio condotto da un gruppo di persone altamente credibili (i cosiddetti "Fact checker"), i quali verificano manualmente singole notizie e ne valutano la veridicità, oppure automatico. Le tecniche di verifica automatica sono utilizzate per raggiungere un numero molto elevato di notizie, specialmente quelle pubblicate sui social media, e si basano su tecniche di Information Retrieval (IR) e Natural Language Processing (NLP).

Come l'analisi knowledge-based, anche lo studio basato sullo stile si focalizza sul contenuto delle notizie. In particolare, tale tipologia mira a valutare se determinati testi siano intenzionati a fuorviare il pubblico o meno. Esempi di analisi che hanno tale obiettivo sono la "deception analysis" e la "detection", che si basano sulla teoria che lo stile di un contenuto ingannevole è differente da quello utilizzato per raccontare la verità, per esempio caratterizzato da espressioni esagerate e forti emozioni.

L'analisi sulla propagazione su incentra invece sulle informazioni legate a come una Fake News viene propagata e diffusa dagli utenti. Si basa sul concetto di "fake news cascade", che consiste in una struttura ad albero che rappresenta la propagazione di un determinato articolo su un social network. Il nodo principale della cascata rappresenta l'utente che per primo ha pubblicato la notizia falsa (il creatore o l'iniziatore), gli altri nodi rappresentano gli utenti che hanno successivamente pubblicato o inoltrato l'articolo.

Infine, gli studi delle Fake News basati sulla credibilità, studiano l'affidabilità delle fonti che pubblicano articoli informativi. Tale analisi implica il rilevamento di Fake News a partire da alcuni aspetti specifici come il titolo dell'articolo, la fonte, i commenti pubblicati e i diffusori. Ad esempio, la valutazione della credibilità dei titoli delle notizie

spesso si riduce alla rilevazione dei cd. “clickbait”, titoli il cui scopo principale è attrarre l’attenzione dei visitatori e incoraggiarli a fare click sul link ad una determinata pagina Web.

L’analisi della credibilità dei commenti, invece, esplora la posizione e l’opinione degli utenti nei confronti degli articoli di notizie. I modelli per valutare la credibilità dei commenti possono essere suddivisi in: modelli basati sul contenuto, i quali utilizzano funzionalità linguistiche estratte dai commenti e applicano strategie simili a quelle del rilevamento di Fake News basate sullo stile; modelli basati sul comportamento, che sfruttano le caratteristiche indicative dei commenti inaffidabili come la tempestività e l’estremismo; e i modelli basati sui grafici che tengono conto delle relazioni tra commenti, prodotti ecc.

2.2. L’analisi dei dati testuali

2.2.1. Caratteristiche dei Big data

Come già discusso nel precedente capitolo, l’avvento dei Social Network e la disponibilità di dispositivi tecnologici come lo smartphone e il tablet hanno modificato le caratteristiche dell’utente medio, che da fruitore passivo e saltuario di informazioni prevalentemente diffuse da fonti istituzionali, è divenuto protagonista attivo, sempre più coinvolto nella produzione di contenuti propri, nella modifica di contenuti altrui e anche in attività relative ad acquisti e vendite di oggetti e servizi. Ciò ha comportato, insieme alle altre conseguenze precedentemente descritte, la generazione di una incredibile mole di dati di natura estremamente eterogenea che insieme portano alla definizione dei cosiddetti “big data”. Tale espressione è impiegata infatti per descrivere quell’insieme di dati caratterizzati dalle cosiddette “5V”, che corrispondono all’espansione del modello delle “3V” introdotto da Douglas Laney (22): il “volume”, poiché appunto vi è un’enorme mole di dati generati al secondo; la “velocità”, termine che si riferisce alla generazione ma anche alla rapidità con cui i dati si “spostano” al giorno d’oggi; la “varietà”, in quanto i dati vengono prodotti in diversi formati tra cui quelli definiti “strutturati”, cioè organizzati per esempio in tabelle e quelli “non strutturati” come le immagini, i video e i testi. Le altre due “v” dei big data corrispondono alle parole “veracità” in quanto fondamentale, nell’utilizzo dei big data a scopi decisionali, è la loro affidabilità e accuratezza; e, ultimo ma non meno importante, il “valore”, termine che indica la capacità

dei dati di essere utili per il supporto di decisioni, tramite la trasformazione degli stessi in conoscenza.

Tali peculiarità dei big data hanno reso necessario lo sviluppo di nuove metodologie e architetture per gestirli, memorizzarli e processarli. Le sfide del volume, velocità e varietà dei big data, infatti, devono essere affrontate dai sistemi di storage e server progettati ad hoc per i big data, in quanto i software e le architetture informatiche tradizionali non sono in grado di gestirli, memorizzarli e processarli in un tempo ragionevole. (23) Si parla, appunto, di "big data storage" per indicare l'architettura di elaborazione e archiviazione che raccoglie e gestisce grandi set di dati e si riferisce a volumi che crescono in modo esponenziale su scala terabyte o petabyte. Un sistema di archiviazione di big data raggruppa un gran numero di server collegati a un disco ad alta capacità per supportare il software scritto per elaborare grandi quantità di dati. Tale sistema si basa sull'utilizzo del cosiddetto "Massively Parallel Processing", ovvero l'utilizzo di un gran numero di processori che eseguono simultaneamente una serie di calcoli coordinati in parallelo. Il framework di analisi più diffuso per i big data è Apache Hadoop (24) , esso consente l'elaborazione distribuita di grandi set di dati su cluster di computer utilizzando semplici modelli di programmazione. Tale software è progettato per lavorare da singoli server a migliaia di macchine, ognuna delle quali offre elaborazione e archiviazione locali. Per il calcolo distribuito viene comunemente utilizzato il framework Apache Spark (25), che richiede un gestore di cluster e un sistema di archiviazione distribuita. Per quest'ultimo si interfaccia con l'Hadoop Distributed File System (HDSF) o altre soluzioni.

2.2.2. Dati testuali e Natural Language Processing

Una delle varietà dei formati in cui i big data compiono è quello testuale, si tratta quindi di quei contenuti espressi in linguaggio naturale, cioè quello comunemente utilizzato dall'uomo per comunicare. In questo contesto l'uso dell'intelligenza artificiale assume rilevanza strategica, in quanto può favorire la realizzazione di soluzioni innovative per l'elaborazione, la comprensione e la produzione automatica di dati testuali. Per esempio, negli ultimi anni, si è assistito alla nascita di nuovi approcci, che integrano l'elaborazione del linguaggio naturale con gli algoritmi di apprendimento profondo, il cosiddetto "**deep learning**", producendo risultati straordinari in differenti scenari applicativi. Grazie ad essi, infatti, è oggi possibile tradurre testi tra lingue differenti in maniera automatica con

prestazioni sorprendenti, dialogare e fare domande alle macchine in linguaggio naturale su domini specifici, estrarre conoscenza e insight rilevanti da enormi quantità di dati testuali, generare contenuto in linguaggio naturale, ad esempio per sintetizzare le informazioni chiave di uno o più documenti, o determinare la polarità di testo che contiene opinioni, ad esempio su prodotti, servizi, individui, eventi (26).

È in questa tipologia di dati che entra in gioco il Natural Language Processing (NLP), un'area di ricerca e applicazione che esplora come i calcolatori elettronici possono essere utilizzati per elaborare e comprendere dei testi in linguaggio naturale. La storia di questo campo di ricerca viene fatta partire solitamente negli anni Cinquanta, quando Alan Turing pubblicò il suo articolo "Machine and Intelligence" (1950), in cui propose il suo famoso test per valutare l'abilità di un computer nel mostrare comportamenti intelligenti, indistinguibili da quelli di un essere umano, conversando in linguaggio naturale. La peculiarità di questo campo di ricerca è l'interdisciplinarità; è infatti noto che essa abbracci numerose discipline come l'informatica, l'intelligenza artificiale, la linguistica e psicologia. Anche i campi di applicazione che l'NLP include sono molteplici: tra questi vi è la traduzione automatica (ad es. il funzionamento di Google Translate), il riepilogo del testo, il riconoscimento vocale e gli agenti conversazionali intelligenti (tra i più famosi oggi ci sono Alexa di Amazon e Siri di Apple) e l'interfaccia utente.

In maggior dettaglio, l'NLP fornisce soluzioni per analizzare la struttura sintattica del testo, associando alle singole parole le rispettive categorie morfologiche (ad es. nome, verbo, aggettivo), identificando entità e classificandole in categorie predefinite (ad es. persona, data, luogo), estraendo dipendenze sintattiche (ad es. soggetti e complementi) e relazioni semantiche. Inoltre, l'NLP consente di comprendere la semantica del testo, identificando il significato delle parole, anche relazionato al contesto e alle modalità di utilizzo (ad es. ironia, sarcasmo, sentimento), classificandolo in categorie predefinite (ad es. sport, geografia, medicina) o sintetizzandone il contenuto.

2.2.3. Tecniche di preprocessing

Il linguaggio naturale è estremamente complesso e lo è ancor di più il processo di comprensione dello stesso. Per questo motivo è molto comune, in questo campo, utilizzare tecniche diverse per "preparare" i dati testuali al fine di ridurre la complessità della loro elaborazione. Si parla, infatti, di documenti di testo come dati di tipo non strutturato, cioè conservati senza alcuno schema, che è necessario trasformare in dati con

una struttura specifica. Verranno di seguito presentate alcune delle principali tecniche utilizzate in NLP con questo obiettivo (27).

2.2.3.1. Suddivisione del testo

Il primo step per la preparazione di un dato di tipo testuale è la sua suddivisione in parti più piccole e più semplici da processare. La tecnica più comunemente utilizzata per svolgere questa operazione è la cosiddetta "**tokenization**", il cui scopo è spezzare il testo contenuto in un determinato documento in "token" che, a seconda dell'applicazione, può indicare singole parole, frasi o sezioni del testo. Questa scelta viene effettuata sulla base dell'obiettivo, in quanto più è piccola la porzione di documento analizzata maggiore è l'impatto delle parole all'interno del documento.
Esistono diversi modi per eseguire la tokenizzazione, generalmente quella più utilizzata consiste nello spezzare il testo in base ai cosiddetti delimitatori come i terminatori di riga *(\n)*, di frase *(.)* o di periodo *(, ; :).* Per specifiche applicazioni può essere utile suddividere il testo in singole parole, considerando gli spazi come delimitatori tra due parole.

2.2.3.2. Rimozione delle stopword e *case normalization*

Nel linguaggio umano esistono dei termini che non aggiungono un contenuto semantico ad un documento testuale, ma servono come intercalare tra periodi e vengono pertanto definite in lingua inglese "**stopword**". Si considerano "stopwords" le preposizioni, gli articoli, gli avverbi, i pronomi, ma anche altre parole che non sono utili all'analisi in quanto si ripetono più volte. Non esiste pertanto un vocabolario di "stopword" universale, una lista di parole può essere costruita ad hoc sul testo che deve essere analizzato, considerando l'obiettivo specifico dell'analisi. L'operazione di rimozione delle stopwords dal testo è effettuata nella fase di preprocessing per diminuire la quantità di parole e di risorse computazionali richieste e per mettere in risalto le parole che danno una maggiore informazione sul contenuto del testo.
La "**case normalization"** è il processo di trasformazione del testo in un'unica forma canonica, come per esempio il carattere minuscolo o minuscolo. Ad esempio, la parola "UnIVerSità" viene trasformata in "università". Ciò garantisce una miglior performance delle operazioni da effettuare e per fare in modo che parole uguali vengano riconosciute come tali indipendentemente dai caratteri utilizzati per rappresentarle.

2.2.3.3. Stemming

Tale tecnica di processazione del testo riporta le parole alla loro forma radice. In questa fase, ad esempio, vengono eliminati i prefissi e i suffissi o le forme singolari, con l'obiettivo di trattare le parole allo stesso modo. Tale operazione impedisce che declinazioni diverse dello stesso verbo o sostantivo vengano interpretate come parole distinte dal modello, evitando quindi di perdere l'associazione semantica tra singolari-plurali, maschili-femminili e diversi tempi verbali. È un processo molto utile per correggere gli errori di ortografia dei token e per velocizzare le prestazioni in quanto, anche in questo caso, viene ridotta la dimensionalità di parole distinte presenti nel testo agevolando la computabilità del modello. Tale tecnica può presentare dei limiti come lo stravolgimento del significato di alcune parole ed il basso livello di interpretabilità di alcune parole. Esistono degli "**stemmer**", vocabolari di riferimento per le diverse lingue, che possono essere utilizzati per facilitare questa attività.

2.2.3.4. La rappresentazione delle parole

I metodi di rappresentazione di dati di tipo testuale sono fondamentali al fine di trasformare questa tipologia di dato in una forma strutturale. Uno dei modelli più comuni di rappresentazione di dati di tipo testuale in forma numerica è il modello definito "**Bag of Words**", tradotto in italiano come "borsa di parole". Si tratta di una rappresentazione che descrive le occorrenze delle parole all'interno di un documento. Il nome di tale tecnica deriva dal fatto che le parole contenute nei documenti vengono elencate ignorando l'ordine con cui esse si presentano. È comune la rappresentazione di tale elenco in una matrice in cui ad ogni riga corrisponde un documento e ad ogni colonna corrisponde il token, associando ad esso un valore specifico. (28)

2.2.3.5. Tecniche di pesatura

Durante la processazione di un corpus di documenti testuali, risulta fondamentale il processo di pesatura delle unità che compongono ciascun documento. È una tecnica che associa a ciascun token un peso che può aiutare la caratterizzazione del documento stesso, facendo emergere determinate parole rispetto ad altre secondo criteri diversi. Nella

letteratura scientifica esistono diverse tecniche di pesatura delle unità: tra le più comuni sono da citare l'associazione binaria, la frequenza semplice e la "TF-IDF". Nel primo caso, a ciascuna parola si assegna "0" in caso di assenza nel documento, e "1" in caso di presenza; la frequenza semplice consiste nel calcolo dell'occorrenza di un token all'interno del documento. Tuttavia, se l'obiettivo è enfatizzare le parole più significative contenute nel testo, lo schema più utilizzato è il "**Term frequency - inverse document frequency**" (**tf – idf**) (29). Questa funzione di peso misura l'importanza di un termine rispetto a una collezione di documenti combinando la "Term frequency" cioè la frequenza del token all'interno del documento calcolata come numero di occorrenze del termine *i* nel documento *j* diviso il numero di termini totali compresi nel documento *j*:

$$tf_{i,j} = \frac{n_{i,j}}{|d_j|}$$

con la "Inverse document frequency" che indica l'importanza generale del termine *i* nella collezione:

$$idf_i = \log \frac{|D|}{|\{d : i \in d\}|}$$

dove $|D|$ è il numero dei documenti nella collezione, mentre il denominatore è il numero di documenti che contengono il termine *i*.

Le due frequenze vengono moltiplicate:

$$(tf - idf)_{i,j} = tf_{i,j} \times idf_i$$

e si ottiene un valore che aumenta proporzionalmente al numero di volte in cui il termine è contenuto nel documento, ma cresce in maniera inversamente proporzionale con la frequenza del termine nella collezione. L'intento di tale funzione di peso è quella di assegnare un valore maggiore di importanza a parole che compaiono nel documento, ma che in generale sono poco frequenti e possono essere maggiormente rilevanti all'estrazione di conoscenza da un dato di tipo testuale.

2.2.4. Topic modelling

Uno dei modelli di estrazione di conoscenza da documenti testuali compresi nel campo dell'elaborazione del linguaggio naturale è il "Topic Model" (30). Tramite questo modello è possibile estrarre, partendo da un set di documenti, gli argomenti che vengono trattati all'interno dei testi, determinando le parole che meglio li descrivono. Si tratta di una tecnica di apprendimento automatico nota come "non supervisionata" in quanto, a

differenza delle tecniche "supervisionate", non richiede un elenco predefinito di tag o dati di addestramento precedentemente classificati dagli esseri umani, cioè non si devono necessariamente conoscere gli argomenti di un insieme di testi prima di analizzarli.
Il "Topic modelling" è uno strumento di estrazione del testo di uso frequente per la scoperta di strutture semantiche nascoste in un corpo di un testo. Si basa sul fatto che, intuitivamente, dato che un documento riguarda un argomento particolare, ci si aspetta che nel documento compaiano parole specifiche più o meno frequentemente. Un tema, un argomento, un "topic" è infatti un insieme di parole che spesso vengono menzionate insieme. Ad esempio, supponiamo che in un testo siano osservate parole come "artista", "canzone" e "concerto". La probabilità di osservare nello stesso testo la parola "chitarra" è molto più alta di quella associata ad una qualsiasi altra parola, come per esempio "pianta". La metodologia del Topic Modelling ha l'obiettivo di formare dei gruppi di parole, ognuno dei quali rappresenta una tematica specifica e di raggruppare un set di documenti sulla base dei topic di cui essi parlano.
Nella letteratura scientifica sono stati proposti diversi modelli utili alla modellazione dei topic: due dei più comuni risultano essere il Latent Semantic Allocation (LSA) e il Latent Dirichlet Allocation (LDA), che saranno di seguito presentati.

2.2.4.1. Latent Semantic Allocation (LSA)

Una delle tecniche utilizzate nell'ambito della modellazione dei topic e del Natural Language Processing è la Latent Semantic Allocation (LSA). Si tratta di una tecnica di analisi semantica che consente di approfondire la conoscenza del contenuto di un documento, oltre ad individuare la relazione tra i termini che lo compongono. Tramite questa tecnica, da singoli documenti vengono estrapolati i concetti rilevanti di cui trattano. Tale metodologia parte da un'ipotesi distributiva secondo cui parole che hanno un significato simile si trovino in parti di testo simili. Viene quindi costruita per ciascun documento una matrice contenente i conteggi delle parole per documento, dove le righe rappresentano le parole uniche e le colonne rappresentano ogni documento e viene utilizzata la tecnica *Singular Value Decomposition* (*SVD*) per ridurre il numero di righe in modo da caratterizzare meglio i documenti contenuti nel corpus, preservando la struttura di similarità tra le colonne. Dopo aver strutturato il dato, i documenti vengono confrontati prendendo il coseno dell'angolo tra i due vettori formati da due colonne qualsiasi. Consegue che i valori vicini a 1 rappresentano documenti molto simili mentre valori vicini a 0 rappresentano documenti molto dissimili. (31)

2.2.4.2. Latent Dirichlet Allocation (LDA)

Uno dei più comuni algoritmi della modellazione dei topic, utilizzato in questo lavoro di ricerca, è il Latent Dirichlet Allocation (LDA), un modello probabilistico generativo che si basa sull'idea che i documenti testuali sono rappresentati come insiemi di parole che, combinate tra loro, formano uno o più sottoinsiemi di argomenti latenti, dove ogni argomento è caratterizzato da una distribuzione di parole. Tale modello è stato presentato per la prima volta in una pubblicazione del "Journal of Machine Learning Research" dal titolo "Latent Dirichlet Allocation" (32).

Di seguito vengono definiti i principali termini del modello:

- La *parola* è l'unità di base dei dati discreti, definita come un elemento di un vocabolario indicizzato da {1, …, V}. Le parole vengono rappresentate utilizzando vettori di base unitaria che hanno un singolo componente uguale a uno e tutti gli altri componenti uguali a zero. Usando gli apici per denotare i componenti, la v-esima parola del vocabolario è rappresentata dal V-esimo vettore w tale che $w^v = 1$ e $w^u = 0$ per $u \neq v$.
- Il *documento* è una sequenza di N parole denotata da $w = (w_1,w_2,...,w_N)$, dove w_n è la n-esima parola nella sequenza.
- Il *corpus* è una collezione di M documenti denotata da $D = \{w_1,w_2,...,w_M\}$.

Dati questi termini, il modello Latent Dirichlet Allocation (LDA) assume il seguente processo generativo per ciascun documento w nel corpus D:

1. Scegliere $N \sim Poisson(\xi)$.
2. Scegliere $\theta \sim Dir(\alpha)$.
3. Per ognuna delle N parole w_n:
   a) Scegliere un topic $z_n \sim Multinomial(\theta)$
   b) Scegliere una parola w_n da $p(w_n|z_n,\beta)$, una probabilità multinomiale condizionata al topic z_n.

Le ipotesi semplificative di questo modello sono le seguenti:

- La dimensionalità k della distribuzione di Dirichlet (cioè la dimensionalità della variabile topic z) si assume nota e fissa.
- Le probabilità della parola sono parametrizzate da una matrice $k \times V$ β dove $\beta_{ij} = p(w^j = 1 \,|\, z^i = 1)$ (la probabilità di un i-esimo topic di contenere la j-esima parola) è trattata come una quantità fissa che deve essere stimata.

- N è indipendente da tutti gli altri dati che generano variabili (q e z).

Una variabile θ k-dimensionale casuale di Dirichlet può assumere valori nel (k-1)-simplesso (un k-vettore θ sta nel (k-1)-simplesso se $\theta_i \geq 0$, $\sum_{i=1}^{k} \theta_i = 1$), e ha la seguente densità di probabilità su questo simplesso:

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma\left(\alpha_i\right)} \theta_1^{\alpha_i - 1} \dots \theta_k^{\alpha_k - 1}$$

dove il parametro α è un k-vettore con componenti $\alpha_i > 0$ e dove Γ(x) è la funzione Gamma. La Dirichlet è una distribuzione sul simplesso – è nella famiglia esponenziale, ha finite statistiche dimensionali sufficienti ed è coniugata alla distribuzione multinomiale.

Dati i parametri α e β, la distribuzione congiunta di una miscela di argomenti θ, un insieme di N argomenti z, e un insieme di N parole w è data da:

$$p(\theta, z, w \mid \alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\, \theta)\, p(w_n|\, z_n, \beta)$$

dove $p(z_n|\, \theta)$ è θ_i per l'unico i tale che $z_n^i = 1$. Integrando su θ e sommando su z, otteniamo la distribuzione marginale di un documento:

$$p(w|\alpha, \beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta)\right)\, d\theta.$$

Infine, considerando il prodotto delle probabilità marginali dei singoli documenti, otteniamo la probabilità di un corpus:

$$p(D|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta)\right)\, d\theta_d\,.$$

Una rappresentazione del modello LDA è possibile visualizzarla nella figura 8:

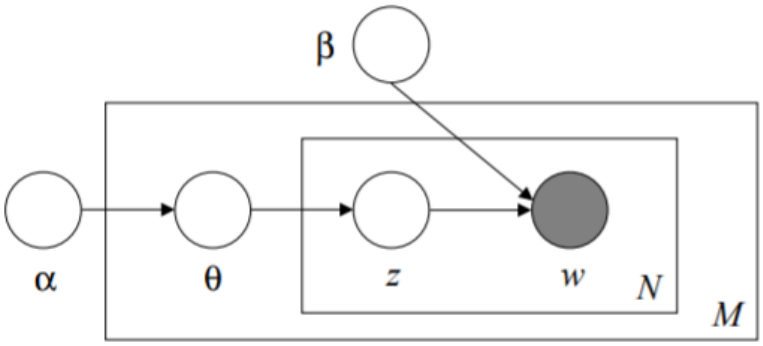

Figura 8 – Rappresentazione grafica del modello LDA (32)

La figura 8 mostra la cosiddetta "notazione su piastra", che è spesso utilizzata per rappresentare modelli grafici probabilistici in modo tale da cogliere in maniera chiara le dipendenze tra le variabili. Le piastre rappresentano entità ripetute: quella esterna rappresenta gli M documenti, mentre la piastra interna rappresenta le posizioni delle N parole ripetute in un dato documento; ogni posizione è associata a una scelta di argomento e parola. I parametri α e β sono parametri a livello di corpus, che si presume vengano campionati una volta nel processo di generazione di un corpus. Le variabili θ_d sono a livello di documento, campionate una volta per documento. Infine, le variabili z_{dn} e w_n sono variabili a livello di parola e vengono campionate una volta per ogni parola in ogni documento. Grazie a questa struttura a tre livelli, i topic possono essere associati a più documenti.

Nell'esempio in figura 9 sono stati estratti quattro topic, ciascuno rappresentato da quindici parole. Nel corpus raffigurato ciascuna parola è stata colorata con il colore corrispondente al topic a cui appartiene con probabilità maggiore.

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figura 9 - Esempio applicativo del modello LDA [25]

2.3. Metodologie di visualizzazione

2.3.2. LDAvis

Coerentemente con la scelta dell'applicazione del modello Latent Dirichlet Allocation, è stata scelta la metodologia di visualizzazione LDAvis, una visualizzazione interattiva di argomenti stimati usando l'LDA. Tale tecnica fornisce una visione globale degli argomenti, delle loro similarità e delle loro differenze, consentendo allo stesso tempo un'ispezione approfondita dei termini più strettamente associati a ogni singolo argomento (33) (34).

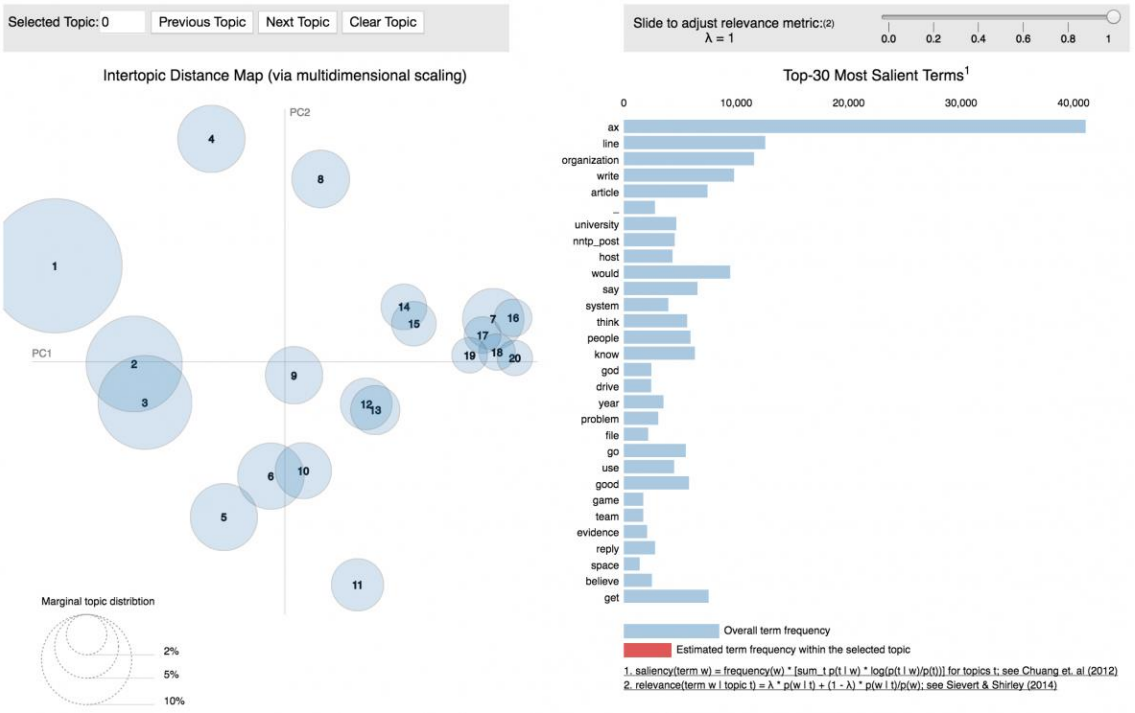

Figura 10 - Rappresentazione LDAvis (34)

Il pannello di sinistra visualizza gli argomenti come cerchi nel piano bidimensionale i cui centri sono determinati calcolando la divergenza Jensen – Shannon tra gli argomenti, quindi utilizzando la scala multidimensionale per proiettare le distanze inter-argomento su due dimensioni. La prevalenza complessiva di ogni argomento è identificata utilizzando le aree dei cerchi. Il pannello di destra rappresenta un grafico a barre orizzontali le cui barre rappresentano i singoli termini più utili per interpretare l'argomento selezionato a sinistra. In particolare, le barre celesti indicano la frequenza delle parole nell'intero corpus, le barre in rosso rappresentano la frequenza specifica dell'argomento del termine. Il cursore λ consente di classificare i termini in base alla rilevanza del termine. Per impostazione predefinita, i termini di un argomento sono classificati in ordine decrescente in base alla loro probabilità specifica dell'argomento (λ = 1).

2.3.1. t-Distributed Stochastic Neighbor Embedding

Un'altra metodologia di visualizzazione utilizzata nell'analisi è stata la t-Distributed Stochastic Neighbor Embedding (t-SNE) (35) (36), una tecnica non supervisionata e non lineare utilizzata principalmente per l'esplorazione dei dati e la visualizzazione di dati ad alta dimensione. Tale rappresentazione fornisce un'intuizione di come i dati sono

organizzati in uno spazio ad alta dimensione. L'algoritmo calcola una misura di somiglianza tra coppie di istanze nello spazio ad alta dimensione e nello spazio a bassa dimensione e cerca di ottimizzare queste due misure di somiglianza utilizzando una funzione di costo. Il funzionamento di tale algoritmo si può suddividere in tre fasi fondamentali:

1. Nella prima fase viene costruita una distribuzione di probabilità che ad ogni coppia di punti nello spazio originale ad alta dimensionalità associa un valore di probabilità elevato se i due punti sono simili, basso se sono dissimili. Per ogni punto in uno spazio dato, viene centrata una distribuzione gaussiana e calcolata la densità di tutti i punti sotto quella distribuzione. Si ottengono quindi una serie di probabilità proporzionali alle somiglianze.
2. Successivamente, l'algoritmo utilizza la distribuzione t di Student con un grado di libertà, nota anche come distribuzione di Cauchy. In questo passaggio si ottiene quindi una seconda serie di probabilità nello spazio dimensionale ridotto. Tale distribuzione ha una struttura diversa rispetto alla gaussiana, in quanto le code pesanti permettono una migliore modellazione delle dissimilarità tra oggetti distanti.
3. Nell'ultimo passaggio, l'algoritmo minimizza la divergenza di Kullback-Leibler delle due distribuzioni di probabilità tramite la discesa del gradiente, riorganizzando i punti nello spazio a dimensione ridotta.

La minimizzazione della divergenza di Kullback-Leibler consente di avere penalità elevate se punti vicini nello spazio originale vengono considerati lontani nello spazio a dimensionalità ridotta, mentre il viceversa ha un'influenza minore, tendendo quindi a preservare la struttura locale della distribuzione dei punti. In questo modo, t-SNE mappa i dati multidimensionali in uno spazio dimensionale inferiore che riflette la similarità tra i punti nello spazio ad alta dimensionalità. Tale modello viene utilizzato in molte applicazioni tra cui la valutazione della segmentazione.

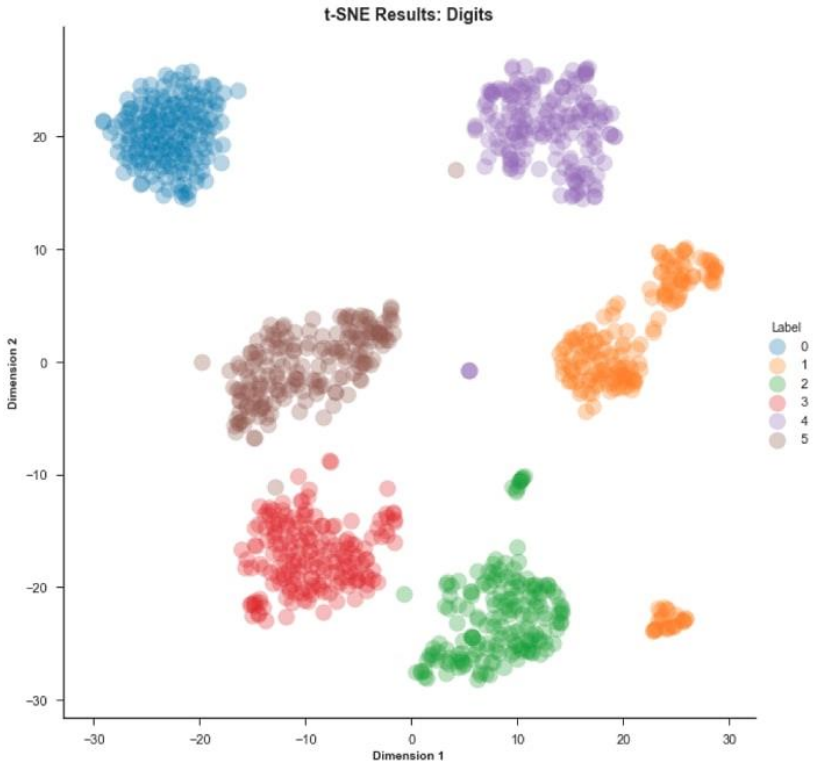

Figura 11 - Rappresentazione t-SNE (35)

2.4. Strumenti di analisi

2.4.1. Linguaggi di programmazione

2.4.1.1. Python

Python è un linguaggio di programmazione orientato a oggetti fortemente utilizzato nell'ambito dell'analisi di dati (37). In particolare, ciò che lo rende ampiamente utile nell'informatica scientifica, è il gran numero di package funzionali che accelerano e semplificano l'elaborazione dei dati, ottenendo un gran risparmio di tempo. Sono infatti disponibili una serie di librerie open-source adatte a scopi diversi che sono diventate molto popolari grazie alla loro leggibilità, flessibilità e scalabilità (38). Sono qui presentate le librerie utilizzate nel corso dell'analisi:

- **Pandas** è una libreria open-source scritta per la manipolazione e l'analisi dei dati. In particolare, offre strutture dati e operazioni per manipolare tabelle numeriche e serie temporali. (39) È stata scelta per le alte prestazioni e facilità di utilizzo nella gestione di un'ampia dimensionalità dei dataset e per la possibilità di eseguire operazioni di comparazione con il linguaggio SQL (40);
- **Matplotlib** è una libreria per la creazione di grafici scritta per il linguaggio di programmazione Python. Tale libreria è risultata di forte supporto all'analisi

condotta tramite Pandas, per poter visualizzare al meglio le statistiche ottenute (41);

- **Natural Language Toolkit** (**Nltk**) è una suite di librerie e programmi per l'analisi simbolica e statistica nel campo dell'elaborazione del linguaggio naturale. Essa fornisce librerie per classificazione, tokenizzazione, stemming, tagging, analisi e ragionamento semantico. È stata utilizzata nell'analisi soprattutto durante la fase di processazione del testo prima dell'applicazione del modello (42);
- **Gensim** è una libreria scritta per topic-modelling non supervisionati e algoritmi di Natural Language Processing. È progettato per gestire raccolte di testo di grandi dimensioni utilizzando lo streaming di dati e algoritmi online incrementali; uno tra questi è il Latent Dirichlet Allocation descritto nei precedenti paragrafi (43).

2.4.1.2. Pyspark

Pyspark è una collaborazione di Apache Spark, il framework di elaborazione cluster open-source introdotto nel 2.2.1., e Python. Esso permette di sfruttare allo stesso tempo la semplicità di Python e la potenza di Apache Spark per la gestione di Big Data (44). In particolare, è stato utilizzato **PySpark SQL**: un modulo di astrazione utilizzato principalmente per l'elaborazione di set di dati strutturati e semi-strutturati, che dà la possibilità di estrarre e di esplorare i dati usando query in linguaggio SQL (45). L'utilizzo di tale framework è stato possibile grazie all'utilizzo del cluster di calcolo Big Data di SmartData@PoliTO. (46)

2.4.1.3. Structured Query Language (SQL)

SQL è un linguaggio standardizzato per database basati sul modello relazionale (RDBMS), progettato per creare e modificare schemi di database, gestire dati memorizzati, interrogare i dati memorizzati e creare e gestire strumenti di controllo e accesso ai dati. Esso richiede la specifica di proprietà logiche delle informazioni ricercate; in particolare mette a disposizione diversi operatori di assegnazione, di confronto, stringa, aritmetici, condizionali, logici e tra bit. (47)

2.4.2. Tool di data mining: RapidMiner

Per il disegno di una pipeline di Machine Learning, è stato utilizzato il tool di data mining Rapid Miner. RapidMiner è una piattaforma di data science che fornisce un ambiente integrato per la preparazione dei dati, l'apprendimento automatico, il deep learning, il text mining e l'analisi predittiva. È uno strumento molto versatile che supporta tutte le fasi del processo di apprendimento automatico, inclusa la preparazione dei dati, la visualizzazione dei risultati e la convalida e l'ottimizzazione del modello. RapidMiner è scritto nel linguaggio di programmazione Java e fornisce una GUI per progettare ed eseguire flussi di lavoro analitici, chiamati "processi" che sono costituiti da più operatori. Ogni operatore esegue una singola attività all'interno del processo e l'output di ciascun operatore costituisce l'input di quello successivo. (48)

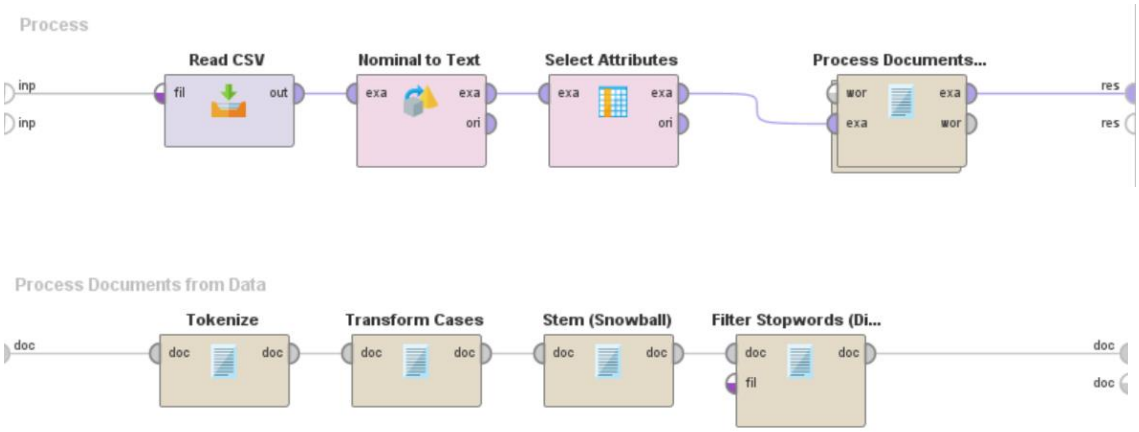

Figura 12 – Esempio di pipeline di ML in RapidMiner

2.4.3. Tool di visualizzazione: PowerBI

Per la visualizzazione dei dati della prima parte di analisi è stato utilizzato il tool di data visualization PowerBI. Tale piattaforma di analisi business consente la raccolta di tutti i dati a disposizione, la loro connessione e permette la creazione di grafici interattivi utili per poter comprendere al meglio le informazioni che i dati possono fornire. (49)

2.5. Lavori correlati

2.5.1. Social Network e influencer politici

Uno dei due dataset utilizzati nell'analisi oggetto di questo lavoro di tesi, in particolare i dati riferiti a Instagram, è stato collezionato e analizzato in un lavoro di ricerca pubblicato da alcuni dottorandi del Politecnico di Torino in collaborazione con l'Universidade Federal de Minas Gerais (50). La ricerca si è incentrata sui post e i commenti pubblicati da febbraio ad aprile del 2019 dai principali esponenti politici italiani, in concomitanza

con il periodo precedente alle elezioni europee del 2019. Sono state monitorate le attività di influencer italiani provenienti da settori come la musica, lo sport, lo spettacolo e la politica, osservando che i profili politici attraggono più interazioni delle altre categorie. Questa categoria è stata analizzata separatamente etichettando i politici con il partito politico di appartenenza, effettuando un confronto sulla base di metriche dei social network tra cui: commenti per 1000 follower, commenti per like, numero di commentatori, caratteri dei commenti, tempo di risposta al post. Sono stati inoltre analizzati nello specifico i commenti e la menzione degli utenti al loro interno, ottenendo come risultato che i commenti sotti i post politici attraggono molti commenti anche non sollecitati.

Il ruolo dei personaggi politici sui social network ha ottenuto anche l'attenzione degli autori di (51). La loro ricerca si è incentrata sull'uso di approcci alla scienza della comunicazione computazionale, che consente di tenere traccia delle conversazioni politiche sui social media. È stata effettuata un'analisi della polarizzazione politica su Facebook, Twitter e Whatsapp per 16 mesi, nello specifico riguardanti una controversia politica in Israele. Gli aspetti affrontati sono stati: la polarizzazione internazionale (omofilia ed eterofilia), la polarizzazione posizionale (posizioni espresse) e la polarizzazione affettiva (le emozioni e gli atteggiamenti espressi dagli utenti).

2.5.2. Contesto: fake news e Covid -19

Gli autori in (52) hanno studiato la proliferazione di notizie false sul Covid-19 in Nigeria. Il fenomeno è stato studiato utilizzando la teoria degli "Uses and Gratification", un approccio che si propone di comprendere la comunicazione di massa (53), ampliato da una motivazione di "altruismo". I dati, analizzati con i minimi quadrati parziali (PLS), hanno mostrato che il fattore più significativo che ha predetto la condivisione di notizie false di Covid-19 è stato il cosiddetto *altruismo*. Tra le altre cause di condivisione di false informazioni sul Covid-19 sono state la *condivisione di informazioni*, la *socializzazione*, la *ricerca di informazioni* e il *passatempo*.

Sempre a proposito della diffusione di informazioni sul Covid-19, gli autori in (54) hanno analizzato i dati di cinque piattaforme di social media diverse: Twitter, Instagram, Youtube, Reddit e Gab, cogliendo il coinvolgimento e l'interesse per l'argomento Covid-19 in ciascuna piattaforma. In particolare, in questo lavoro di ricerca è stato caratterizzato

per ciascuna piattaforma l'indice di riproduzione di base (R_o), cioè il numero medio di utenti che iniziano a postare sul Covid-19 che un individuo che già crea post sul Covid-19 crea. È stato cioè messo a paragone il significato di "infettività" epidemiologica con quella "infodemica". Sono state inoltre confrontate le fonti di informazione attendibili e quelle discutibili, non riscontrando particolari differenze nei modelli di diffusione. La misurazione delle interazioni come l'*engagement* che gli utenti delle piattaforme hanno nei confronti del topic Covid-19 è stato calcolato per ciascuna piattaforma rispetto ai parametri presenti: nello specifico, su Instagram, con il numero di likes e di commenti nel tempo.

Gli autori in (55) si sono focalizzati su una specifica Fake News diffusa nei mesi della pandemia: la teoria complottista che ha collegato la tecnologia del 5G alla diffusione del Covid-19. La ricerca si è proposta di comprendere i cosiddetti "driver" di tale teoria, analizzando il contenuto dei dati di Twitter nel periodo in cui l'hashtag #5GCoronavirus è stato di tendenza su Twitter nel Regno Unito. In particolare, si sono analizzate le strutture di rete di gruppi di utenti di tale social network: una corrispondente al gruppo di "isolati" e una al gruppo di "trasmissione", osservando inoltre l'assenza di una figura autorevole che combattesse tale disinformazione in maniera attiva.

2.5.3. Metodologia: modellazione di topic

Il modello Latent Dirichlet Allocation è stato già impiegato nel lavoro di ricerca degli autori in (56). Tale ricerca si è incentrata sugli articoli condivisi via Whatsapp o pubblicati su Twitter riferiti all'argomento Coronavirus e dichiarati falsi da una società di Fact-Checking nel contesto africano. Il lavoro si è proposto come punto di partenza per la comprensione delle tematiche più comuni che si sono diffuse tra le comunità nei vari social, creando la base per individuare le tematiche su cui le organizzazioni della salute e governative dovrebbero diffondere maggior informazione.

Un nuovo modello per l'identificazione dei topic è stato introdotto dagli autori in (57). Essi, ispirandosi alle catene di Markov, hanno analizzato i tweet pubblicati, raggruppati ogni 15 giorni, da account Twitter di Fact-Checking brasiliani cercando di identificare gli argomenti che sono diventati all'ordine del giorno durante i mesi della pandemia. Sono state inserite tali tendenze in un'analisi di serie temporali per monitorare la diffusione degli argomenti nel tempo. Sono state confrontate organizzazioni di Fast-Checking

identificando somiglianze e differenze nei contenuti pubblicati. Il lavoro ha ottenuto dei risultati ottimali per raggruppare gli argomenti degli scenari relativi al Covid-19, rilevando per esempio un intreccio tra la politica e la crisi sanitaria. L'algoritmo LDA, insieme al LSA, sono serviti in questo lavoro nell'ottimizzazione della fase di preelaborazione dei dati, valutando i termini e le espressioni più comuni.

L'articolo (58) ha analizzato circa 700 mila post sulla piattaforma Weibo (una piattaforma di microblogging cinese) durante i mesi della prima ondata di Coronavirus (dal 1 Gennaio 2020 al 30 Giugno 2020), utilizzando la modellazione degli argomenti LDA e l'analisi del sentiment. Sono stati estratti i principali topic di risposta alla crisi pandemica come il sostegno al personale in prima linea, l'incoraggiamento tra gli utenti, l'espressione di preoccupazione per il ripristino economico e della vita quotidiana. Per la visualizzazione dei topic estratti e dei 30 termini più frequenti nella collezione è stato utilizzato il metodo LDAvis.

Il modello LDA è stato utilizzato anche dagli autori in (59). Il lavoro si è concentrato sull'analisi di tematiche nei testi del social network cinese Weibo durante i primi mesi della pandemia da Coronavirus. Sono stati estratti i topic più rilevanti discussi dagli utenti, focalizzando l'attenzione sulle caratteristiche del cambiamento emotivo analizzate da prospettive spaziotemporali diverse. Si è riscontrato che il tema dell'epidemia su Weibo si è progressivamente ridotto dal 24 gennaio, nonostante la percentuale di ricerche per argomento epidemico è gradualmente aumentata. Tra i risultati è stato inoltre notato che i social media sono maggiormente utilizzati nel centro politico ed economico della Cina; a tal proposito, è stato riscontrato che sono stati pubblicati più testi nel blog nelle città come Pechino e Shanghai.

Capitolo 3

Metodologie di analisi

In questo capitolo saranno presentati i dati dei Social Network analizzati sia nella fase di analisi descrittiva sia nella fase di analisi testuale. Saranno spiegate le modalità di raccoglimento e di estrazione di tali dati e le metodologie utilizzate per il raggiungimento dei risultati, che saranno illustrati nel successivo capitolo.

3.1. Presentazione dei dataset

Prima di procedere con la descrizione dei dataset utilizzati, si è ritenuto opportuno indagare le peculiarità dei due Social Network su cui tale lavoro si basa, al fine di poter ottenere una migliore comprensione delle metriche successivamente presentate.

3.1.1. Facebook e Instagram

Facebook e **Instagram** sono due dei principali Social Network per numero di persone iscritte e di persone attive. Secondo i dati raccolti da (60) aggiornati al 2020, Facebook oggi è la piattaforma più popolare nel mondo, con 2,74 miliardi di utenti attivi; Instagram è il quarto Social Network nel mondo (dopo Youtube, Whatsapp e Facebook Messenger), con 1,22 miliardi di utenti attivi. La distribuzione degli utenti è leggermente differente nei due Social, in quanto Instagram ha una percentuale media di utenti maggiore nelle tre fasce d'età 13-17, 18-24, 25-34, mentre Facebook ha una percentuale di utenti inferiore nella fascia d'età 18-24 e più marcata nelle fasce d'età più adulte.

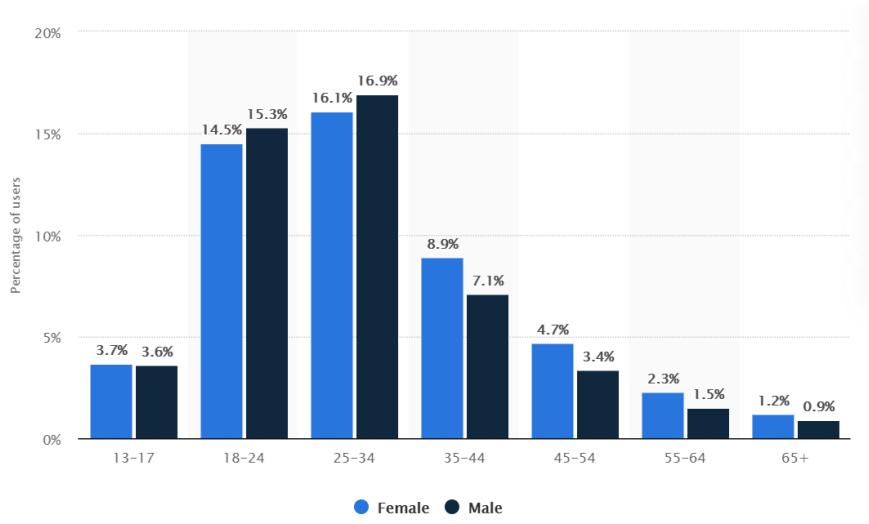

Figura 13 – Distribuzione di utenti su Instagram [Fonte: statista 2020]

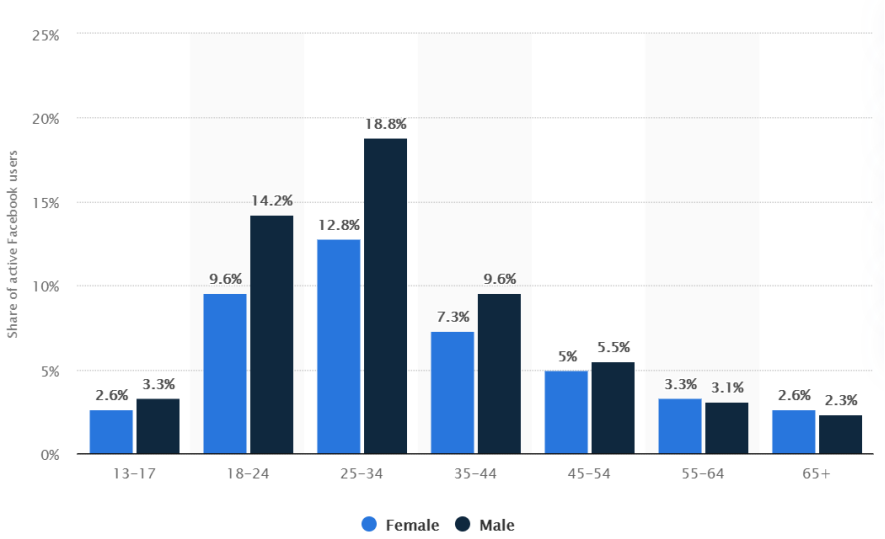

Figura 14 – Distribuzione di utenti su Facebook [Fonte: statista 2020]

I due Social Network presentano due storie e funzionalità diverse, ma anche alcune similarità. Facebook è una piattaforma destinata al networking, nata come strumento utilizzato dalle persone per "restare in contatto". Non a caso, le relazioni che si creano all'interno del Social sono propriamente dette "amicizie". Ogni utente ha una propria pagina su cui condividere post, foto, video e link esterni con i propri amici. Ciascun post, compatibilmente con le impostazioni sulla privacy impostate da ciascun utente, può ricevere un commento e/o una reazione e può eventualmente essere ri-condiviso sul "diario" personale di un "amico". In aggiunta ai profili personali, tale piattaforma permette la creazione di pagine pubbliche utilizzate prevalentemente dalle aziende o dalle personalità popolari come i personaggi dello spettacolo o della politica. Per poter seguire i contenuti creati da tali pagine è opportuno cliccare il cosiddetto "Mi Piace" alla pagina.

Instagram è nato invece come una applicazione per la condivisione di foto e video. I contenuti condivisibili su tale piattaforma sono, infatti, esclusivamente multimediali. Alle foto e ai video pubblicati, però, è possibile aggiungere anche una didascalia testuale, definita "caption". Anche in questa piattaforma è possibile aggiungere una reazione e/o un commento al post pubblicato. Differentemente da Facebook, le relazioni in essere su tale piattaforma si basano esclusivamente sul concetto di "Follower". A questo proposito, per poter visualizzare i contenuti di un certo profilo sul proprio feed personale è infatti necessario diventare un "follower" di un altro profilo, vale a dire scegliere di "seguire" i contenuti creati da tale utente.

3.1.2. Raccolta dei dati

Grazie alla collaborazione con i prof. Luca Vassio e prof. Luca Cagliero, membri del centro di ricerca SmartData@PoliTO, è stato possibile accedere al cluster nel quale sono stati raccolti i dataset relativi ai Social Network di Instagram e di Facebook. I dati sono stati ottenuti dal team di Big Data del centro di ricerca del Politecnico di Torino tramite l'utilizzo di un *crawler* messo in atto dal dicembre 2018. Un crawler è un software che analizza i contenuti di una rete o di un database in un modo metodico e automatizzato e che acquisisce una copia testuale di tutti i documenti presenti in una pagina web creando un indice che ne permetta, successivamente, la ricerca e la visualizzazione (61). Tale operazione è stata effettuata sulle pagine web riferite a Facebook e Instagram. Il *crawler* ha raccolto le attività di un numero selezionato di profili pubblici che sono stati monitorati in tempo reale, scaricandone i metadati e tutti i contenuti generati come, per esempio, i loro post. Per tali post, il crawler ha scaricato tutti i commenti scritti da ciascun utente nelle prime 24 ore dopo la pubblicazione del medesimo post. Per ottenere dati relativi esclusivamente a figure pubbliche rilevanti, è stato scelto dal team di monitorare solo i profili aventi almeno 10.000 followers. La raccolta dei dati è stata guidata da un ulteriore filtro riferito alla lingua: sono stati infatti considerati solo i profili i cui post erano composti da almeno il 40% delle parole in lingua italiana.

I dati raccolti sono stati collocati in dataset diversi a seconda dei riferimenti su cui si basano. Nell'analisi oggetto di tale tesi sono stati utilizzate tre sorgenti dato, per ciascun Social Network, che sono stati opportunamente uniti da operazioni di *join* in SQL. Di seguito sono descritti i dataset di partenza:

- **"posts"** e "**medias"** sono i dataset rispettivamente riferiti ai post pubblicati su Facebook e Instagram. Contengono tutti i metadati relativi ai

post raccolti dal crawler. Ad esempio, in questo dataset, sono stati raccolti l'username del creatore del post, il testo, il codice univoco, il numero di reazioni ricevute, la data di pubblicazione.

- **"comments"**, separatamente per i due Social Network, è il dataset contenente tutti i metadati relativi ai commenti raccolti dal crawler nelle prime 24 ore dalla pubblicazione del post, come per esempio l'username del creatore del commento, gli eventuali username menzionati, il codice univoco del post a cui si riferisce tale commento, il codice univoco del commento e la data di pubblicazione.
- **"profiles_periodic"** è il dataset in cui sono state raccolte le osservazioni giornaliere del crawler sulla pagina dei profili, separatamente di Instagram e di Facebook. Per ogni giorno e per ogni profilo sono state collezionate le informazioni sulla biografia, il numero di follower, il numero di post pubblicati e tutti i dettagli estrapolabili dalla pagina pubblica.

3.1.3. Metodologie di estrazione

Partendo da queste sorgenti dato, al fine di porre l'attenzione su aspetti diversi durante le fasi di analisi, sono state effettuate diverse estrazioni tramite interrogazioni SQL con l'utilizzo di PySpark SQL. Il filo conduttore di tali estrazioni è stata la scelta di focalizzarsi sull'analisi dei contenuti creati esclusivamente da profili di categoria "politica" nel periodo corrispondente ai mesi della prima ondata da Coronavirus in Italia. Tale scelta è stata portata avanti anche con lo studio del colore politico di ciascuna personalità di cui è stata analizzata la "vita" social, motivo per cui è stata manualmente effettuata una ricerca dei partiti politici di appartenenza degli influencer presenti nei dataset di Instagram e di Facebook. Tali partiti sono stati opportunamente suddivisi in tre fazioni principali: "Lega + FdI + FI", "Centrosinistra + PD" e "M5S". Tale scelta, effettuata in maniera arbitraria, ha avuto come scopo la riduzione del numero di comparazione nei grafici.

Tutti i campi, provenienti da ciascuno dei dataset precedentemente descritti, che sono stati utilizzati nelle varie fasi dell'analisi e che sono risultati utili al raggiungimento degli obiettivi preposti sono stati raccolti nella seguente tabella:

Campo	Descrizione
post_code (Facebook)/ short_code (Instagram)	Codice univoco del post
post_created_day	Data di pubblicazione del post
post_username	Username del creatore del post
post (Facebook)	Testo del post
caption (Instagram)	Testo del post
reactions (Facebook)	Numero di reazioni al post
num_like (Instagram)	Numero di like al post
comment_created_day	Data di pubblicazione del commento
comment_username	Username del creatore del commento
comment_id	Codice univoco del commento
party	Aggregazione di partiti politici di appartenenza
category	Categoria del profilo creatore del post
media_follower	Media mensile del numero di follower del creatore del post

Tabella 1 – Descrizione dei campi dei dataset in uso

È possibile individuare due principali estrazioni che hanno portato ad ottenere i dati che sono stati successivamente analizzati:

- **Estrazione periodo Covid**

Per poter osservare le interazioni delle personalità politiche con il loro pubblico sui Social Network e per presentare i profili maggiormente attivi e maggiormente seguiti durante il periodo di prima ondata di pandemia Covid -19 in Italia, è stato scelto di filtrare solo i post pubblicati nell'arco temporale compreso tra il 1° gennaio 2020 e il 30 giugno 2020. Inoltre, sono state calcolate le metriche di interazione degli utenti in risposta a tali post, come il totale dei commenti, delle reazioni e la media del numero di follower. Viene di seguito riportato l'esempio di codice pySpark SQL utilizzato per l'estrazione dai dataset relativi a Facebook:

```
returnDF = spark.sql("SELECT \
                        p.post_code, \
                        p.created_day as post_created_day,\
                        MONTH(p.created_day) as month, \
                        p.owner as post_username, \
                        MAX(c.post_reactions) as tot_reactions, \
                        COUNT(c.comment_id) as tot_comments,\
                        COUNT(DISTINCT c.username) as num_commentatori \
                    FROM posts p, comments c, fb_list l \
                    WHERE (l.username = p.owner)  \
                    AND (p.post_code = c.post_code) \
                    AND (l.category = 'politics') \
                    AND (p.created_day >= '2020-01-01') \
                    AND (p.created_day < '2020-07-01')\
                    GROUP BY p.post_code, p.created_day, p.owner")
resultPandas =returnDF.toPandas()
```

Il seguente codice Pyspark è stato utile per il calcolo della media mensile dei follower di ciascun profilo politico:

```
returnDF = spark.sql("SELECT \
                        profile as post_username, \
                        month(day) as month,\
                        cast(avg(likes) as int) as media_followers \
                    FROM profiles pr \
                    WHERE (pr.day >= '2020-01-01')\
                          AND (pr.day < '2020-07-01') \
                    GROUP BY profile, month(day)")
resultPandas =returnDF.toPandas()
```

Nella seguente tabella sono raccolti i valori corrispondenti a tale estrazione:

	#post_username	#post	#commenti	#reazioni
Facebook	215	68 102	31 618 360	163 984 256
Instagram	97	25 803	6 823 577	106 671 424

Tabella 2 – Caratteristiche dataset periodo Covid-19

- **Estrazione argomento "Covid"**

Poiché la ricerca si è focalizzata sugli argomenti diffusi relativamente al tema "Coronavirus", si è scelto di filtrare soltanto i post che contenessero alcune delle keywords principali che potessero riferirsi a tale tematica. In particolare, sono stati estratti

i post pubblicati dai profili politici tra il 1° gennaio 2020 e il 30 giugno 2020 contenenti almeno una tra le parole contenute nella seguente lista:

['Covid','Coronavirus', 'Pandemia', 'Virus', 'Contagi', 'Epidemia,' 'Quarantena']

L'esempio del codice utilizzato per tale estrazione per i dati di Instagram è riportato di seguito:

```
returnDF = spark.sql("SELECT m.short_code, \
                             m.created_day as post_created_day,\
                             MONTH(m.created_day) as month,\
                             m.owner_username as post_username,\
                             m.likes_count,\
                             m.comments_count,\
                             COUNT(DISTINCT c.owner_id) as num_commentatori \
                      FROM medias m, comments c \
                      WHERE m.short_code = c.media_code \
                            AND (m.caption LIKE '%covid%' \
                            OR m.caption LIKE '%coronavirus%'\
                            OR m.caption LIKE '%pandemia%'\
                            OR m.caption LIKE '%virus%' \
                            OR m.caption LIKE '%contagi%'\
                            OR m.caption LIKE '%epidemia%') \
                            AND m.created_day >= '2020-01-01'\
                            AND m.created_day < '2020-07-01' \
                      GROUP BY m.short_code, m.created_day, m.owner_username,\
                               m.likes_count, m.comments_count")
df =returnDF.toPandas()
```

Nella seguente tabella sono riassunte le caratteristiche dei due dataset estratti:

	#post_username	#post	#commenti	#reazioni
Facebook	204	6 793	3 439 008	16 998 084
Instagram	88	4 895	908 221	14 166 396

Tabella 3- Caratteristiche dataset argomento"Covid"

3.2. L'analisi descrittiva

Dopo aver ottenuto i dataset su cui focalizzare lo studio, è stata effettuata un'analisi di caratterizzazione dei dati di entrambi i Social Network. (62) In particolare, sono state analizzate le metriche più comunemente utilizzate per calcolare l'attività dei profili e il

loro seguito. Saranno di seguito descritte le metriche utilizzate per il calcolo dell'attività e del seguito dei profili oggetto di tale analisi e le metodologie con cui sono state costruite con i dati a disposizione. Tali metriche sono state calcolate e visualizzate tramite il tool di *data visualization* Power BI.

3.2.1. Analisi periodo Covid

- *Numero medio di followers*

In primis, per individuare le personalità le cui pagine risultano più seguite sui Social Network, sono stati rappresentati i primi dieci politici per numero più elevato di follower. È da sottolineare che il numero di follower preso in considerazione in questo caso corrisponde alla media del numero medio di follower mensili da gennaio e giugno per ciascun profilo.

- *Numero medio giornaliero di post pubblicati*

Al fine di individuare i personaggi politici che producono più contenuti giornalieri sui Facebook e Instagram, è stata calcolata la media giornaliera di post pubblicati nei sei mesi presi in considerazione nell'analisi. È stata cioè calcolata la somma dei post pubblicati da gennaio a giugno per ciascun profilo e tale somma è stata suddivisa per un numero fisso di 180 giorni, in modo tale da considerare anche i giorni con 0 condivisioni.

- *Numero medio di commenti per post*

La risposta degli utenti a ciascun post pubblicato dai profili politici è un indice interessante per la conoscenza del seguito di ciascuna personalità in quanto, da questo numero si comprende quanto ogni post alimenti eventuali discussioni in merito alla tematica trattata. Tale metrica è stata calcolata raggruppando, per ciascuna personalità politica, la somma del numero di commenti ricevuti durante i sei mesi individuati e suddividendo per la somma del numero di post pubblicati. La stessa misura è stata poi calcolata raggruppando i commenti e i post per fazione politica.

- *Numero medio di likes/reazioni per post*

Tale metrica è stata calcolata allo stesso modo per entrambi i Social Network, nonostante il significato di tale indicatore sia leggermente diverso. Il "likes" di Instagram dimostra

infatti, un apprezzamento al post (la piattaforma prevede un'unica reazione che viene raffigurata con l'icona di un cuore), mentre le "reazioni" di Facebook possono essere molteplici e indicare, appunto, emozioni diverse. Ciò significa che la reazione ad un post può denotare sia un significato positivo, quindi un'approvazione, sia un'emozione negativa. La metrica è stata calcolata raggruppando prima per singolo profilo e poi per fazione politica, e suddividendo la somma dei likes/reazioni ottenute per tutto il periodo per la somma del numero di post pubblicati.

- *Media di commenti ogni mille follower*

Un'altra metrica utilizzata per conoscere il grado di coinvolgimento medio degli utenti sui Social Network è il calcolo della percentuale di followers che interagiscono con i post pubblicati da un profilo. Tale indicatore è stato calcolato sommando per ciascun post di ciascun profilo il numero dei commenti ricevuti e suddividendo tale somma per la media del numero di follower del mese in cui il post è stato pubblicato. Per ciascun profilo, quindi, è stata calcolata la media dei valori ottenuti per ciascun post. Tale valore è stato normalizzato per 1000, per ottenere quindi il numero di commenti che vengono pubblicati sotto un post ogni mille follower. Per ottenere il dato riferito all'aggregazione per fazione politica, in questo caso, è stata calcolata la media della media calcolata per i valori delle personalità appartenenti a ciascuna categoria.

- *Media di like ogni mille follower*

Stesso calcolo effettuato per il numero di commenti ogni mille follower, sia per quanto riguarda i singoli profili sia i partiti politici, è stato effettuato per il numero medio di like.

- *Andamento temporale di pubblicazione post*

È stato anche misurato l'andamento di pubblicazione dei post da parte dei personaggi politici raggruppati per fazione partitica. Per ottenere una visualizzazione chiara sono state raggruppate le date di pubblicazione dei post per settimana in modo da rappresentare in un grafico di trend l'andamento settimanale della pubblicazione dei contenuti su entrambi i social.

3.2.2. Analisi dei contenuti sul Coronavirus

L'idea alla base di tale analisi è stata quella di effettuare un confronto tra i due Social Network in quanto a creazione di contenuti sull'argomento Coronavirus. Inoltre, l'obiettivo che ha guidato l'analisi è stato quello di visualizzare le personalità che hanno portato ad una maggiore diffusione di tale tematica sulle piattaforme social, e quelle che hanno raggiunto un maggior numero di utenti ottenendo il loro coinvolgimento.
Con l'intento di confrontare i due social network è stato misurato l'andamento temporale di pubblicazione post con argomento "Coronavirus" e i relativi commenti nelle due piattaforme. Ciò è stato possibile raggruppando per settimana i post pubblicati dai profili presi in considerazione e, separatamente, i commenti da loro ricevuti su Instagram e su Facebook. Per poter confrontare gli andamenti nei due differenti Social Network è stato scelto di normalizzare i valori per il numero di post complessivi pubblicati su ciascuna piattaforma. Stesso ragionamento è stato portato avanti nel caso dell'andamento temporale dei commenti, i cui valori sono stati normalizzati per il numero complessivo di commenti pubblicati. Tutti i grafici che mostrano, quindi, un confronto tra i due Social Network, rappresentano l' **"Empirical Probability Density Function"** (63) e la loro somma è pari a 1.
Per l'individuazione delle personalità che hanno creato un maggior numero di contenuti sulla tematica e che hanno riscontrato maggior successo, è stato calcolato il numero complessivo di post pubblicati, il numero complessivo di commenti ricevuti, il numero medio di commenti per post, il numero medio di likes/reazioni per post e il numero medio di utenti distinti che hanno commentato ciascun post. Anche in questo caso, le metriche sono state calcolate successivamente anche tenendo conto dell'aggregazione in fazioni politiche.

3.2.3. Analisi dei contenuti potenzialmente "fake"

Dopo aver analizzato l'attività e il seguito delle personalità politiche sui Social Network e, successivamente, la loro creazione di contenuti sulle tematiche legate alla pandemia Covid-19, si è deciso di osservare se, in risposta a tali contenuti pubblicati dalle personalità politiche, gli utenti diffondessero tematiche di disinformazione. Con tale intento, si è pertanto reso opportuno ricercare, tra i commenti sotto i post con argomento "Coronavirus", i testi che citassero in maniera esplicita tematiche riferite alle Fake News maggiormente diffuse durante i primi mesi di pandemia. Nello specifico, è stato scelto un set di keywords ricavate dalle Fake News più diffuse sul Coronavirus secondo il documento relativo all'Osservatorio sulla disinformazione online dell'Autorità per le

Garanzie nelle Comunicazioni del 24 aprile 2020 e il sito www.butac.it. Sono stati quindi estratti i commenti contenenti almeno una delle parole contenute nella seguente lista:

['spie cinesi', 'complotto', 'tgr leonardo', 'tg leonardo', 'arma biologica', '5g', 'vitamina', 'bill', 'gates', 'billgates', 'bill gates', 'arance', 'limoni', 'antenne', 'aglio', 'argento', 'laboratorio', 'brevetto', 'brevettato', 'vitamina C', 'vitamina D', 'abidol', 'arbidol', 'esercitazione', 'esercitazioni', '500 leoni', '5G', 'Bill', 'Bill Gates', 'BillGates', 'Gates', 'candeggina'].

Per ciascun commento è stata aggiunta una colonna denominata "**fake_flag**" a cui è stato assegnato il valore "*True*" qualora il testo comprendesse almeno una delle parole nella lista, "*False*" qualora nessuna delle parole fosse presente all'interno del commento. In questo modo è stato possibile osservare quali sono stati i profili e i partiti politici che hanno ottenuto maggiori commenti con contenuti riguardanti le tematiche relative alle Fake News su entrambe le piattaforme social. Inoltre, è stato anche valutato il numero dei post maggiormente commentati con i commenti relativi alle tematiche di disinformazione.

Successivamente, è stata effettuata una comparazione tra le varie tematiche relative alle Fake News, separando le parole contenute nella lista precedentemente mostrata e raggruppandole per argomenti. Ad esempio, la parola "5g" è stata aggregata con la parola "antenne" per individuare il topic relativo alla Fake News relativa alla correlazione dei sintomi da Coronavirus con l'installazione delle antenne per la diffusione della tecnologia di quinta generazione. Tale step è stato svolto "manualmente" effettuando una ricerca puntuale di keywords all'interno dei testi, tuttavia è da sottolineare che questo passaggio può anche essere automatizzato in quanto, per definizione, ciascun topic è descritto da un set di parole. Ciò è stato svolto con l'intento di estrarre le tre Fake News che sono state maggiormente nominate nei testi dei commenti ricevuti dai post pubblicati dalle personalità politiche. Di tali commenti estratti ne è stato analizzato l'andamento temporale rispetto alla data in cui la notizia è stata diffusa e, anche in questo caso, sono stati osservati i profili e le fazioni politiche che hanno ricevuto maggiori commenti riguardo tali specifici argomenti rispetto al numero totale dei commenti con "fake_flag" = True.

3.3. L'analisi di caratterizzazione del contenuto dei post e commenti

La seconda parte dell'analisi è stata incentrata principalmente sui testi dei contenuti creati sui Social Network, quindi, nello specifico, sui post e sui commenti. L'intento che ha portato alla scelta del modello da applicare è stato quello di automatizzare l'estrazione degli argomenti che, nella prima parte dell'analisi, è stata condotta manualmente. Infatti, inizialmente ci si è chiesti se le keywords selezionate a priori corrispondessero effettivamente ad argomenti "Fake" e a topic distinti. È stato quindi applicato il modello Latent Dirichlet Allocation (LDA) sul set di commenti filtrati durante l'analisi dei contenuti "Fake" per poter soddisfare tale richiesta.
Le questioni che hanno guidato le metodologie di analisi in questa fase sono state: l'individuazione di differenze e similarità tra gli argomenti discussi dagli utenti nei due Social Network e l'osservazione delle argomentazioni affrontate nei post pubblicati dai politici rispetto a quelle citate tra i commenti. Entrambe le analisi sono state condotte tramite l'utilizzo dell'algoritmo di modellazione dei topic LDA e hanno entrambe necessitato di una adeguata processazione dei testi. A questo proposito, è da sottolineare che, poiché i testi analizzati provengono dal contesto dei Social Network, dove quindi i caratteri utilizzati (es. emoticon, hashtag, link) sono comuni per le liste di documenti su cui è stato applicato il modello, la fase di pre-processing è stata comune.

3.3.1. Pre-processing del testo

La fase di processazione del testo è stata caratterizzata dalle seguenti fasi:

- Tokenization;
- Rimozione di parole rispetto al numero di caratteri (uguale a 1 e maggiore di 21);
- Rimozione di stopwords;
- Rimozione di simboli, emoticon e punteggiatura.

Di seguito è mostrato il codice Python con cui è stata implementata:

```
def removeWords(listOfTokens, listOfWords):
    return [token for token in listOfTokens if token not in listOfWords]

def Letters(listOfTokens):
    LetterWord = []
    for token in listOfTokens:
        if len(token) == 1 or len(token) >= 21:
            LetterWord.append(token)

    return LetterWord


def processCorpus(corpus, language):
    stopwords = nltk.corpus.stopwords.words(language)
    stopwords.extend(['poi', 'per', 'pi','com','www'])

    for document in corpus:

        index = corpus.index(document)
        corpus[index] = corpus[index].replace(u'\ufffd', '8')
        corpus[index] = corpus[index].rstrip('\n')
        corpus[index] = corpus[index].casefold()
        corpus[index] = re.sub('<[^<]+?>',' ', corpus[index])
        corpus[index] = re.sub(r'[^\w\s]', ' ', corpus[index])
        corpus[index] = re.sub('\W_',' ', corpus[index])
        corpus[index] = ' '.join([word for word in corpus[index].split(" ") if not (any
(map(str.isdigit, word)) and len(word)>5)])
        corpus[index] = re.sub("\S*@\S*\s?"," " , corpus[index])
        corpus[index] = re.sub(r'http\S+', ' ', corpus[index])
        corpus[index] = re.sub(r'https\S+', ' ', corpus[index])
        corpus[index] = re.sub(r'www\S+', ' ', corpus[index])
        corpus[index] = re.sub("["
        u"\U0001F600-\U0001F64F"
        u"\U0001F300-\U0001F5FF"
        u"\U0001F680-\U0001F6FF"
        u"\U0001F1E0-\U0001F1FF"
        u"\U0001F1F2-\U0001F1F4"
        u"\U0001F1E6-\U0001F1FF"
        u"\U0001F600-\U0001F64F"
        u"\U00002702-\U000027B0"
        u"\U000024C2-\U0001F251"
        u"\U0001f926-\U0001f937"
        u"\U0001F1F2"
        u"\U0001F1F4"
        u"\U0001F620"
        u"\u200d"
        u"\u2640-\u2642"
        "]+", ' ', corpus[index])

        listOfTokens = word_tokenize(corpus[index])
        LetterWord = Letters(listOfTokens)

        listOfTokens = removeWords(listOfTokens, stopwords)
        listOfTokens = removeWords(listOfTokens, LetterWord)
        corpus[index]= " ".join(listOfTokens)

    return corpus
```

3.3.2. Applicazione del topic modelling

A seguito della fase di processazione del testo, si è resa necessaria, per l'applicazione del modello di LDA, la scelta dei parametri da inserire in input, che vengono brevemente descritti di seguito:

- **Corpus.** Flusso di vettori di documenti o matrice sparsa in forma (num_documents, num_terms);
- **Id2word.** Mappatura dagli ID delle parole alle parole. Viene utilizzato per determinare la dimensione del vocabolario, per il debug e la stampa di argomenti;
- **Num_topics.** Numero di argomenti latenti richiesti da estrarre dal corpus;
- **Random_state.** Parametro utile per la riproducibilità;
- **Update_every.** Numero di documenti da scorrere per ogni aggiornamento;
- **Chunksize.** Numero di documenti da utilizzare in ciascun blocco di training;
- **Passes.** Numero di passaggi nel corpus durante il training;
- **Alpha.** Parametro che può essere impostato su un array 1D di lunghezza uguale al numero di argomenti previsti che esprime la convinzione a priori per la probabilità di ciascun argomento. In alternativa, si possono utilizzare strategie di selezione preliminari fornendo la stringa: "symmetric", se si decide di utilizzare una priorità simmetrica fissa per argomento; "asymmetric", se si vuole utilizzare un valore asimmetrico normalizzato fisso di 1.0/ (topic_index + sqrt(num_topics)); oppure "auto", in modo da far apprendere una priorità asimmetrica al corpus;
- **Iterations.** Numero massimo di iterazioni attraverso il corpus per inferire la distribuzione dell'argomento di un corpus;
- **Per_word_topics.** Parametro a cui assegnare il valore True se si vuole ottenere in output anche un elenco di argomenti ordinati in ordine decrescente di argomenti più probabili per ogni parola, insieme ai valori moltiplicati per il conteggio delle parole.

La scelta di tali parametri per ciascuna applicazione è avvenuta effettuando diverse ispezioni grafiche dei risultati, utilizzando come metro di verifica sia la comprensione dei topic in output, sia la marcata suddivisione dei punti associati a topic diversi mediante l'utilizzo della visualizzazione t-SNE e LDAvis.

Ciascuna analisi è stata pertanto considerata separatamente dalle altre. Nello specifico, nella fase di confronto degli argomenti di cui si è discusso all'interno dei commenti dei due differenti Social Network è stato scelto lo stesso numero di topic in maniera tale da

effettuare un confronto sulla base di un egual numero di topic, nel caso specifico, pari a 10.

Nella fase di confronto tra gli argomenti discussi nei commenti e nei post dei due differenti Social Network, la scelta è stata differente per ciascuna piattaforma.

Di seguito è mostrato un esempio di codice Python con cui è stata eseguita una delle applicazioni del modello LDA:

```
lda_model = gensim.models.ldamodel.LdaModel(corpus=bow_corpus,
                                            id2word=dictionary,
                                            num_topics=num_topics,
                                            random_state=1000,
                                            update_every=1,
                                            chunksize=100,
                                            passes=100,
                                            alpha='symmetric',
                                            iterations=1000,
                                            per_word_topics=True)
```

Capitolo 4

Risultati

4.1. Analisi di descrizione dei dataset

4.1.1. Analisi del periodo Covid-19

Con l'intento di introdurre il contesto "social" delle personalità politiche nel periodo della prima ondata di pandemia Covid-19, la prima parte dell'analisi presentata va a mostrare i risultati ottenuti sulle statistiche degli influencer politici che hanno creato un numero maggiore di contenuti nel periodo di riferimento e che hanno anche ottenuto un maggior seguito su Facebook e Instagram. Come affermato nel precedente capitolo, l'estrazione dei dataset analizzati in questa fase si riferisce ai post e ai relativi commenti pubblicati nel periodo di prima ondata di Coronavirus in Italia, dal 1° gennaio al 30 giugno.
Nella seguente tabella sono stati raccolti il numero di profili per ciascuna fazione politica di appartenenza e il numero di post pubblicati da ciascuno visualizzati come somma, media, minimo e max per profilo.

		Facebook			
Partito politico	**#profili**	**#post**	**media**	**min**	**max**
Centrosinistra + PD	76	16 637	442	1	966
Lega + FdI + FI	61	29 692	788	10	1 907
M5S	78	21 773	426	2	964
		Instagram			
Partito politico	**#profili**	**#post**	**media**	**min**	**max**
Centrosinistra + PD	31	5 708	381	1	914
Lega + FdI + FI	39	14 283	780	1	1 664
M5S	27	5 812	700	1	1 653

Tabella 4 - Somma, media, min, max dei post per profilo

Come si può notare dai dati raccolti in tabella, nei due Social Network si nota una differenza tra il numero di profili politici presenti. Su Facebook, infatti, sono presenti più del doppio dei profili rispetto all'altra piattaforma di proprietà di Zuckerberg. Ciò può

essere dovuto sia al fatto che Instagram sia un Social nato più di "recente", sia per le differenti caratteristiche della piattaforma stessa. Per quanto riguarda il numero di post pubblicati dagli influencer, in entrambi i network spicca il numero dei contenuti pubblicati dall'aggregazione del centrodestra, soprattutto nella piattaforma di Instagram, in cui il numero di contenuti pubblicati da questa fazione triplica quello dei contenuti creati dalle altre.

Nella seguente tabella sono stati raccolti i dettagli dei commenti ricevuti dai profili di ciascuna coalizione, anch'essi rappresentati in forma di somma, media, minimo e massimo.

Una delle differenze sostanziali tra i due Social Network sembra essere il numero dei commenti ricevuti dalla fazione del centrosinistra, di gran lunga inferiore su Instagram sia rispetto a quelli ricevuti su Facebook (a parità di post pubblicati), sia rispetto alle altre aggregazioni partitiche.

			Facebook		
Partito politico	**#profili**	**#commenti**	**media**	**min**	**max**
Centrosinistra + PD	76	4 388 830	136 302	5	813 616
Lega + FdI + FI	61	16 027 056	545 649	50	5 151 498
M5S	78	11 202 474	257 912	39	2 484 072
			Instagram		
Partito politico	**#profili**	**#commenti**	**media**	**min**	**max**
Centrosinistra + PD	31	495 060	29 563	0	111 947
Lega + FdI + FI	39	5 264 606	480 996	2	3 037 237
M5S	27	1 063 911	115 021	69	393 468

Tabella 5 – Somma, media, min, max dei commenti per profilo

Considerando invece l'ordine di grandezza delle reazioni e dei likes rilasciate dagli utenti, rappresentato nella tabella 6, esso risulta piuttosto simile tra i due Social: ad esempio, il numero di likes ricevuti dall'aggregazione di centrodestra è molto vicina al numero di reazioni ricevute dagli stessi partiti politici su Facebook, considerando un numero di post pari al doppio. Questo dato risulta molto interessante in quanto ci dimostra una peculiarità che è propria delle due diverse piattaforme: su Facebook prevale la discussione tra gli

utenti in risposta ai post politici, su Instagram prevale l'approvazione comunicata tramite il "like".

			Facebook		
Partito politico	**#profili**	**#reazioni**	**media**	**min**	**max**
Centrosinistra + PD	76	26 191 761	824 903	18	4 520 756
Lega + FdI + FI	61	84 676 768	2 956 278	563	30 530 504
M5S	78	53 079 727	1 205 219	466	14 036 228
			Instagram		
Partito politico	**#profili**	**#likes**	**media**	**min**	**max**
Centrosinistra + PD	31	10 424 238	614 457	2	3 003 860
Lega + FdI + FI	39	80 674 235	7 546 074	388	49 053 205
M5S	27	15 572 951	1 421 438	1 395	7 736 446

Tabella 6 – Somma, media, min, max dei likes per profilo

4.1.1.1. Numero di followers

La prima metrica che è stata calcolata è il numero medio di followers nel periodo selezionato. Questo valore ci fornisce una panoramica sulle personalità aventi un maggior seguito nelle due piattaforme. Nelle seguenti figure sono mostrati i primi 10 influencer per tale indicatore.

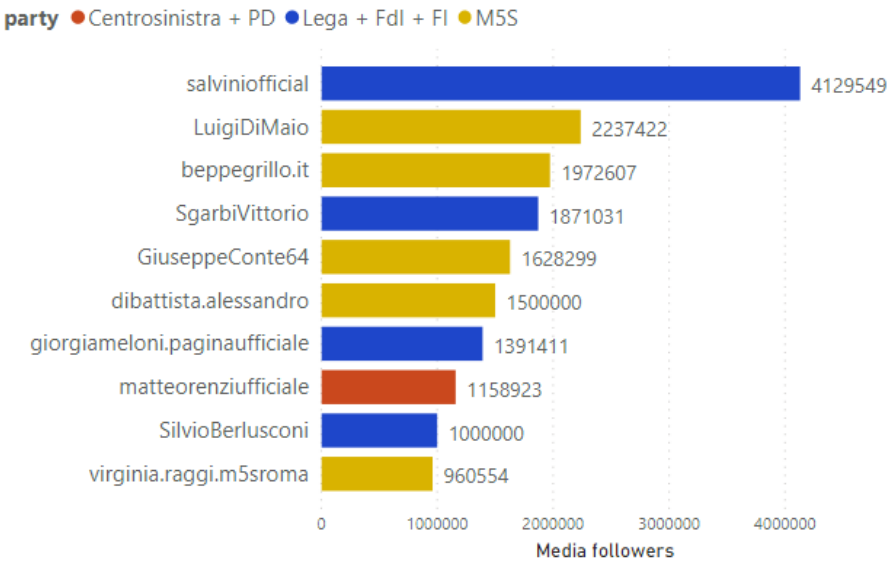

Figura 16 - Top 10 profili per media di followers su Facebook

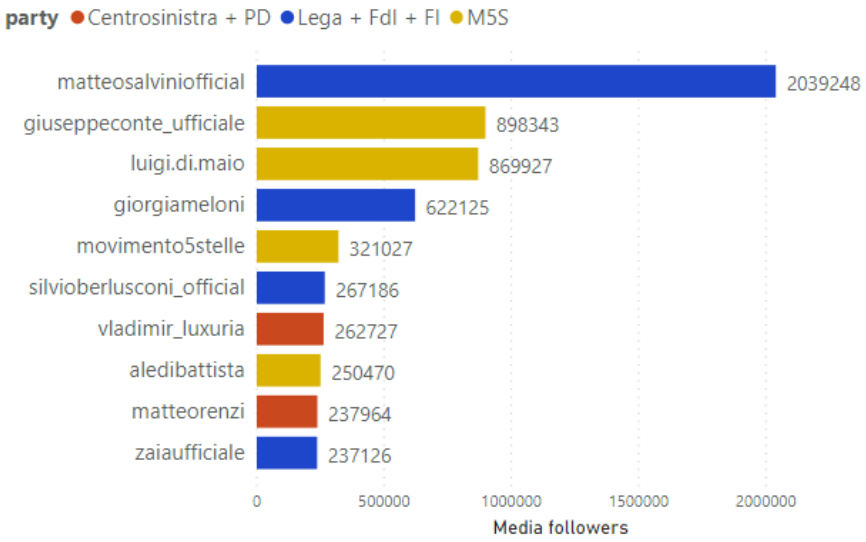

Figura 17 - Top 10 profili per media di followers su Instagram

Come si può notare, in entrambi i Social Network, le figure con maggior numero medio di followers sono per la maggior parte appartenenti alle fazioni di centrodestra e della compagine pentastellata. La figura con il maggior seguito nelle due piattaforme è Matteo Salvini, uno dei maggiori esponenti dell'opposizione e Segretario del partito di Lega Nord. Seconda e terza posizione, rispettivamente per Facebook e Instagram, per il Ministro degli Affari Esteri del Governo Conte II, appartenente al partito del M5S. In entrambi i grafici compare il nome di Giuseppe Conte, Presidente del Consiglio in carica

durante i mesi delle prime fasi della pandemia, di Giorgia Meloni, fondatrice del partito Fratelli d'Italia, Silvio Berlusconi, eurodeputato appartenente al gruppo parlamentare di Forza Italia, e Matteo Renzi, fondatore del partito di centrosinistra Italia Viva.
La maggioranza, all'interno dei profili più seguiti, di personalità appartenenti ai gruppi colorati di blu e di giallo, è confermata da quanto mostrato nelle figure 18 e 19. Il partito con, in media, un numero maggiore di follower medi per profilo politico, è l'aggregazione di Lega+FdI+FI. Segue il partito del Movimento 5 Stelle. In misura inferiore, soprattutto nel social di Instagram, vi è l'aggregazione di centrosinistra.

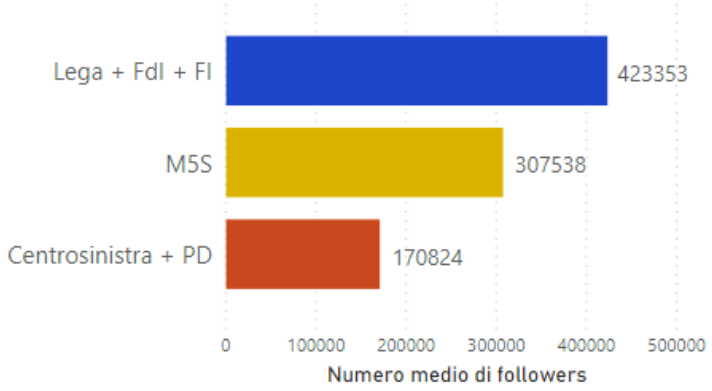

Figura 18 - Numero medio di followers per fazione politica su Facebook

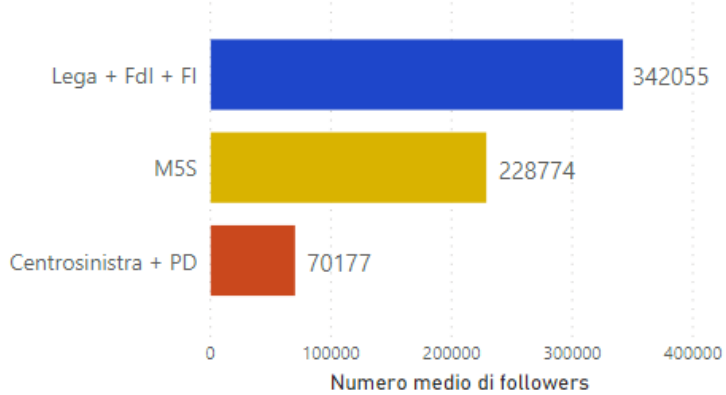

Figura 19 - Numero medio di followers per fazione politica su Instagram

4.1.1.2. Produzione giornaliera di post

Successivamente è stata analizzata la produzione giornaliera di contenuti da parte dei profili, distinguendo le figure per ciascun partito politico. Sulla base di questi valori è

stato possibile conoscere sia il numero medio di post pubblicati dai singoli profili giornalmente, sia i profili che risultano maggiormente attivi per ciascun partito politico. Nella piattaforma di Facebook, la figura che risulta più attiva nella pubblicazione di contenuti è Filippo Rossi, leader del partito Buona Destra, il quale ha pubblicato in media circa undici post al giorno durante il periodo della pandemia. Segue Matteo Salvini con circa otto post al giorno, il quale risulta il più attivo nella piattaforma di Instagram, con una media di nove post al giorno pubblicati. Spicca, in entrambi i Social Network, anche il profilo di Luca Zaia, presidente della Regione Veneto, con una media di 6-7 post al giorno condivisi.

Per quanto riguarda l'aggregazione "Centrosinistra + PD", l'influencer più attivo su Facebook durante i primi sei mesi del 2020 risulta essere Vincenzo De Luca, Presidente della Regione Campania. Spicca inoltre, in entrambi i Social, il nome di Stefano Bonaccini, Presidente della regione Emilia-Romagna, e di Nicola Zingaretti, segretario del Partito Democratico e Presidente della Regione Lazio.

Per quanto riguarda il M5S, il profilo che ha pubblicato un numero medio maggiore di post giornalieri è Virginia Raggi, sindaca della città di Roma. Con un valore leggermente inferiore, c'è Gianluigi Paragone, conduttore televisivo e iscritto, dal 4/01/2020 al Gruppo Misto. Sulla piattaforma di Instagram si osserva che il profilo maggiormente attivo risulta essere quello riferito alla pagina del partito "movimento5stelle", con circa 9 post pubblicati ogni giorno. In generale, è possibile notare che in questa piattaforma le pagine riferite ai partiti politici come "legaofficial", "partitodemocratico", "azione.it" e, appunto "movimento5stelle" sono tra le più attive.

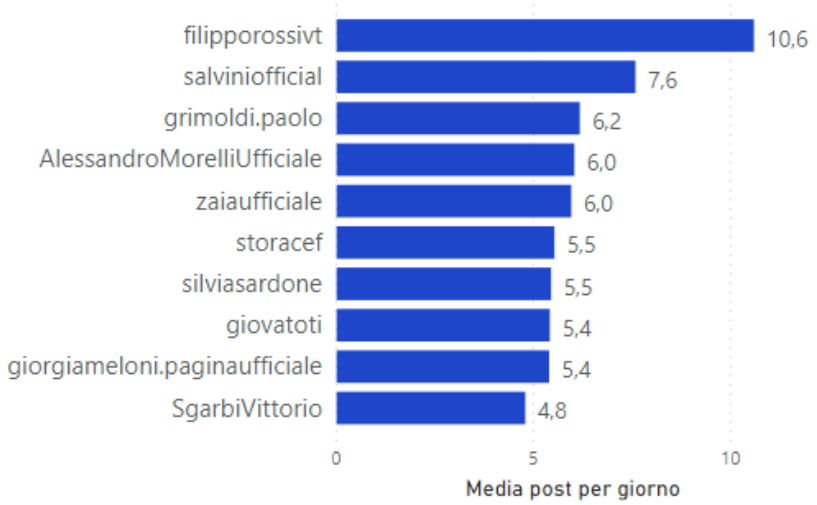

Figura 15 – Facebook: top 10 profili per produzione giornaliera di post – Lega + FdI + FI

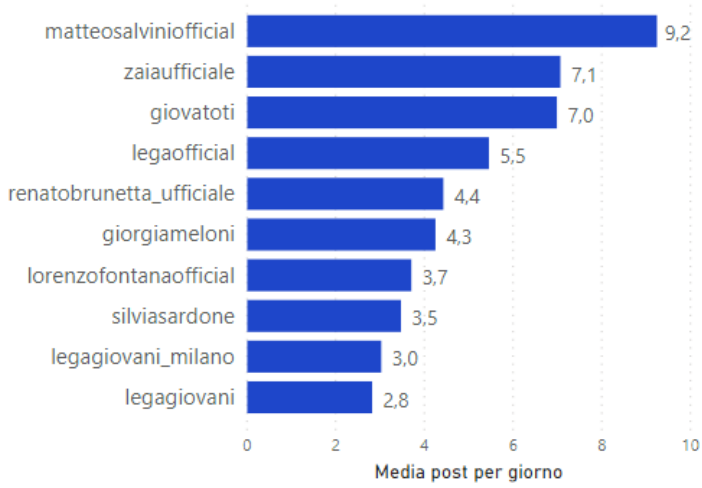

Figura 16 - Instagram: top 10 profili per produzione giornaliera di post – Lega + FdI + FI

Figura 17 – Facebook: top 10 profili per produzione giornaliera di post - Centrosinistra + PD

Figura 18 – Instagram: top 10 profili per produzione giornaliera di post - Centrosinistra + PD

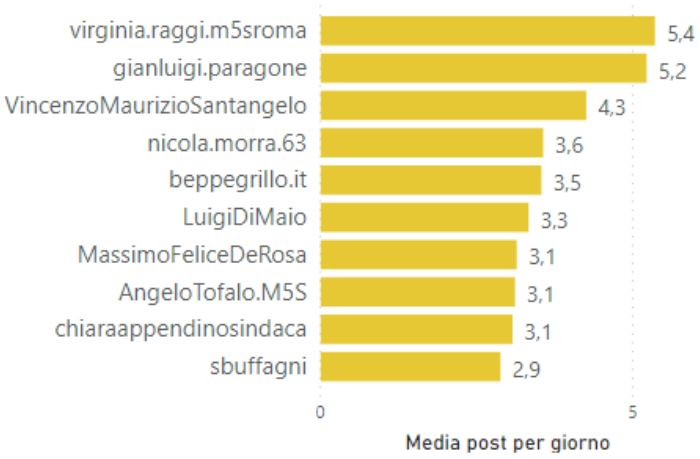

Figura 19 – Facebook: top 10 profili per produzione giornaliera di post – M5S

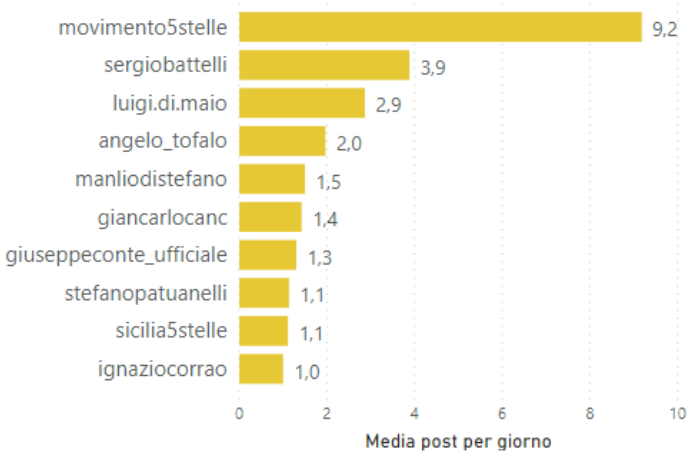

Figura 20 - Instagram: top 10 profili per produzione giornaliera di post – M5S

4.1.1.3. Misure di coinvolgimento per profilo politico

Dopo aver osservato le personalità che hanno creato più contenuti sui due Social Network analizzati, è opportuno osservare il grado di coinvolgimento degli utenti in risposta a tali contenuti, in termini di commenti e di likes e reazioni rilasciate. Nelle seguenti figure si possono notare i primi dieci politici per numero medio di commenti e di likes ricevuti per ciascun post.

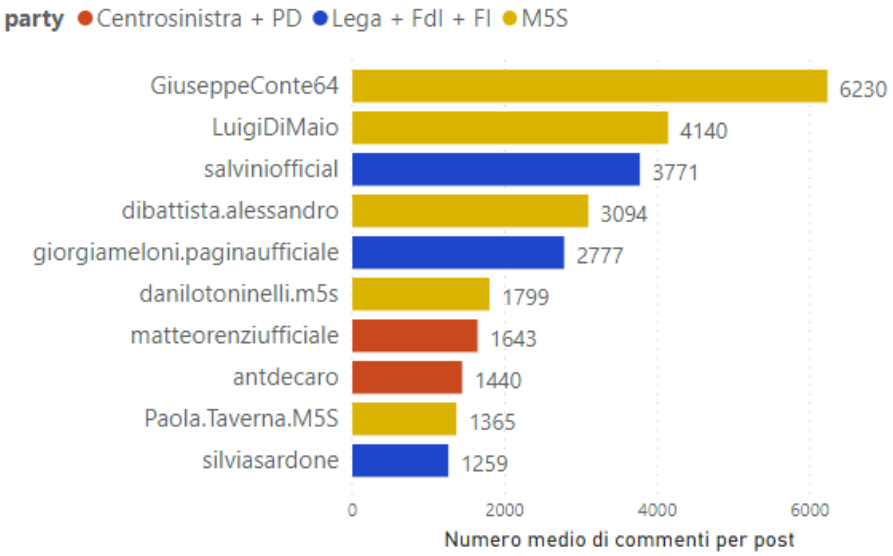

Figura 21 – Top 10 profili per numero medio di commenti per post su Facebook

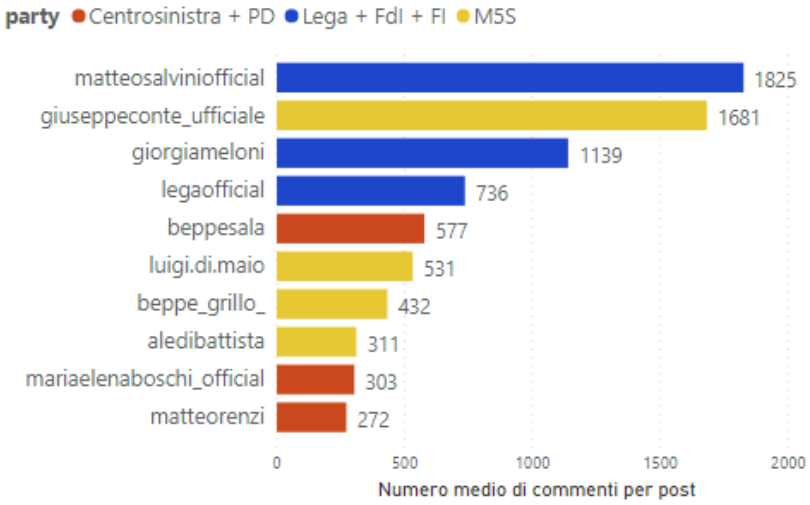

Figura 22 – Top 10 profili per numero medio di commenti per post su Instagram

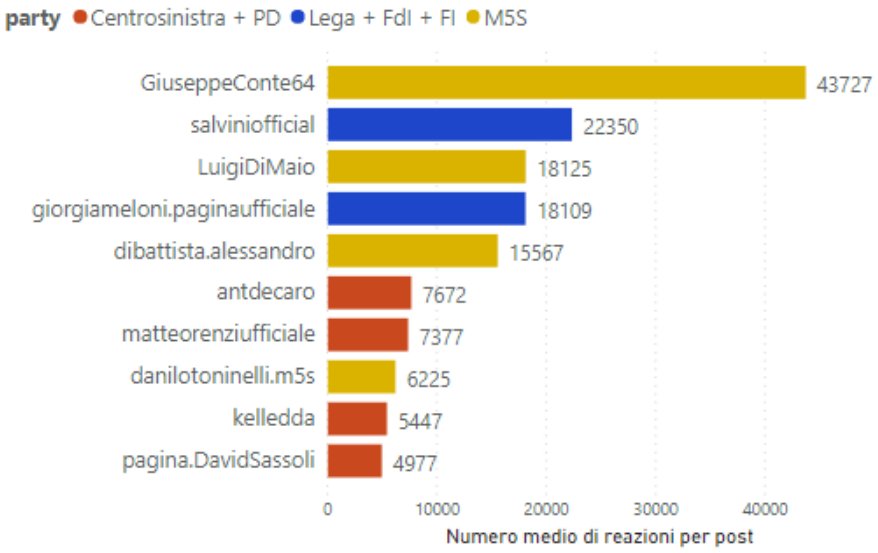

Figura 23 – Top 10 profili per numero medio di reazioni per post su Facebook

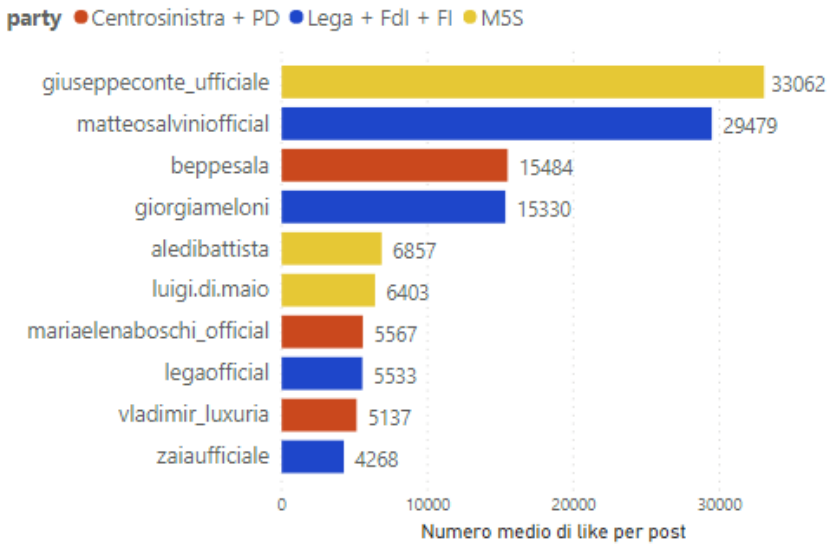

Figura 24– Top 10 profili per numero medio di like per post su Instagram

Nei seguenti diagrammi a barre sono invece rappresentati i profili che hanno ottenuto un coinvolgimento maggiore rispetto al numero di follower, che è stato normalizzato per mille. Il valore mostrato corrisponde alla media della metrica calcolata per ciascun post pubblicato da ciascun profilo suddividendo la somma dei commenti ricevuti da ciascun post per la media mensile dei follower (normalizzata per mille) di ciascun profilo. Spicca, ad esempio, il nome di Antonio Decaro, sindaco della città di Bari, su Facebook per aver ricevuto un alto numero di commenti e di reazioni ogni mille follower. Su Instagram, sia per numero di commenti sia per numero di likes ogni mille follower troviamo nella top 10 il sottosegretario di Stato Lucia Borgonzoni, appartenente al partito di Lega Nord.

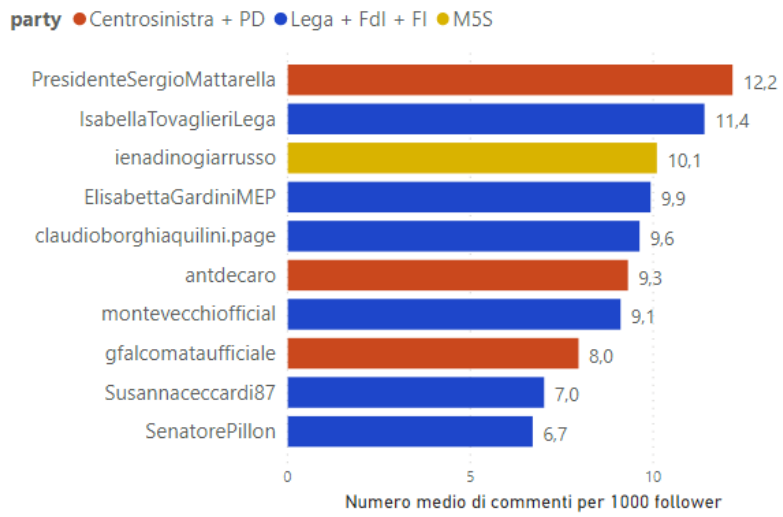

Figura 25 – Top 10 profili per numero medio di commenti ogni 1000 follower su Facebook

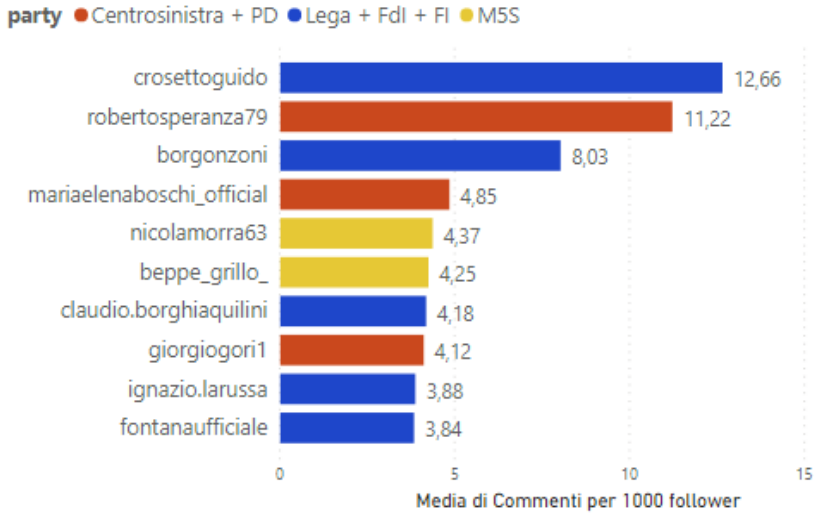

Figura 26 – Top 10 profili per numero medio di commenti ogni 1000 follower su Instagram

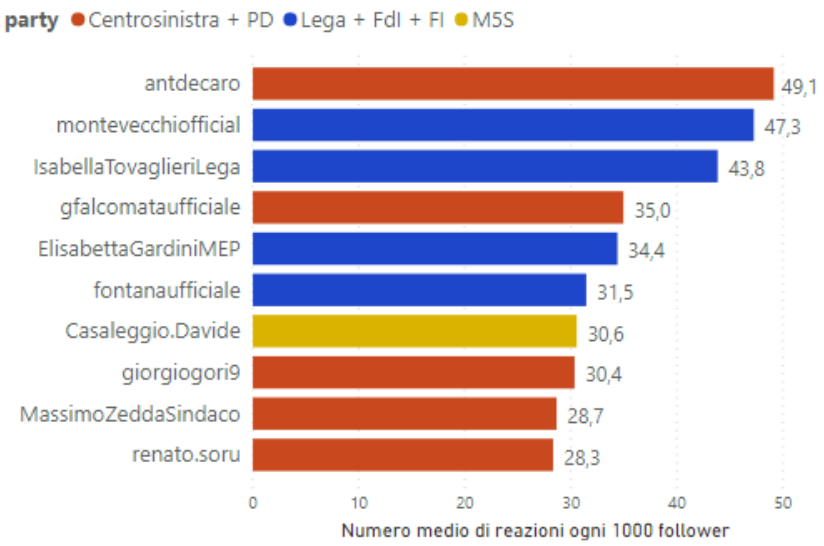

Figura 27 – Top 10 profili per numero di reazioni ogni 1000 follower su Facebook

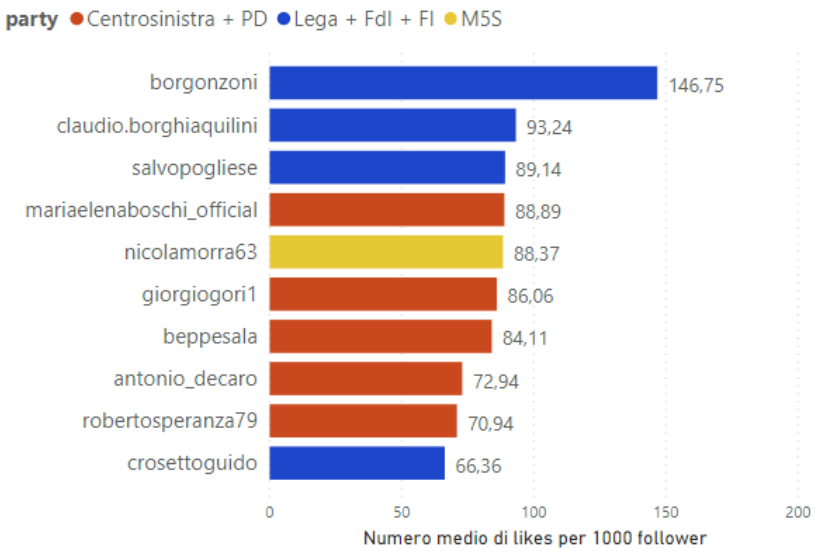

Figura 28 – Top 10 profili per numero di likes ogni 1000 follower su Instagram

4.1.1.4. Misure di coinvolgimento per partito politico

Aggregando i post per settimana e per coalizione politica, è stato possibile visualizzare l'andamento della pubblicazione di tali contenuti da ciascuno schieramento in entrambi i Social Network. Da questi grafici in figura 29 e 30 si può constatare che le curve dei due Social hanno comportamenti piuttosto differenti. Su Facebook l'andamento sembra essere più altalenante, soprattutto nel caso della fazione del Centrodestra che presenta più picchi soprattutto durante i mesi di marzo, aprile e maggio. Su Instagram, invece, i valori dei post pubblicati settimanalmente nei mesi della pandemia risultano essere stazionari per

tutte e tre le categorie. In quest'ultimo caso il distacco tra l'aggregazione dei partiti di Centrodestra e le due aggregazioni sembra essere più marcato.

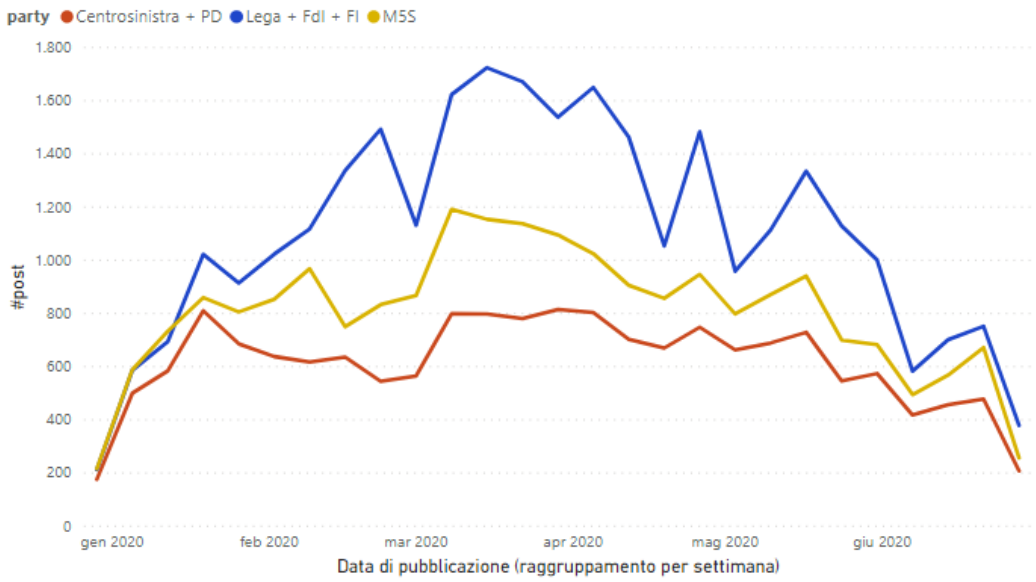

Figura 29 – Andamento pubblicazione post per fazione politica su Facebook

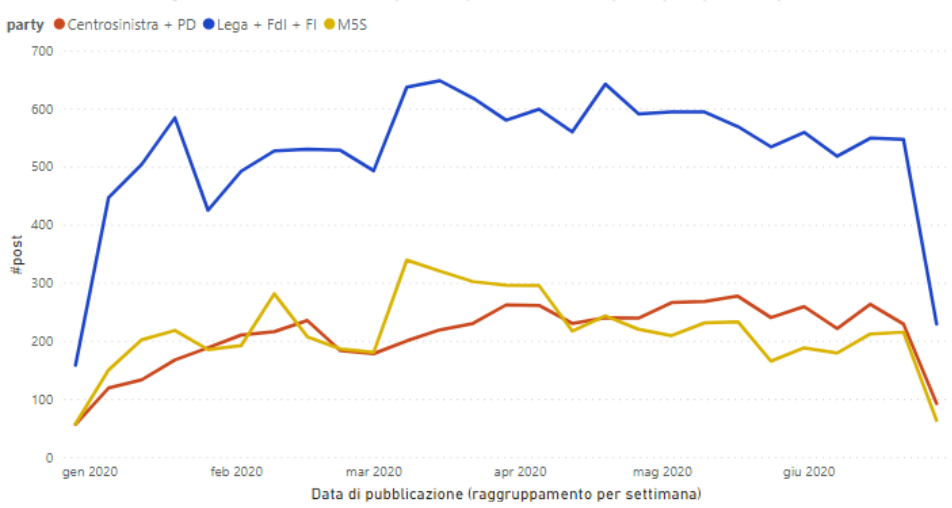

Figura 30 - Andamento pubblicazione post per fazione politica su Instagram

Nei seguenti bubble chart sono stati rappresentati i valori di engagement ottenuti per ciascuna fazione. In particolare, sull'asse delle ascisse è stata inserita la metrica riferita al numero medio di commenti per post pubblicati, sull'asse delle ordinate il numero medio di likes ricevuti. L'ampiezza delle bolle indica il numero di post condivisi da

ciascun raggruppamento. Come si può notare, in entrambi i Social Network, l'aggregazione del centrodestra è posizionata in alto a destra, mostrando un maggior engagement rispetto agli altri partiti. Il Movimento 5 Stelle risulta più vicino a tale fazione sul social di Facebook, mentre più spostata verso il basso su Instagram. La bolla di colore rosso, indicante l'aggregazione "Centrosinistra + PD" risulta essere quella che raggiunge meno utenti in entrambe le piattaforme.

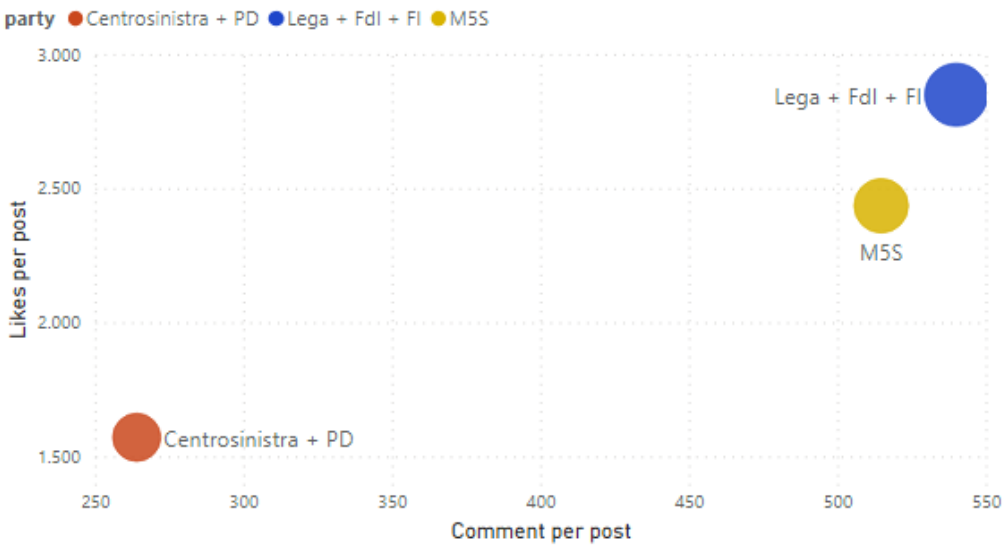

Figura 31 – Engagement per fazione politica su Facebook

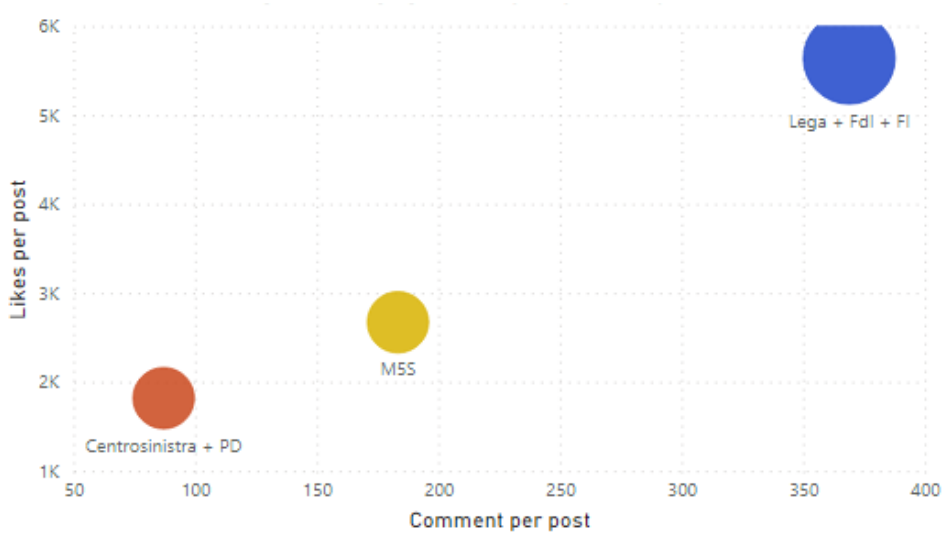

Figura 32 – Engagement per fazione politica su Instagram

Nelle seguenti figure sono stati inseriti i diagrammi a barre che rappresentano il confronto tra coalizioni politiche rispetto al numero medio di reazioni e di commenti ogni mille follower. Tale metrica è stata calcolata effettuando, per ciascun profilo una media di commenti/reazioni per mille follower e, per ciascuna fazione, è stata fatta una media di tale valore dei profili politici. Si può facilmente notare che, l'aggregazione politica che ha un maggior numero medio di interazioni per numero di follower è la stessa che ha un numero medio più basso di seguaci. Questo fenomeno è maggiormente accentuato su Instagram, mentre è totalmente ribaltato nel caso del numero medio di commenti per follower su Facebook, un dato che mostra quanto sia frequente in questo social la discussione sotto i post pubblicati da questa compagine.

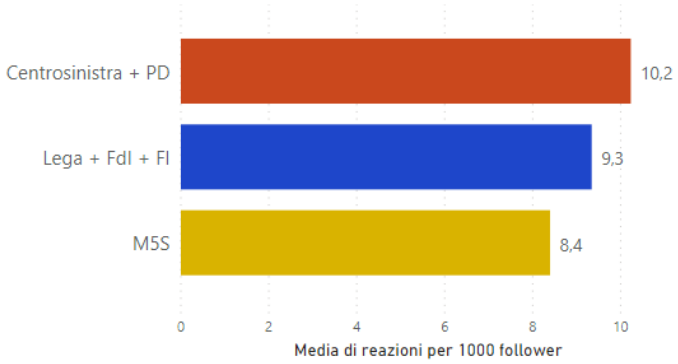

Figura 33 - Numero medio di reazioni ogni 1000 follower per fazione politica su Facebook

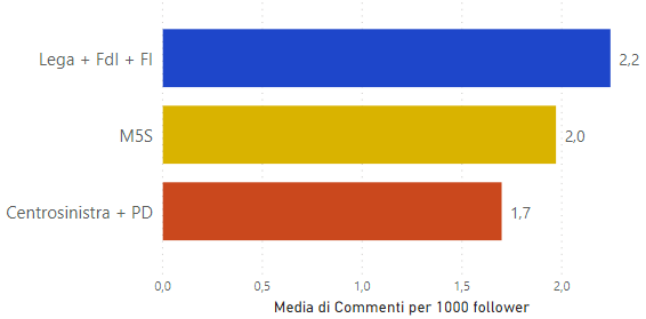

Figura 34 - Numero medio di commenti ogni 1000 follower per fazione politica su Facebook

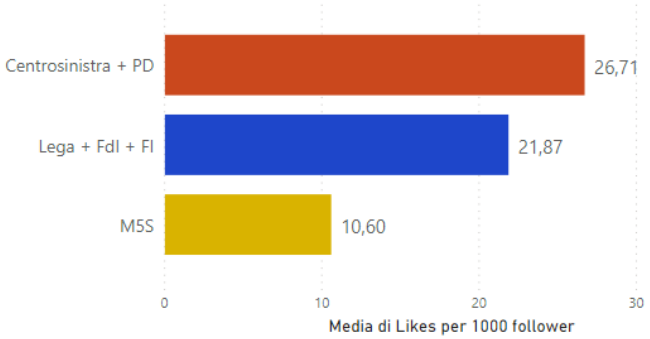

Figura 35 - Numero medio di likes ogni 1000 follower per fazione politica su Instagram

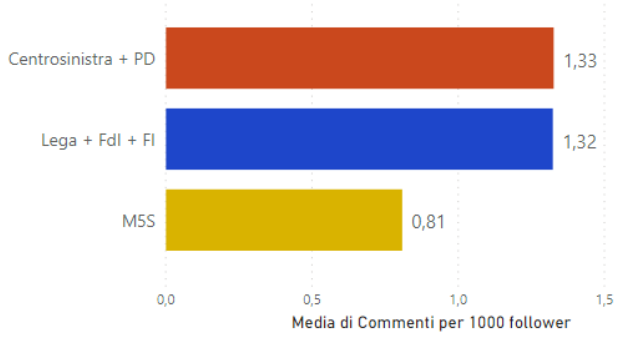

Figura 36 - Numero medio di commenti ogni 1000 follower per fazione politica su Instagram

4.1.2. Analisi dei contenuti sul Coronavirus

Passando all'estrazione di contenuti esclusivamente riguardanti l'argomento "Coronavirus", nella seguente tabella viene presentato una overview di dati che caratterizzano il dataset con questo filtro, suddividendo i contenuti per categoria politica:

			Facebook		
Partito politico	**#profili**	**#post**	**media**	**min**	**max**
Centrosinistra + PD	68	1 904	62	1	140
Lega + FdI + FI	60	2 631	85	1	199
M5S	76	2 258	49	1	119
			Instagram		
Partito politico	**#profili**	**#post**	**media**	**min**	**max**
Centrosinistra + PD	28	1 039	104	2	204
Lega + FdI + FI	36	2 395	210	1	456
M5S	24	1 461	265	1	544

Tabella 7 – Somma, media, min, max dei post con argomento Coronavirus per fazione politica

La distribuzione dei profili per le categorie dei partiti politici italiani è differente nei due Social Network. Facebook riscontra un numero pressoché simile di profili appartenenti alle diverse fazioni con una leggera dominanza del Movimento 5 Stelle.

Su Instagram invece i profili presenti in misura maggiore sono appartenenti all'aggregazione dei tre partiti di centrodestra Lega Nord, Fratelli d'Italia e Forza Italia.

Sulla base dei valori assoluti la categoria aggregata di partiti di centrodestra sembra essere quella che ha creato maggiori contenuti riferiti all'argomento "Coronavirus" su entrambi i social. Tale fazione ha ricevuto un numero di commenti molto superiore ai suoi avversari su entrambi social.

Da questi dati iniziali si può già evincere una sostanziale differenza tra i due social, dettata anche dalla loro diversa natura. I profili politici iscritti a Instagram fanno parte di un gruppo ristretto ma creano più contenuti. Come è stato riscontrato nella prima parte dell'analisi, ciò è anche dovuto alla presenza di pagine ufficiali dei partiti che risultano essere i più attivi (es. Movimento5stelleofficial e Legaofficial).

		Facebook			
Partito politico	**#profili**	**#commenti**	**media**	**min**	**max**
Centrosinistra + PD	68	709 910	31 851	15	200 743
Lega + FdI + FI	60	1 568 557	57 876	5	320 356
M5S	76	1 160 541	29 270	8	439 918
		Instagram			
Partito politico	**#profili**	**#commenti**	**media**	**min**	**max**
Centrosinistra + PD	28	81 121	7 131	9	15 328
Lega + FdI + FI	36	591 276	36 295	21	285 283
M5S	24	235 824	33 993	6	76 830

Tabella 8 – Somma, media, min, max dei commenti sotto i post con argomento Coronavirus per fazione politica

		Facebook			
Partito politico	**#profili**	**#reazioni**	**media**	**min**	**max**
Centrosinistra + PD	76	4 016 375	180 545	156	1 209 279
Lega + FdI + FI	61	7 661 931	274 856	109	1 880 126
M5S	78	5 319 778	124 834	45	1 582 100
		Instagram			
Partito politico	**#profili**	**#likes**	**media**	**min**	**max**
Centrosinistra + PD	28	1 430 591	129 253	279	339 054
Lega + FdI + FI	36	9 531 214	697 192	320	4 454 625
M5S	24	3 204 591	388 338	205	1 418 085

Tabella 9 – Somma, media, min, max dei likes ai post con argomento Coronavirus per fazione politica

Nella seguente tabella è stata calcolata la percentuale di post riferiti al Coronavirus rispetto al totale dei post pubblicati nel periodo preso in considerazione. Stesso calcolo è stato effettuato per i commenti e le reazioni.

		Facebook	
Partito politico	**post**	**commenti**	**reazioni**
Centrosinistra + PD	11,44%	16,18%	15,33%
Lega + FdI + FI	8,86%	9,79%	9,05%
M5S	10,37%	10,36%	10,02%
		Instagram	
Partito politico	**post**	**commenti**	**reazioni**
Centrosinistra + PD	18,20%	16,39%	13,72%
Lega + FdI + FI	16,77%	11,23%	11,81%
M5S	25,14%	22,17%	20,58%

Tabella 10 – Percentuale delle interazioni riferite al Coronavirus sul totale per fazione politica

La figura 37 raffigura l'andamento percentuale della quantità di post creati da parte dei profili monitorati in entrambi i social rispetto al tempo. La misura è normalizzata rispetto al numero di post totali pubblicati sui rispettivi social network. Nel grafico sono state inoltre annotate le date dei principali eventi che si sono susseguiti nei primi mesi della pandemia.

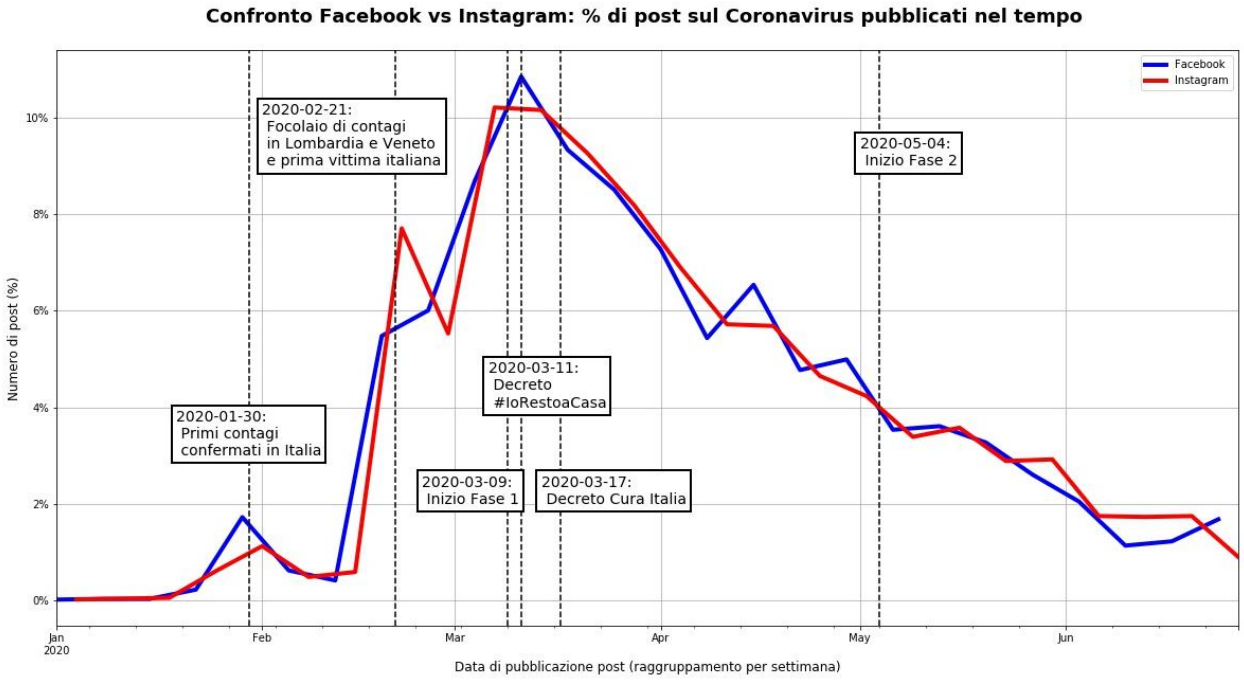

Figura 37 – Andamento pubblicazione post sul Coronavirus: confronto Facebook vs Instagram

Si può notare una sostanziale somiglianza delle due curve che vede la presenza di 3 picchi: il primo, il più basso, in corrispondenza della prima notizia sul Covid riguardante il territorio italiano; il secondo, che risulta in maniera più marcata su Instagram, a seguito del primo focolaio di contagi in Lombardia e Veneto e della notizia della prima vittima italiana e il terzo che corrisponde al momento in cui su entrambi i social si sono creati più contenuti (più del 10% del totale dei post creati) su tale argomento con il susseguirsi dei vari DPCM e con l'avvio della Fase 1 della pandemia. In seguito, le due curve presentano un andamento nel complesso decrescente, tranne nel mese di aprile in cui si possono notare lievi cambiamenti di rotta della curva riguardante Facebook.

La figura 38 mostra l'interazione dei media in risposta ai contenuti creati dalle varie figure tramite la pubblicazione di commenti nel tempo. Tale misura è stata normalizzata per il numero di commenti totali dei due dataset. Il comportamento delle due curve in questo

grafico risulta piuttosto simile seppur con determinate differenze, ad esempio la curva in rosso riferita ai commenti pubblicati dal popolo Instagram presenta un picco più elevato nel periodo di maggior interazione per entrambi i social e successivamente procede in maniera decrescente a gradini, mentre la curva in blu presenta più picchi nel periodo marzo–aprile, per poi decrescere in maniera vertiginosa a inizio maggio. Dall'osservazione di queste due curve si può notare che il social Instagram è stato maggiormente utilizzato all'inizio della pandemia e l'attività dei commentatori, al passo con i creatori di post, è stata via via meno intensa; mentre l'interazione su Facebook risulta essere stata attiva per un periodo più lungo.

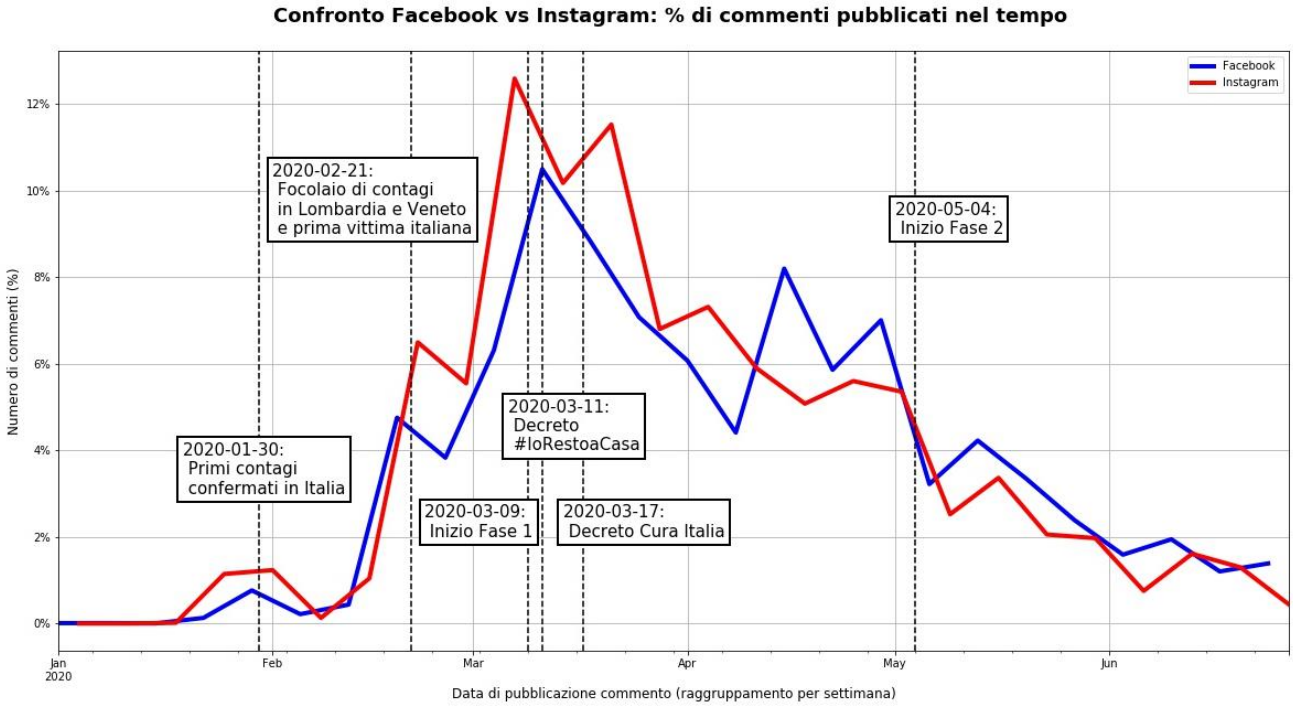

Figura 38 – Andamento pubblicazione commenti sul Coronavirus: confronto Facebook vs Instagram

4.1.2.1. Misure di interazione per profilo politico

Tra i politici più attivi sulla tematica riferita al Coronavirus ritroviamo su Facebook i due politici che erano stati individuati come tra i primi dieci politici della fazione del centrodestra, Filippo Rossi e Giovanna Toti, presidente della Regione Liguria e Nello Musumeci, presidente della Regione Sicilia. Di seguito ritroviamo Vincenzo de Luca,

influencer più attivo del Centrosinistra. Su Instagram il profilo corrispondente al partito del Movimento 5 Stelle si conferma molto attivo anche sulla tematica specifica, pubblicando ben 544 post contenenti parole relative al Coronavirus. Seguono Giovanna Toti, profilo quindi attivo in entrambi i social, e il Presidente Zaia.

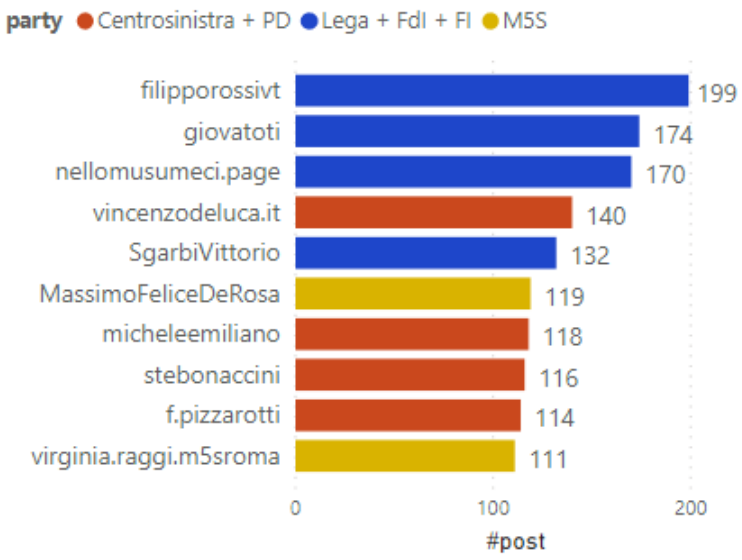

Figura 39 - Top 10 profili per numero di post sul Coronavirus su Facebook

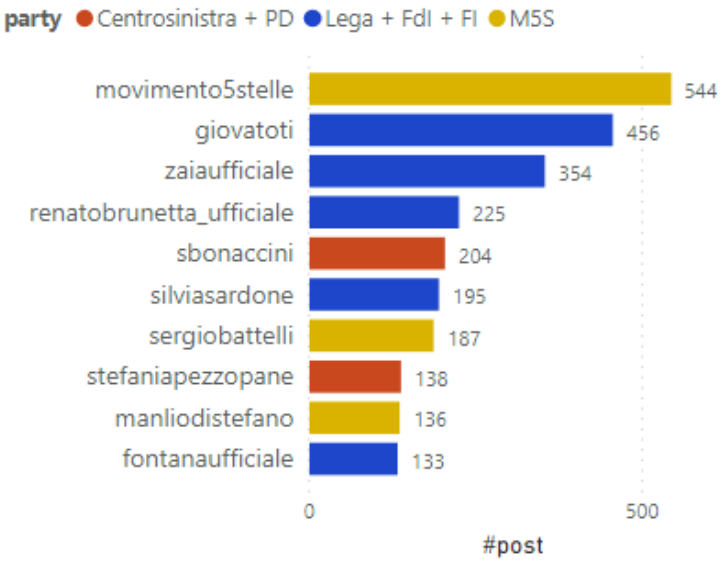

Figura 40 - Top 10 profili per numero di post sul Coronavirus su Instagram

Considerando invece il numero di interazioni ricevute sotto post riferiti alla tematica della pandemia, è possibile riscontrare che essi non corrispondono, sia nei valori assoluti che nei valori medi, alle personalità che hanno pubblicato più contenuti su tale argomento.

Nei primi 10 profili per numero assoluto e medio di interazioni si notano infatti politici come Luigi Di Maio, Matteo Salvini, Giorgia Meloni e Giuseppe Conte in entrambe le piattaforme.

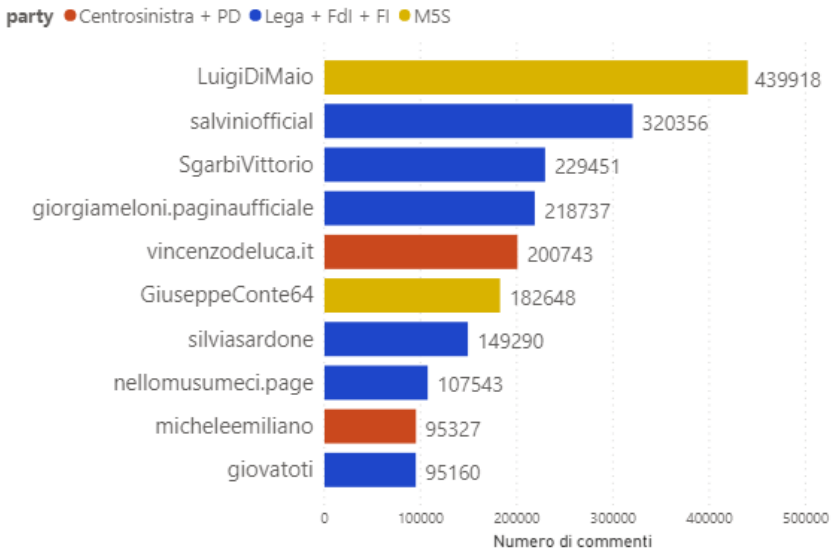

Figura 41 – Top 10 profili per numero di commenti su Facebook

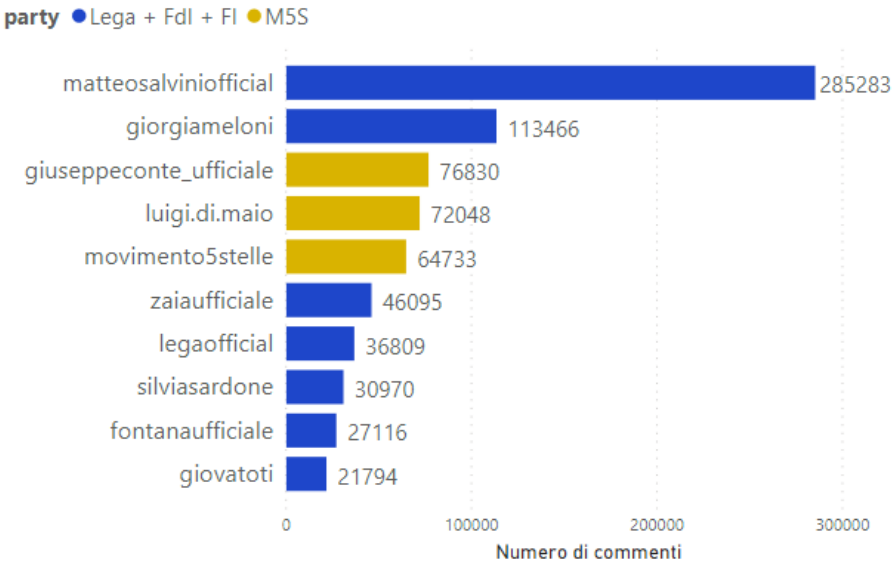

Figura 42 – Top 10 profili per numero di commenti su Instagram

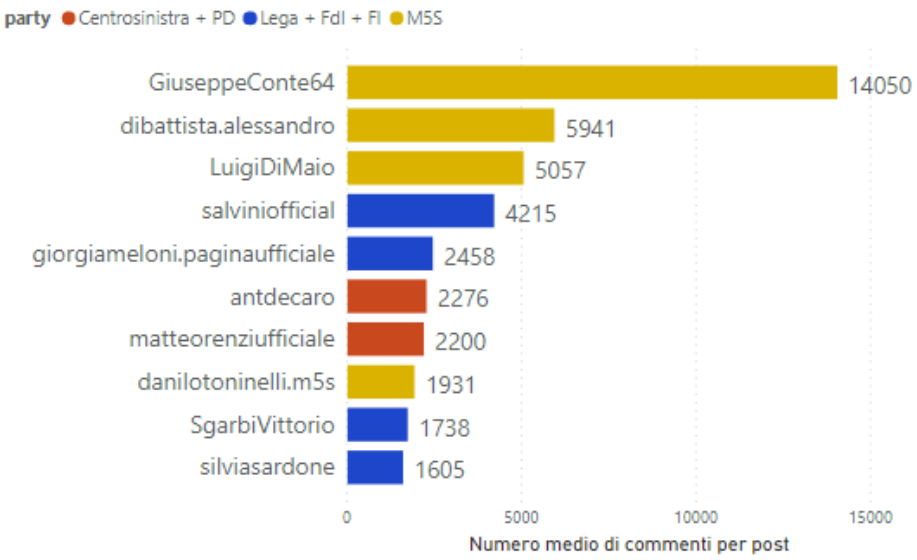

Figura 43 – Top 10 profili per numero medio di commenti per post su Facebook

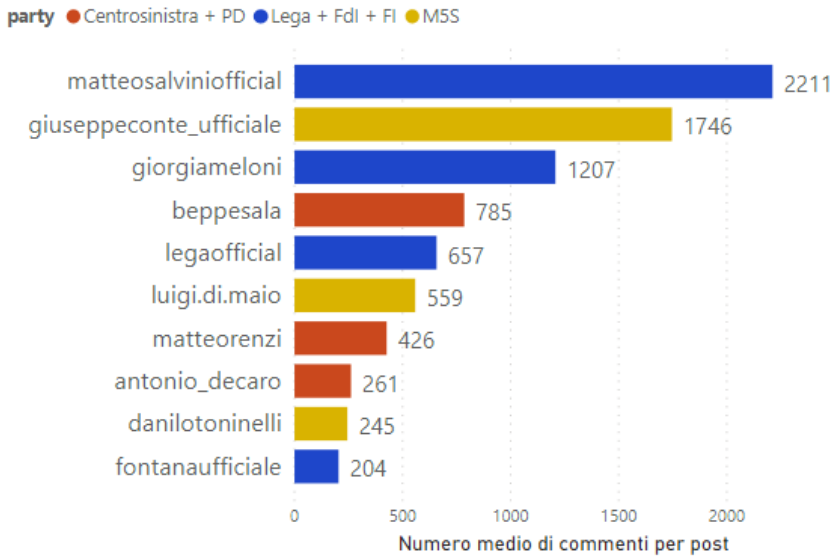

Figura 44 – Top 10 profili per numero medio di commenti per post su Instagram

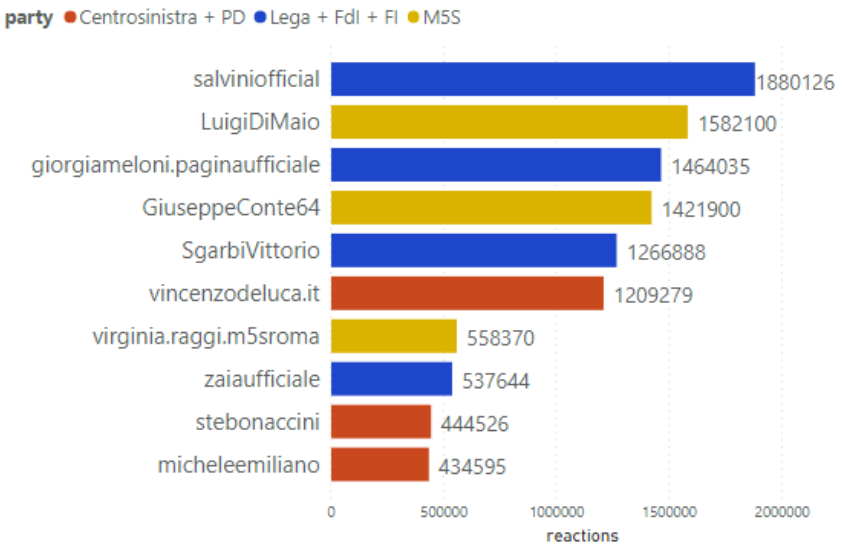

Figura 45 – Top 10 profili per numero di reazioni su Facebook

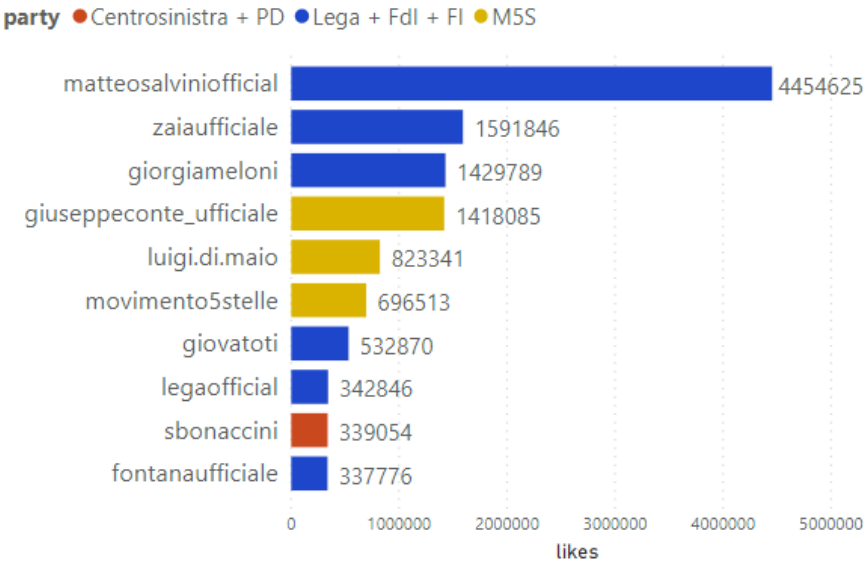

Figura 46 – Top 10 profili per numero di likes su Instagram

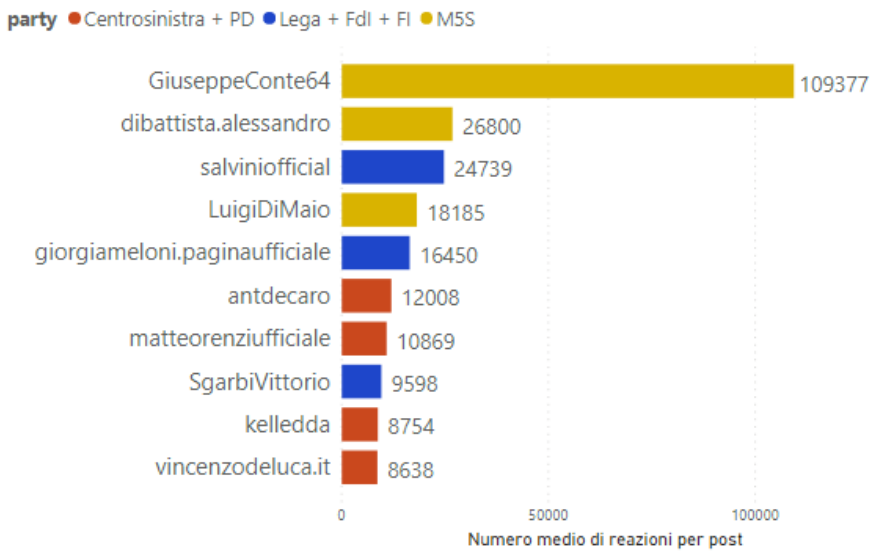

Figura 47 – Top 10 profili per numero medio di reazioni per post su Facebook

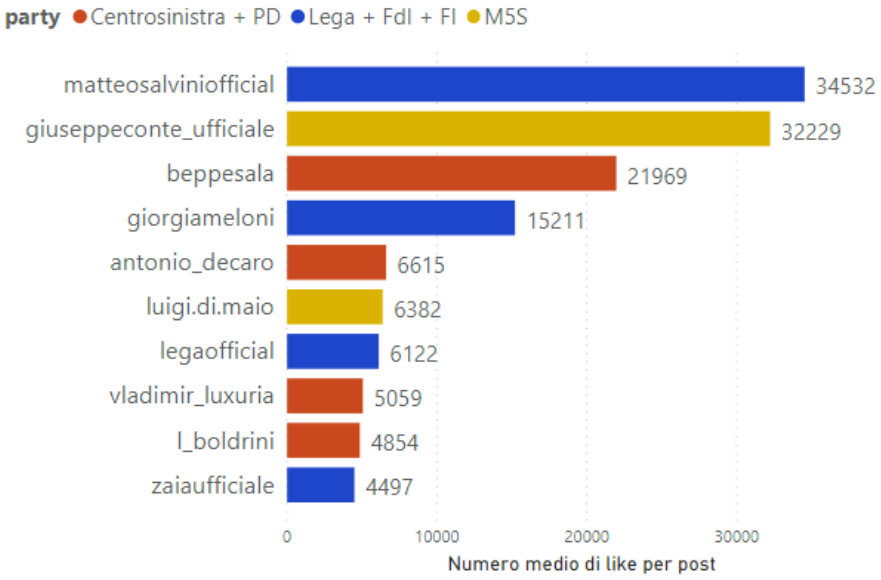

Figura 48 – Top 10 profili per numero medio di likes per post su Instagram

È interessante notare, nelle figure sottostanti, che ci sia un vero e proprio distacco tra le personalità in termini di persone coinvolte nelle discussioni. Su Facebook tale fenomeno è evidenziato, si può notare per esempio che Giuseppe Conte, Presidente del Consiglio in carica, ha coinvolto in media 9,5 mila utenti per post; segue Luigi Di Maio con circa 3,4 mila commentatori. Su Instagram vi è una distribuzione simile per numero di commentatori tra i due maggiori esponenti del Centrodestra e per Giuseppe Conte.

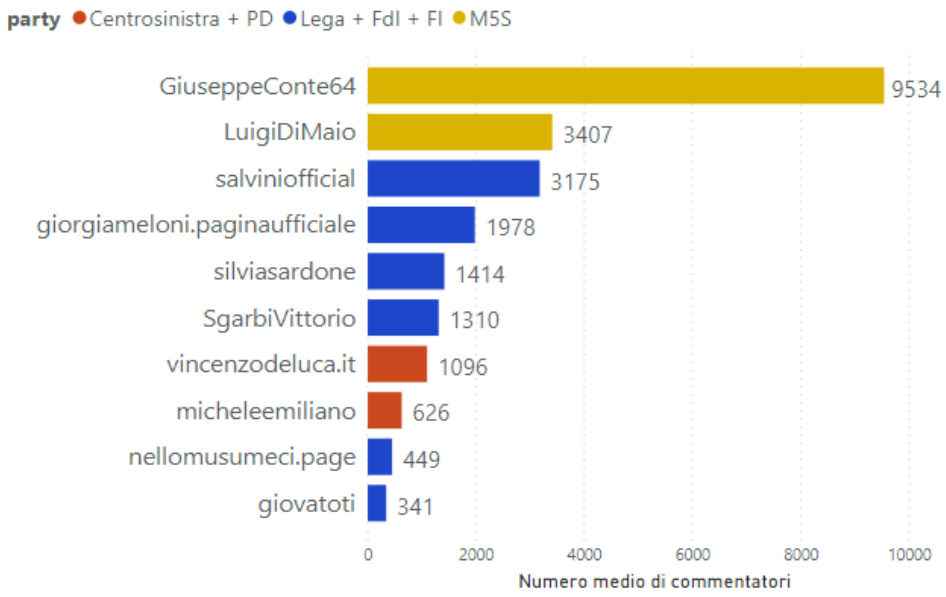

Figura 49 – Top 10 profili per numero medio di commentatori su Facebook

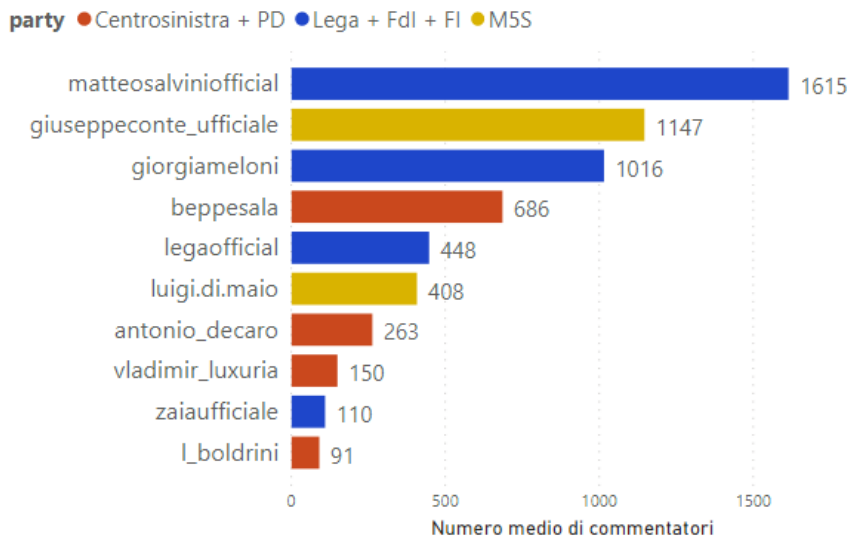

Figura 50 – Top 10 profili per numero medio di commentatori su Instagram

Nelle seguenti figure sono state raccolte le metriche normalizzate per numero di follower che confermano quanto detto in precedenza: i profili di centrosinistra producono più engagement a parità di numero di follower. Si può notare che su Facebook, sia per quanto riguarda i commenti sia le reazioni, spicca il nome di Elisabetta Gardini, conduttrice televisiva e facente parte del partito di Fratelli d'Italia. Su Instagram, invece, il politico che ha ottenuto un maggior numero di commenti per mille follower è Roberto Speranza,

Ministro della Speranza, molto attivo nelle comunicazioni sulle direttive per il contrasto dell'epidemia e quindi, come si può immaginare, anche soggetto a numerose discussioni al di sotto dei suoi post. Giuseppe Sala, Giorgio Gori (sindaco di Bergamo) e Antonio De Caro sono le tre personalità che presentano un maggior numero di likes ricevuti per mille follower su Instagram, tutti e tre facenti parte della fazione politica del Centrosinistra.

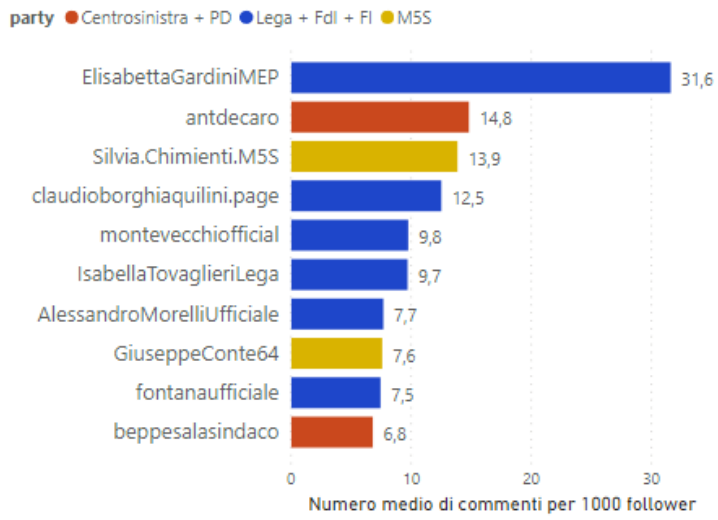

Figura 51 – Top 10 profili per numero medio di commenti per 1000 follower Facebook

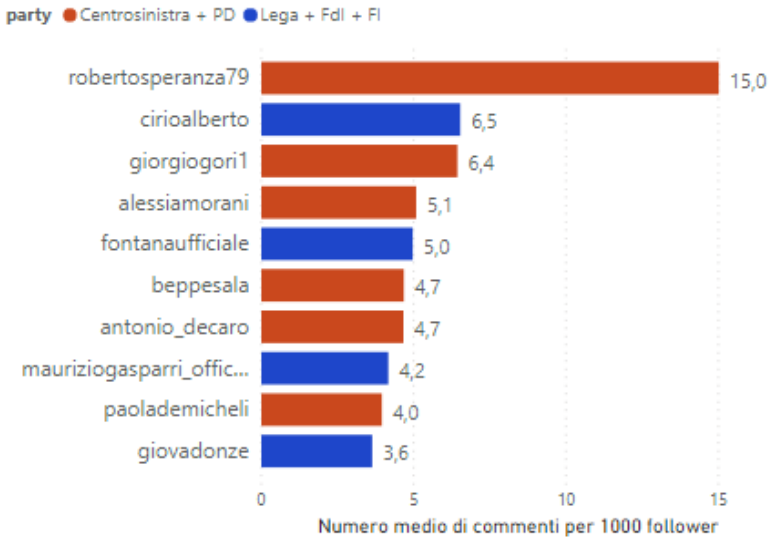

Figura 52 – Top 10 profili per numero medio di commenti per 1000 follower Instagram

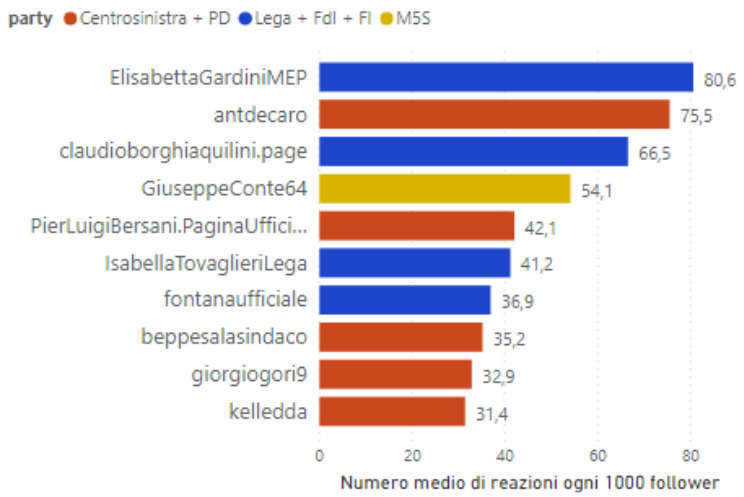

Figura 53 – Top 10 profili numero medio di reazioni ogni 1000 follower su Facebook

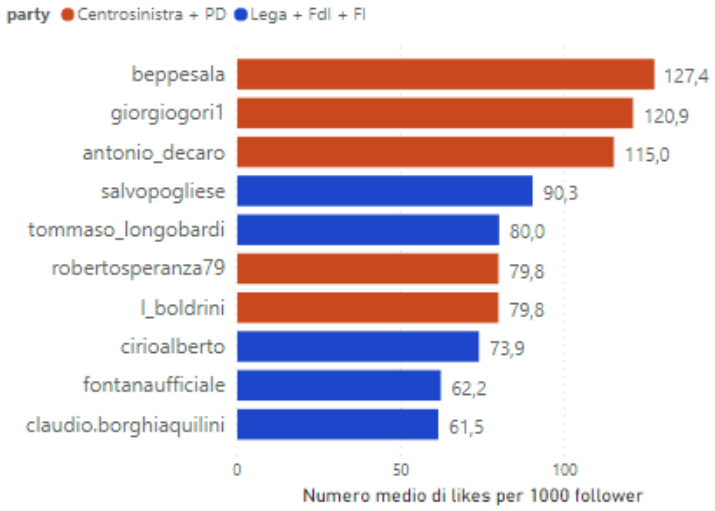

Figura 54 – Top 10 profili per numero medio di likes ogni 1000 follower su Instagram

4.1.2.2. Misure di interazione per partito politico

Nei seguenti grafici è possibile evincere l'andamento temporale dell'interazione social delle tre categorie durante i primi mesi di pandemia.

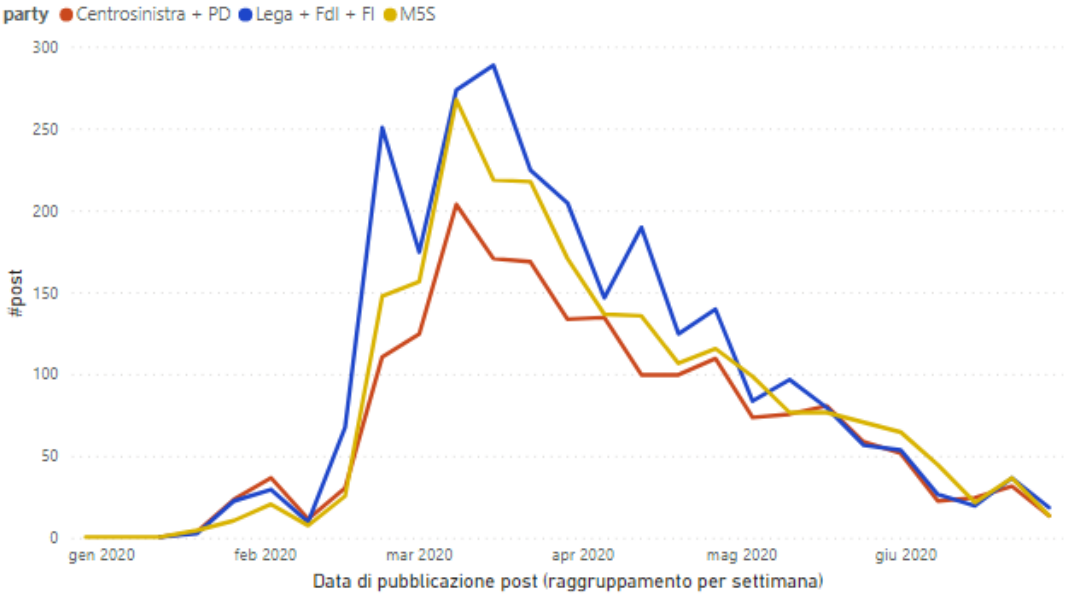

Figura 55 – Andamento temporale di pubblicazione post per fazione politica su Facebook

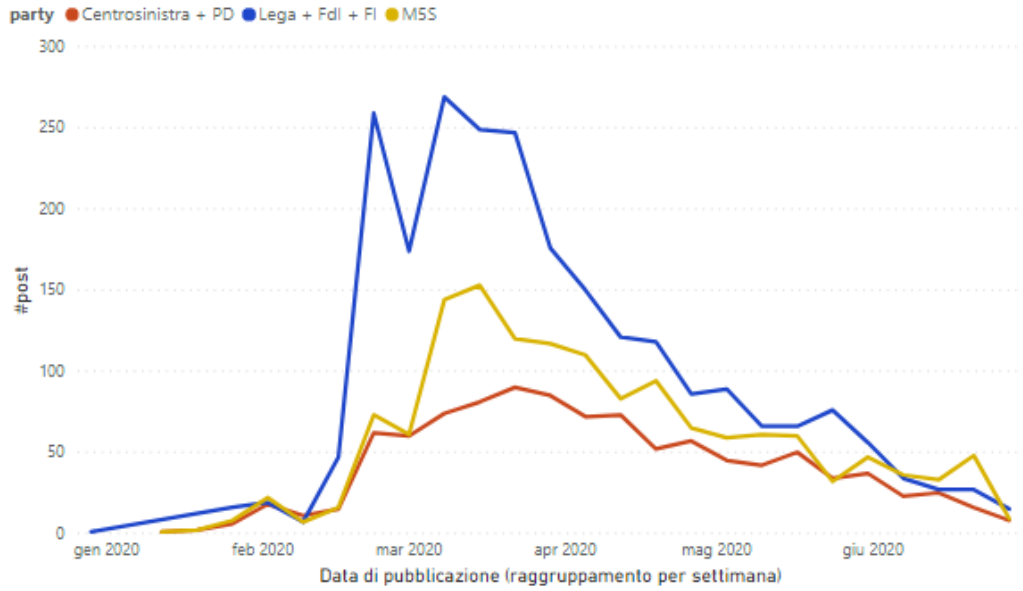

Figura 56 – Andamento temporale di pubblicazione post per fazione politica su Instagram

I dati relativi confermano quanto detto precedentemente. La categoria appartenente all'aggregazione dei partiti di centrodestra, nonostante il numero inferiore di profili, è quella che ha creato più post con argomento Coronavirus. Come si può notare infatti dal

seguente grafico a bolle, è il partito con un numero medio di post creati maggiore rispetto agli altri partiti su entrambi i social. La categoria "Lega+FdI+FI" è anche quella che riceve maggior risposta da parte del popolo media. Infatti, si può notare che i profili politici appartenenti a tale categoria ricevono una media di circa 600 commenti per post pubblicato su Facebook e 250 su Instagram. Il centrosinistra risulta essere invece la categoria meno seguita ricevendo una media di 300 commenti su Facebook e circa 70 su Instagram.

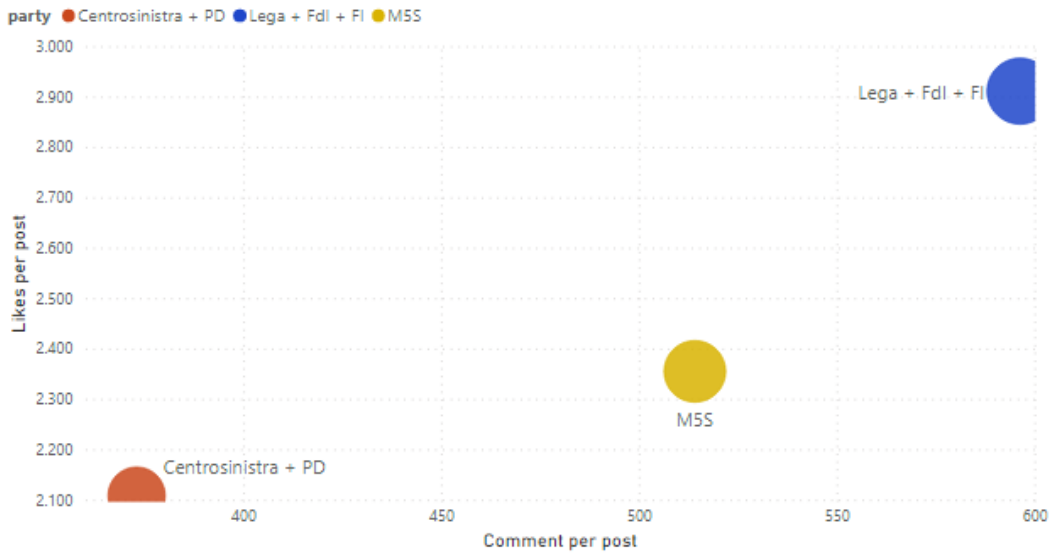

Figura 57 – Engagement per fazione politica su Facebook

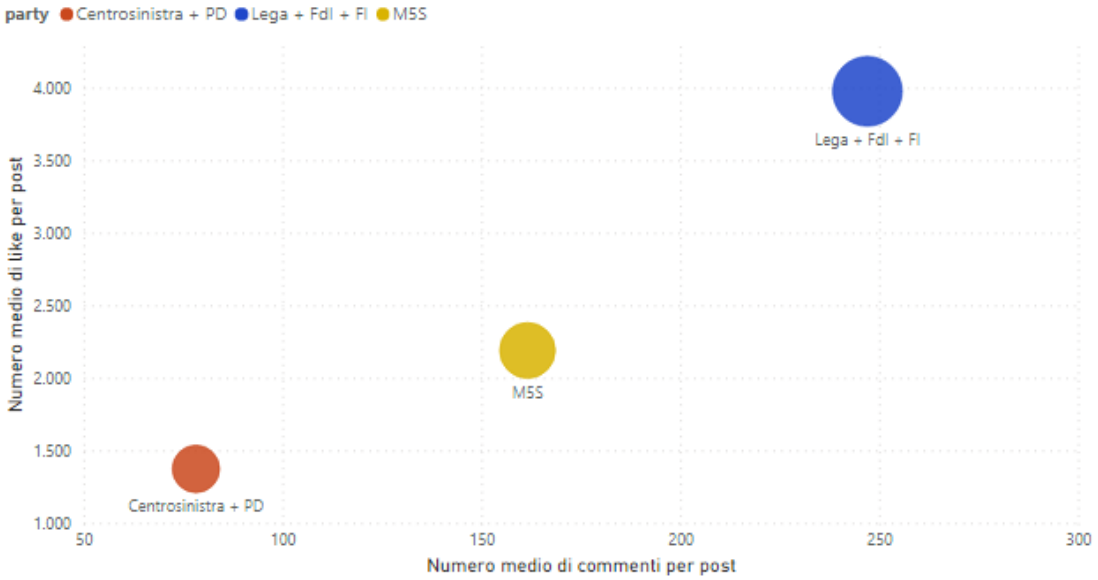

Figura 58 – Engagement per fazione politica su Instagram

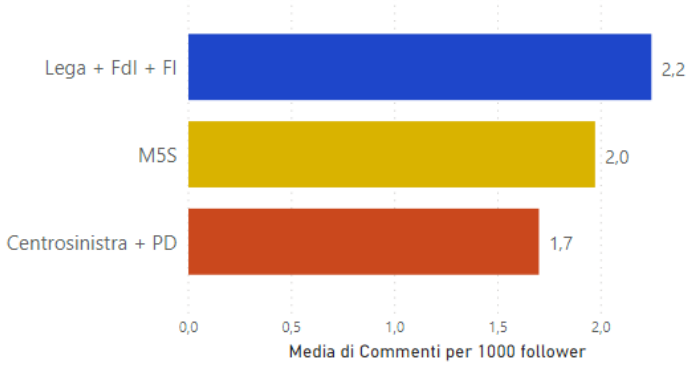

Figura 59 – Numero medio di commenti per 1000 follower per fazione politica su Facebook

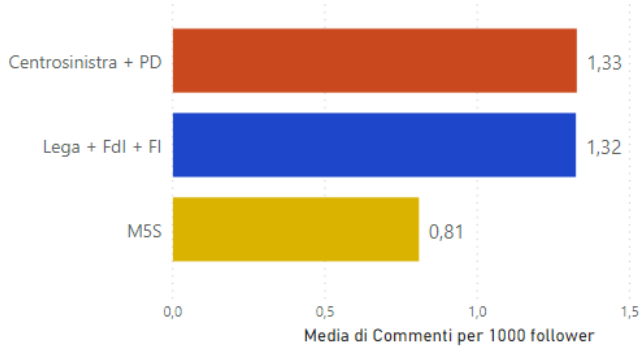

Figura 60 -– Numero medio di commenti per 1000 follower per fazione politica su Instagram

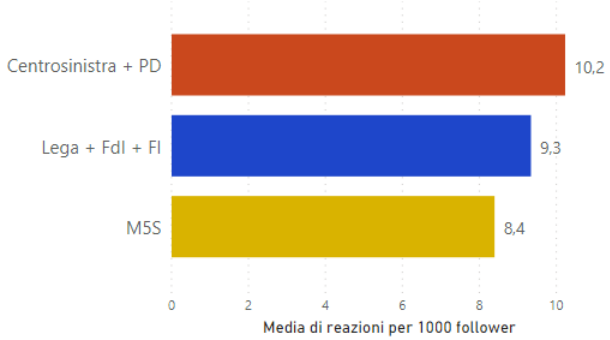

Figura 61 - – Numero medio di reazioni per 1000 follower per fazione politica su Facebook

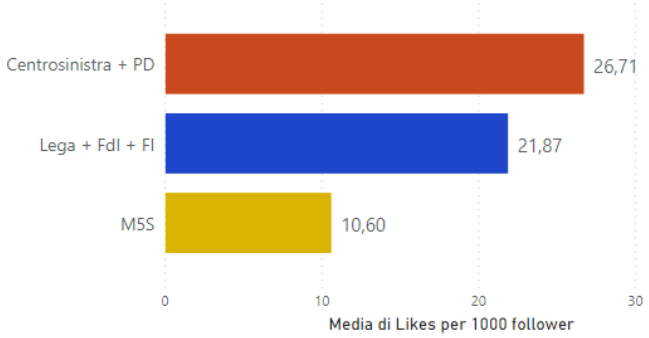

Figura 62 - – Numero medio di likes per 1000 follower per fazione politica su Instagram

4.2. Analisi dei contenuti potenzialmente "fake"

Come argomentato nel capitolo 3, in questa fase è stata effettuata una ricerca di un set di keywords all'interno dei commenti pubblicati dagli utenti sotto i post pubblicati dai profili politici. Nella seguente tabella è mostrato il numero di commenti riscontrati sulle due piattaforme, il numero di utenti distinti che hanno scritto i commenti contenenti tali parole e il numero di post sotto cui questi commenti sono presenti:

	Commenti	**Username commenti**	**Post**
Facebook	31 217	24 991	3 348
Instagram	5 790	4 334	1 443

Tabella 11 – Caratteristiche commenti riferiti ad argomenti potenzialmente "fake"

È senz'altro possibile individuare una diffusione decisamente più marcata di tali topic sul social di Facebook rispetto a quello di Instagram.

Di seguito è mostrato l'andamento temporale della pubblicazione di tali commenti normalizzato sulla base del numero totale dei commenti filtrati. Ciò è stato svolto in modo tale da confrontare il comportamento delle due curve rappresentati i due Social. Nel trend chart si può individuare il primo e più marcato picco nella seconda metà di marzo che vede coinvolta in maniera più significativa la piattaforma di Instagram. Facebook, d'altro canto, vede la presenza di tre picchi nei mesi di marzo e aprile.

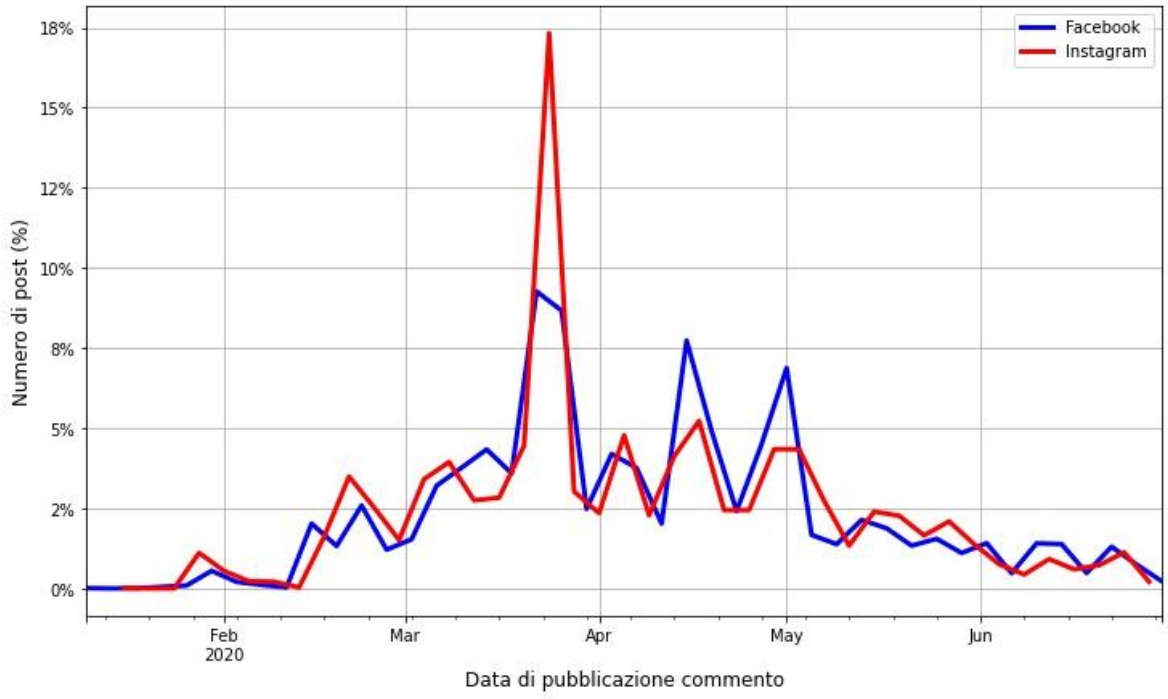

Figura 63 – Andamento di pubblicazione commenti con argomenti "fake"

Nei seguenti diagrammi a torta è stato mostrato il calcolo percentuale della distribuzione dei commenti contenenti parole riferite ai topic potenzialmente "fake" sotto i post dei profili appartenenti alle diverse fazioni politiche. In entrambi i social la porzione più significativa corrisponde alla cateogoria "Lega + FdI + FI", su Instagram tale percentuale corrisponde a più della metà dei commenti. Segue la fazione del Movimento 5 Stelle e in percentuale molto minore, soprattutto sul social di Instagram dove è stata notata in generale una minore interazione con la coalizione, i commenti contenenti tematiche potenzialmente "fake" rivolte ai post pubblicati dal Centrosinistra, con appena il 7,7% del totale dei commenti.

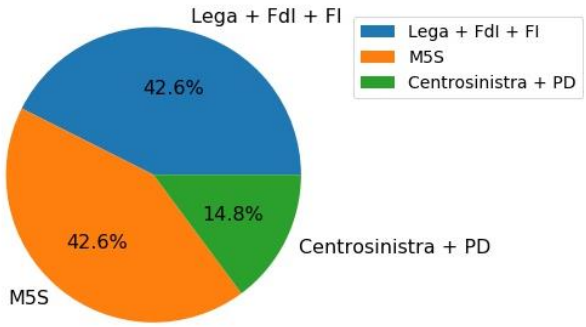

Figura 64 - Distribuzione di commenti con argomento "fake" per fazione politica su Facebook

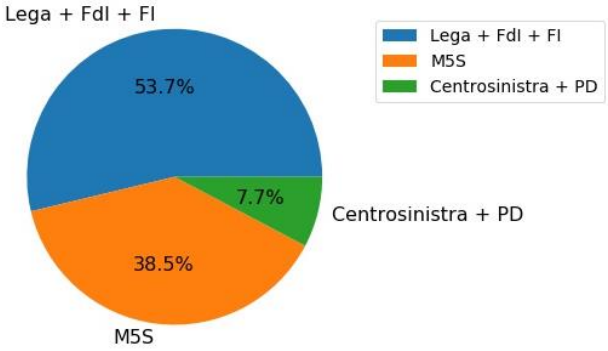

Figura 65 - Distribuzione di commenti con argomento "fake" per fazione politica su Instagram

Nel diagramma a barre sottostante è stato fatto un focus sui profili politici che hanno ottenuto un numero maggiore di commenti con argomenti potenzialmente "fake". Si può notare che su Facebook il profilo di Luigi Di Maio sembra prevalere su tutti gli altri ricevendo più del 20% dei commenti contenenti le parole filtrate in partenza. Seguono i profili di Matteo Salvini, Giorgia Meloni, Sgarbi Vittorio e Giuseppe Conte con percentuali molto minori. Su Instagram il profilo che ha ottenuto il numero maggiore dei commenti contenenti tali tematiche è quello di Matteo Salvini; Luigi Di Maio si trova in terza posizione subito dopo il profilo del partito pentastellato che, come è stato più volte citato, si conferma essere il profilo prevalente su tale Social in termini di attività e di interazioni.

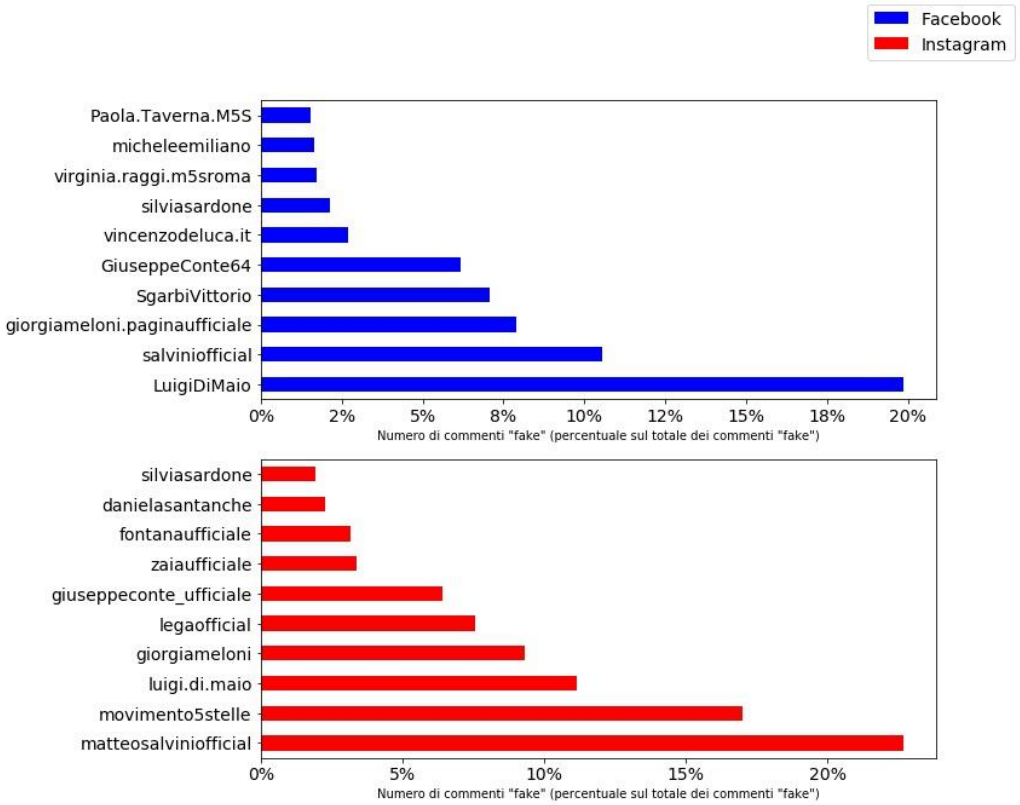

Figura 66 – Top 10 profili per maggior numero di commenti con argomenti "fake"

Nei successivi sottoparagrafi sarà descritto l'andamento di tre specifiche tematiche denominate "potenzialmente fake" che si sono manifestate sotto forma di specifiche keywords all'interno dei commenti sotto i post pubblicati dai politici sul tema "Coronavirus". In particolare, nella seguente tabella è stato raccolto il numero di commenti in cui le parole riferite a tali tematiche si sono presentate:

	Commenti Facebook	**Commenti Instagram**
5g	2 414	231
Laboratorio	4 958	737
Bill Gates	5 368	458

Tabella 12 – Commenti riferiti alle tre tematiche potenzialmente "fake" su Facebook e Instagram

4.2.1. Fake news: Coronavirus – Tecnologia 5g

In questa sezione è stato fatto un focus sulla Fake News che ha coinvolto una correlazione tra la tecnologia 5g e la pandemia da Coronavirus. Nel seguente grafico è possibile osservare l'andamento dei commenti contenenti le parole riferite alla tematica del 5g sotto

i post riferiti al Coronavirus nel tempo in entrambe le piattaforme. I dati sono normalizzati rispetto al totale commenti in tabella 11 per poter effettuare un confronto tra i Social. Le due curve hanno comportamenti molto simili, ma è evidente la presenza di un maggiore picco di commenti su Facebook all'inizio del mese di aprile. In entrambi i casi, le curve iniziano a crescere in maniera vertiginosa poco dopo la diffusione della notizia.

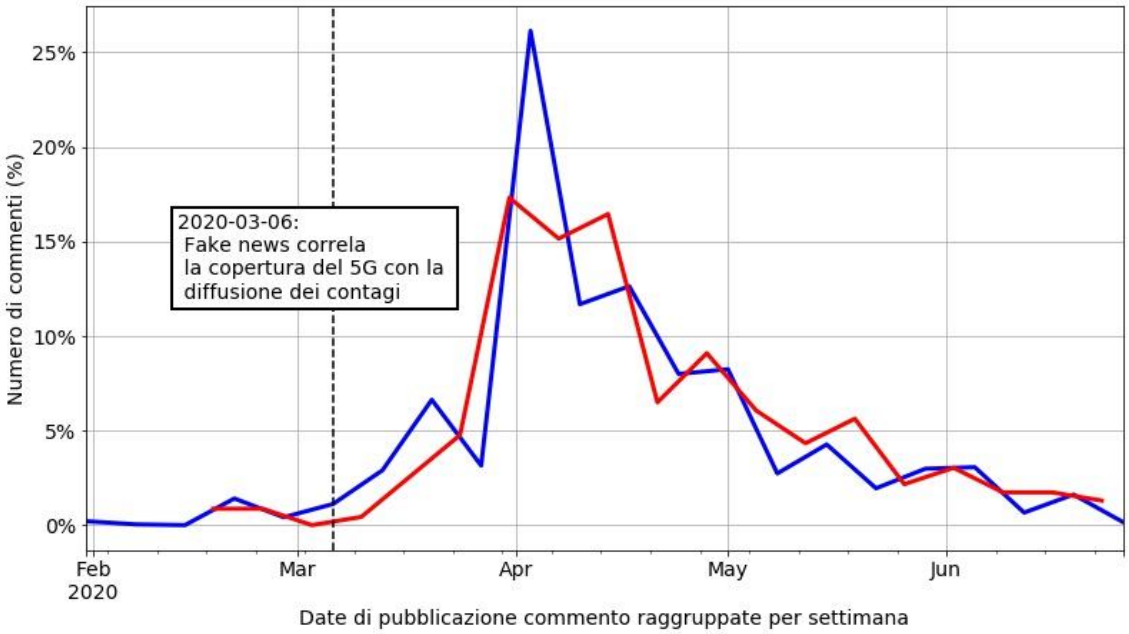

Figura 67 – Andamento pubblicazione commenti con argomento "5g" nel tempo: confronto Facebook vs Instagram

Il partito che su Facebook ha ricevuto più commenti riguardanti tale tematica è il Movimento 5 Stelle. Infatti, più del 60% dei commenti sono stati pubblicati sotto il post di tale schieramento. Segue la coalizione del Centrodestra con il 30% circa. Su Instagram vi è un distacco molto significativo tra le tue categorie appena citate e la fazione del Centrosinistra, che ha ricevuto una percentuale di commenti inferiore al 10%.

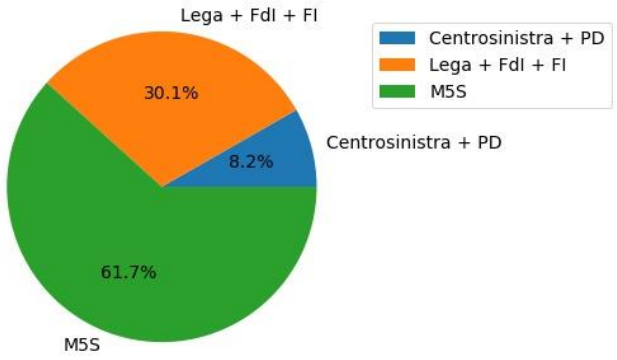

Figura 68 – Distribuzione di commenti con riferimenti al "5g" per fazione politica su Facebook

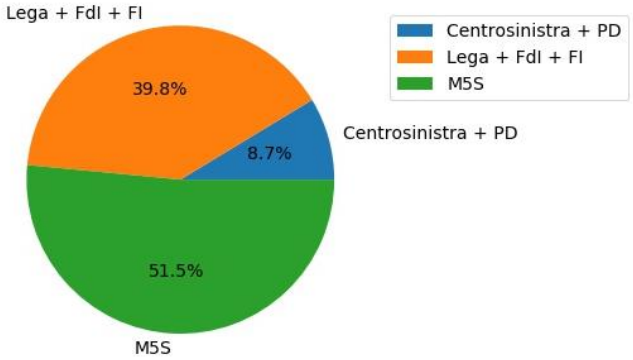

Figura 69 - Distribuzione di commenti con riferimenti al "5g" per fazione politica su Instagram

Nelle seguenti figure si possono osservare gli andamenti temporali della pubblicazione dei commenti raggruppate per coalizione politica.

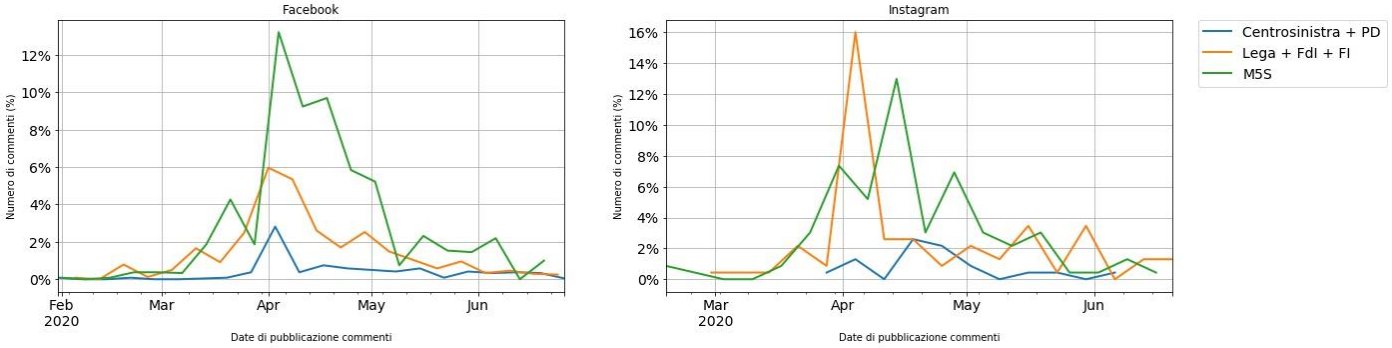

Figura 70 – Andamento di pubblicazione commenti con riferimenti al "5g" per fazione politica

Per quanto riguarda il focus sui singoli profili, è interessante notare come su Facebook il profilo di Luigi Di Maio compaia in maniera preponderante rispetto al resto dei profili: quasi il 35% dei commenti riferiti all'argomento "5g" sotto post che citano parole relative al Covid-19 sono stati pubblicati sotto i suoi post. Il suo nome è il primo della top 10 anche su Instagram, dove però il distacco con gli altri politici è meno accentuato.

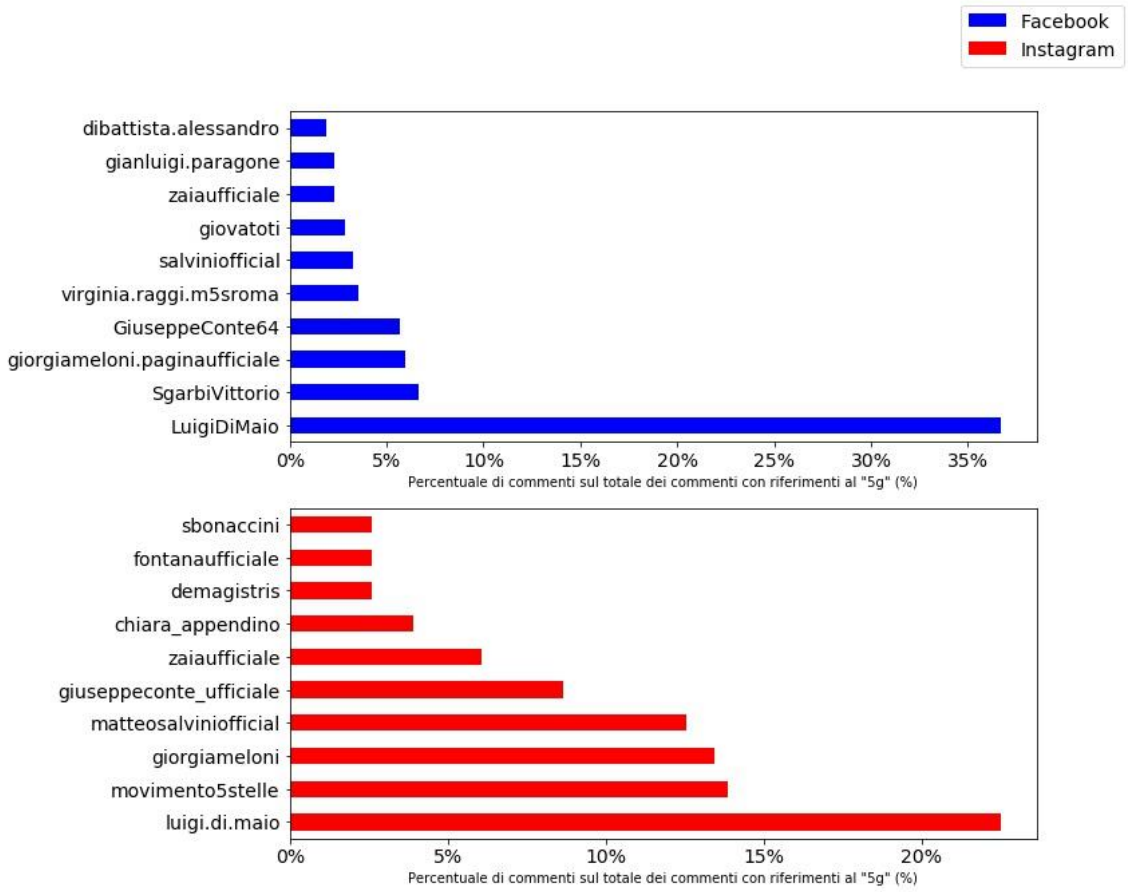

Figura 71 – Top 10 profili per numero di commenti con riferimenti al "5g" ricevuti

4.2.2. Fake news: Coronavirus creato in laboratorio

Filtrando i commenti pubblicati sotto i post sul Coronavirus contenenti le parole "laboratorio" e "tgr leonardo" è stato possibile osservare l'andamento temporale di condivisione di tali contenuti e i profili sotto cui sono stati maggiormente pubblicati. Nella seguente figura si nota come il comportamento delle due curve, la blu riferita a Facebook e la rossa riferita a Instagram, sia lo stesso. Il picco coincide perfettamente con la metà di marzo, pochi giorni dopo la diffusione di tale notizia su Internet. È interessante osservare come il ciclo di vita della discussione di tale tematica su entrambe le piattaforme sia stato molto breve, infatti dalla fine di marzo entrambe le curve si appiattiscono.

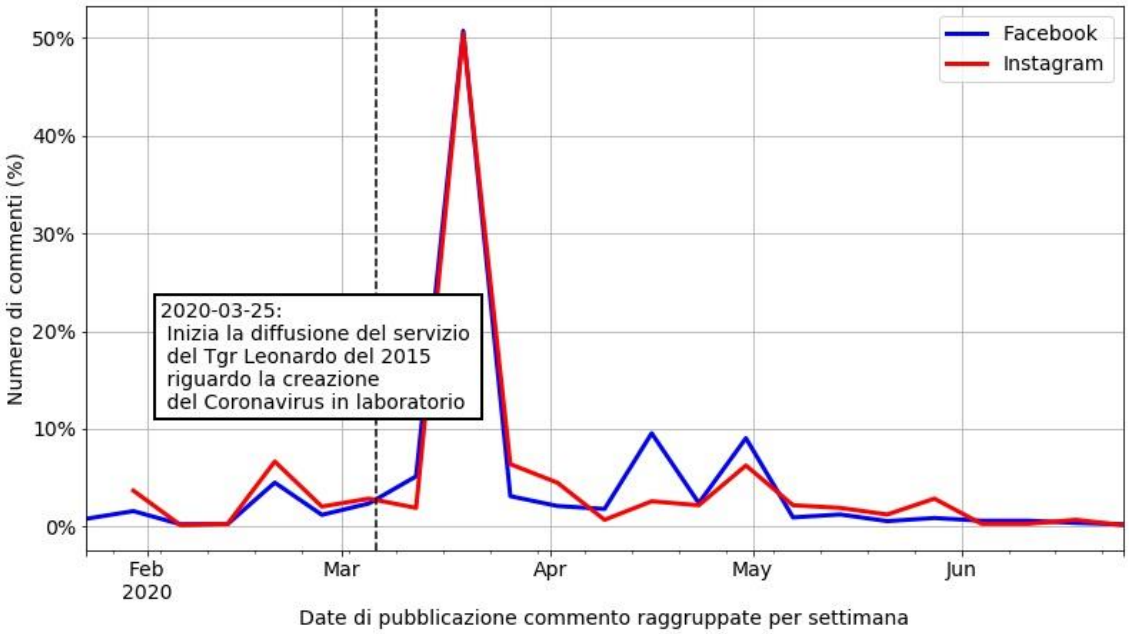

Figura 72 – Andamento pubblicazione commenti riferiti alla creazione del Coronavirus in laboratorio: confronto Facebook vs Instagram

In entrambi i Social Network la categoria politica che ha ricevuto più commenti relativi a tale tematica, come si può ben notare dai seguenti diagrammi a torta, è quella che aggrega i partiti di Lega Nord, Fratelli d'Italia e Forza Italia. Il Centrosinistra, in entrambe le piattaforme ha ricevuto meno del 10% dei commenti sul totale.

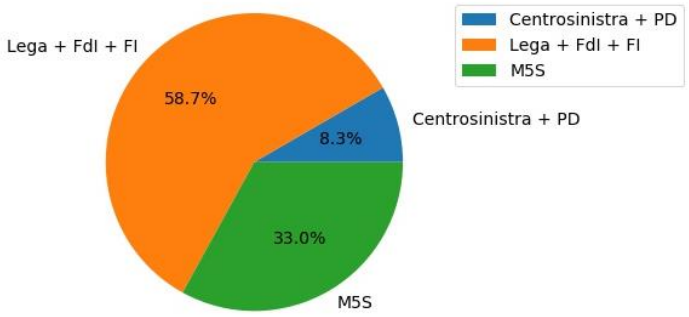

Figura 73 – Distribuzione di commenti con riferimenti alla creazione del Coronavirus in laboratorio per fazione politica su Facebook

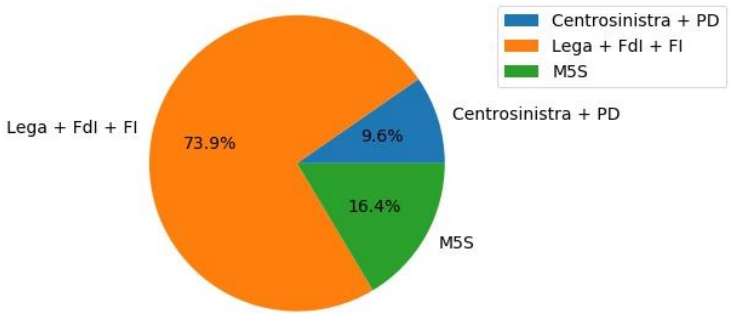

Figura 74 - Distribuzione di commenti con riferimenti alla creazione del Coronavirus in laboratorio per fazione politica su Instagram

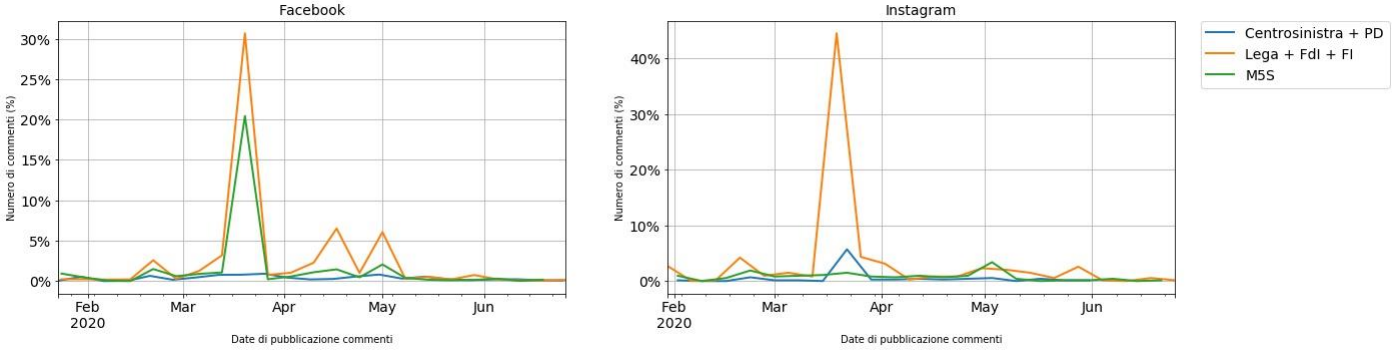

Figura 75 – Andamento pubblicazione commenti con riferimenti alla creazione del Coronavirus in laboratorio per fazione politica

I profili sotto i cui post sono stati pubblicati la maggior parte dei commenti relativi a tale argomento sono, su Facebook, Luigi Di Maio e Matteo Salvini e con una percentuale poco più bassa (12%) Giorgia Meloni. Su Instagram tale argomento sembra essere stato discusso prevalentemente sotto i post di Matteo Salvini con il 45%.

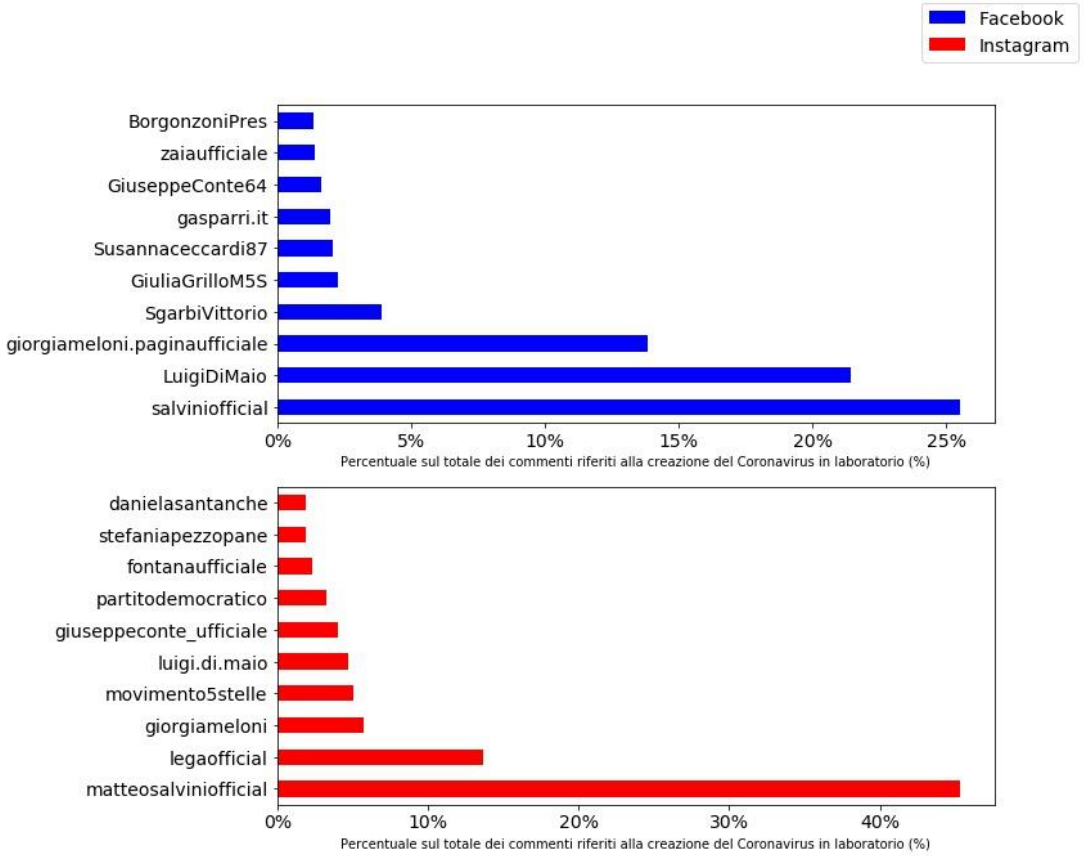

Figura 76 – Top 10 profili per percentuale di commenti con riferimenti alla creazione del Coronavirus in laboratorio ricevuti

4.2.3. Fake News: Covid-19 e Bill Gates

Come introdotto nel Capitolo 1, molte sono state le Fake News che hanno visto come protagonista il fondatore di Microsoft Corporation Bill Gates. È il motivo per cui, in questo trend chart non è stata inserita l'etichetta relativa ad una ipotetica data di inizio della diffusione della notizia. L'andamento delle curve che rappresentano la pubblicazione di commenti con riferimenti a Bill Gates conferma questo fenomeno. Diversamente dalle altre tematiche, infatti, le curve non presentano un andamento regolare, ma mostrano vari picchi durante il periodo relativo alla prima ondata. Ciò vuol dire che il nome di Bill Gates riferito al Coronavirus è stato citato in tanti commenti in diversi momenti della pandemia.

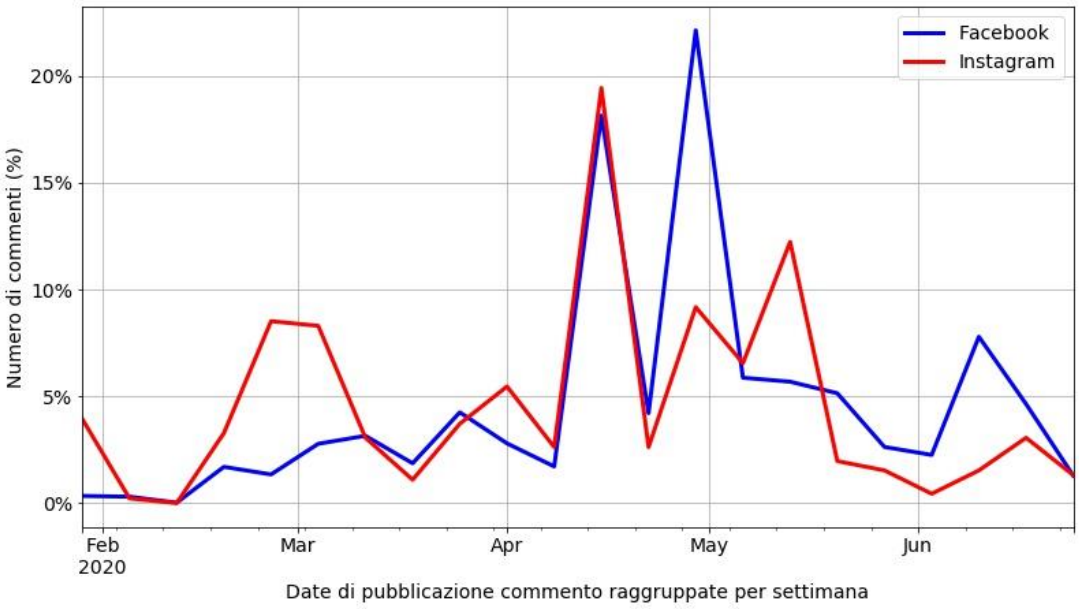

Figura 77 – Andamnto pubblicazione commenti con riferimenti a Bill Gates: confronto Facebook vs Instagram

Diversamente da quanto osservato nelle precedenti sezioni, il partito politico che ha ottenuto maggiori commenti relativi a tale topic su Facebook è quello del Movimento 5 Stelle. Su Instagram la percentuale dei commenti sotto i post pentastellati risulta molto simile a quello della fazione del Centrodestra. I partiti del Centrosinistra e il Partito Democratico si confermano anche in questo terzo caso politici sotto i cui post queste tematiche sono state poco affrontate.

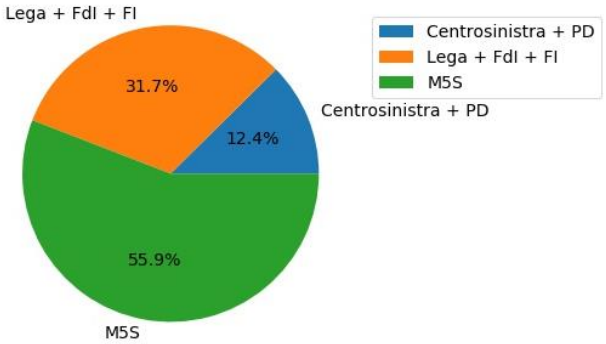

Figura 78 – Distribuzione dei commenti con riferimenti a Bill Gates per fazione politica Facebook

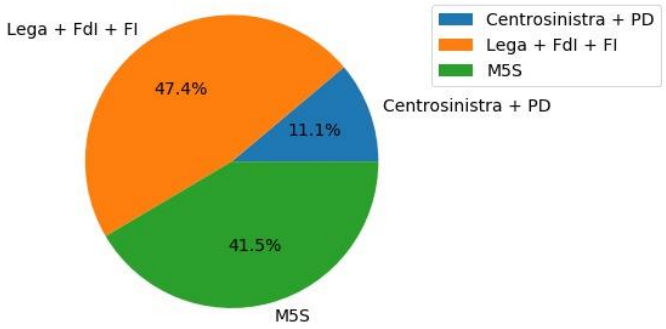

Figura 79 - – Distribuzione dei commenti con riferimenti a Bill Gates per fazione politica su Instagram

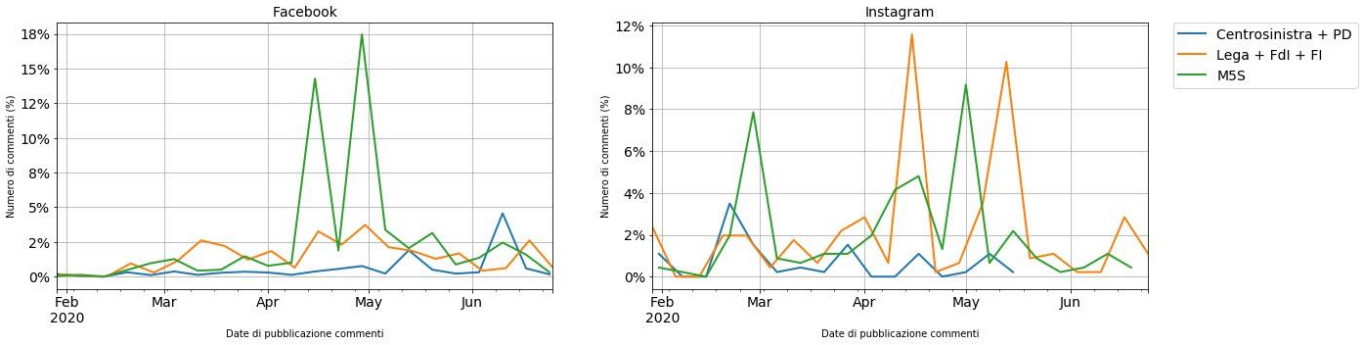

Figura 80 – Andamento pubblicazione commenti con riferimenti a Bill Gates per fazione politica

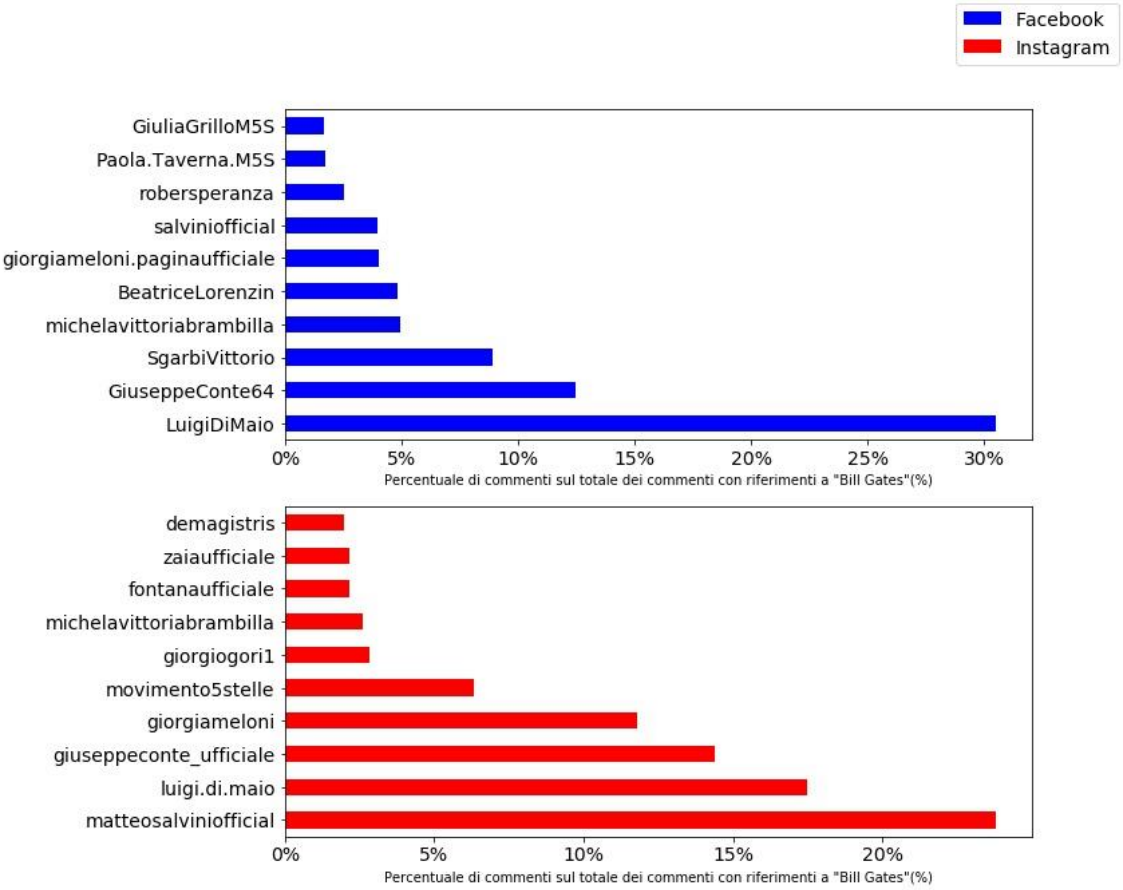

Figura 81 – Top 10 profili per commenti con riferimenti a Bill Gates ricevuti

Il confronto dell'andamento della comparsa di queste tre tematiche all'interno dei commenti dei due Social Network è rappresentato nelle figure 82 e 83 in cui i valori sono normalizzati per il totale dei commenti contenenti tali parole. In entrambe le piattaforme la curva che ha per prima avuto un elevato picco è quella relativa alla Fake News riguardante la creazione del Coronavirus in laboratorio.

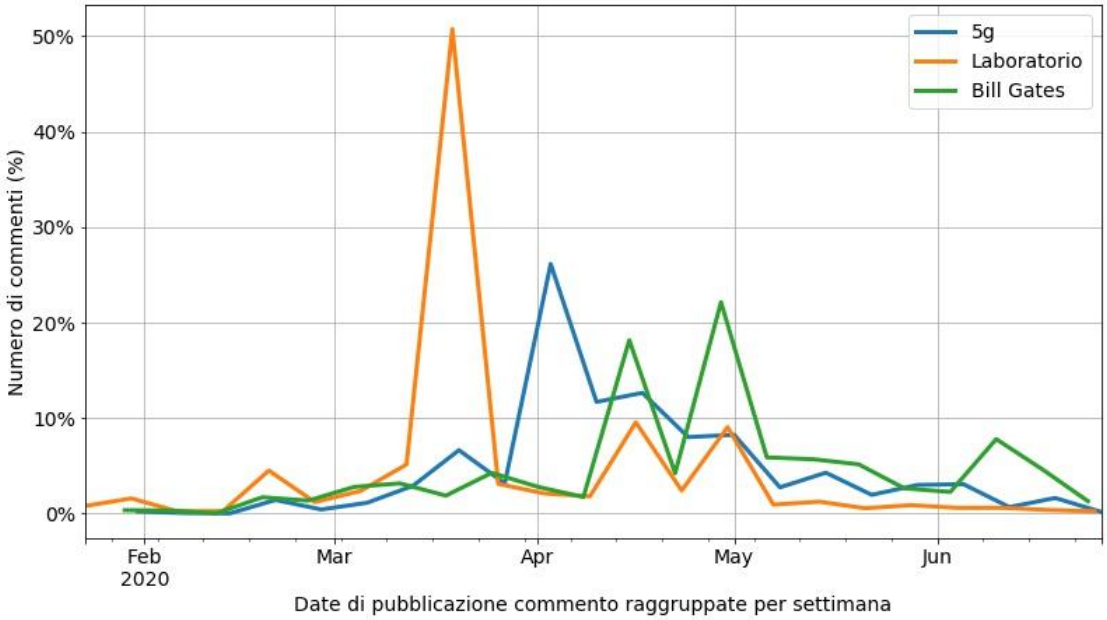

Figura 82 – Andamento pubblicazione commenti con riferimenti ai tre topic "fake" Facebook

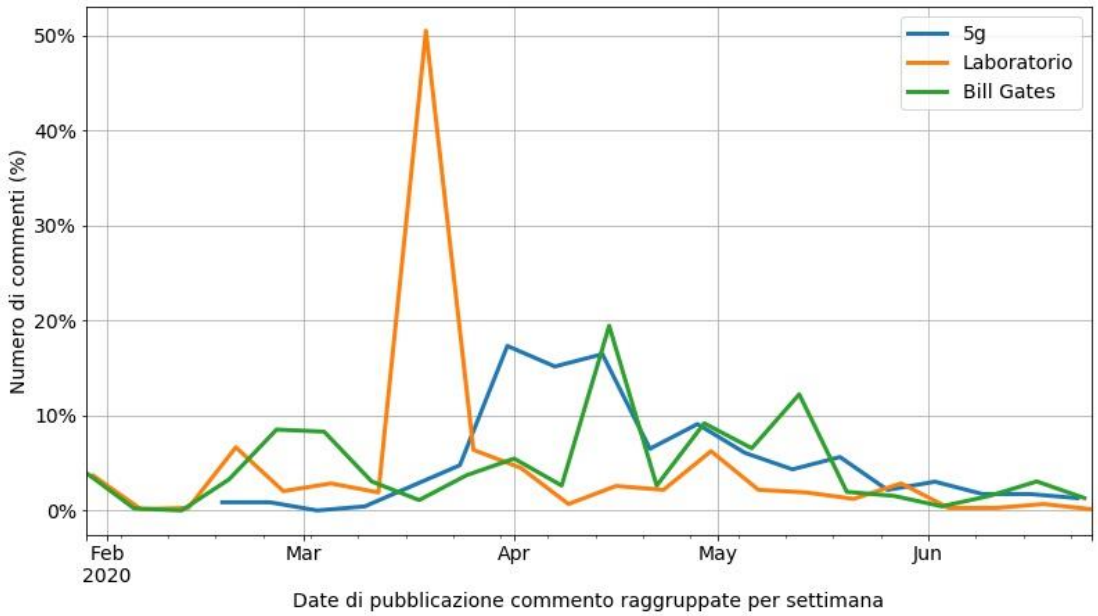

Figura 83 - Andamento pubblicazione commenti con riferimenti ai tre topic "fake" Instagram

4.3. Analisi di caratterizzazione del contenuto testuale

Nella seguente sezione saranno presentati i risultati dell'applicazione del modello Latent Dirichlet Allocation sul set di commenti preso in considerazione e descritti nel paragrafo 4.2. Successivamente saranno mostrati i risultati dell'applicazione dello stesso modello sui post sotto cui questi commenti sono stati pubblicati.

4.3.1. Applicazione del modello sui commenti

Inizialmente, per ciascun documento compreso nel *corpus* in input del Latent Dirichlet Allocation, sono state conteggiate i caratteri di cui è composto. Ciò è stato svolto separatamente per i due Social Network in modo da creare, come è possibile notare nelle figure 84 e 85, due distribuzioni distinte. I commenti pubblicati su Facebook osservano una distribuzione più spostata verso destra rispetto a quella dei commenti di Instagram, indice del fatto che su tale piattaforma si tende ad essere coinvolti in discussioni più argomentate, ma anche che le limitazioni del Social Network più "multimediale" sono più restrittive: il limite massimo dei caratteri delle didascalie testuali e dei commenti è di 2.200, mentre su Facebook tale restrizione è di 63.206 caratteri.

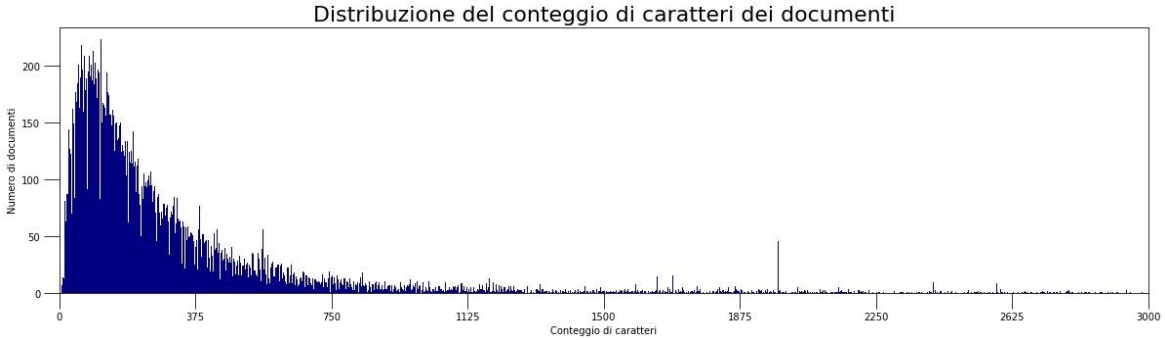

Figura 84 – Distribuzione del conteggio di caratteri dei commenti Facebook

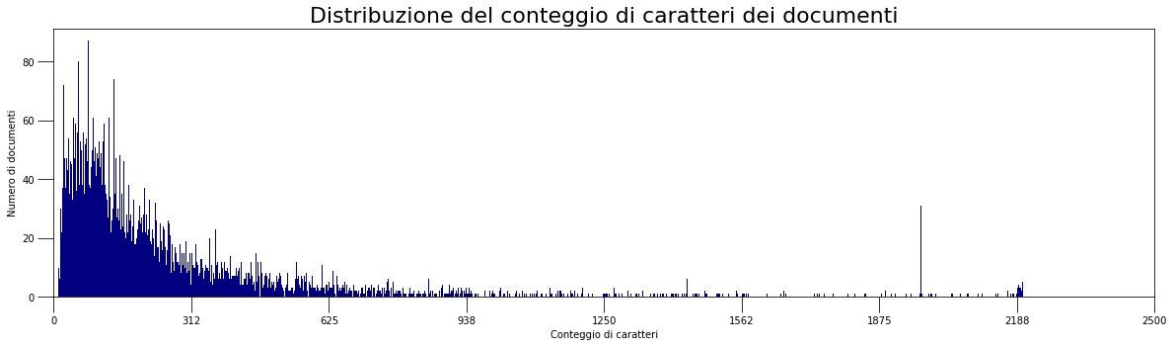

Figura 85 - Distribuzione del conteggio di caratteri dei commenti Instagram

L'obiettivo dell'applicazione del modello sui set di commenti ottenuti dal filtro delle argomentazioni potenzialmente "fake" è stato quello di confrontare i topic di cui ci è discusso nei due differenti Social Network. Per questo motivo, nonostante le dimensionalità dei dataset fossero diverse, è stato scelto di confrontare i contenuti testuali condivisi sulle due piattaforme, sulla base dello stesso numero di topic. In questo modo è stato possibile osservare come, a parità di topic, fossero distribuite le parole più frequenti all'interno dei commenti, pubblicati sui due social, dal modello. Il numero di topic scelto come parametro di input del modello è stato individuato effettuando servendosi sia della interpretazione infografica delle Word Cloud rappresentanti i vari topic sia le tecniche di visualizzazione t-SNE e LDAvis. Si è infatti scelto il numero di topic che ha permesso di individuare topic distinti sia da un punto di vista interpretativo, sia grafico come sarà illustrato successivamente. Per questa prima parte è stato scelto il numero di topic pari a 10. Per quanto riguarda gli altri parametri richiesti in input dal modello, sono state effettuate diverse prove e sono stati scelti i parametri che garantivano una performance ottimale tenendo conto del trade-off tra l'interpretabilità e il tempo computazionale necessario.

Dopo aver scelto il numero di topic, per visualizzare il contenuto degli argomenti è stata utilizzata la tecnica della Word Cloud che rappresenta i termini che, secondo il modello LDA, descrivono con alta probabilità gli argomenti. Le "nuvole" rappresentano le matrici topic-termini. Tale visualizzazione sottolinea i termini con le più alte probabilità con un carattere di dimensione più grande. In questo modo è possibile osservare subito se il modello ha prodotto dei risultati accettabili o meno. Conoscendo a priori gli eventi e le tematiche principali relative alla pandemia Covid-19, è stato possibile intuire gli argomenti di cui parlano i rispettivi topic. Nelle seguenti figure è possibile osservare 20 parole che descrivono ciascuno dei 10 topic estratti dal modello applicato sui commenti rispettivamente di Facebook e Instagram. Il peso che stabilisce la grandezza del carattere con cui ciascuna parola è rappresentata all'interno della nuvola corrisponde alla probabilità tale termine descriva il topic, secondo il Latent Dirichlet Allocation.

Figura 86 – Word Cloud commenti Facebook

Figura 87 – Word Cloud commenti Instagram

Nelle figure 88 e 89 sono mostrati, per ogni topic, i grafici a barre rappresentanti le prime dieci parole che presentano un "peso" maggiore per la descrizione del topic secondo l'LDA; valore che determina le Word Cloud sopra mostrate. Degli stessi termini è stata rappresentata sull'asse sinistro la loro frequenza all'interno dei documenti.

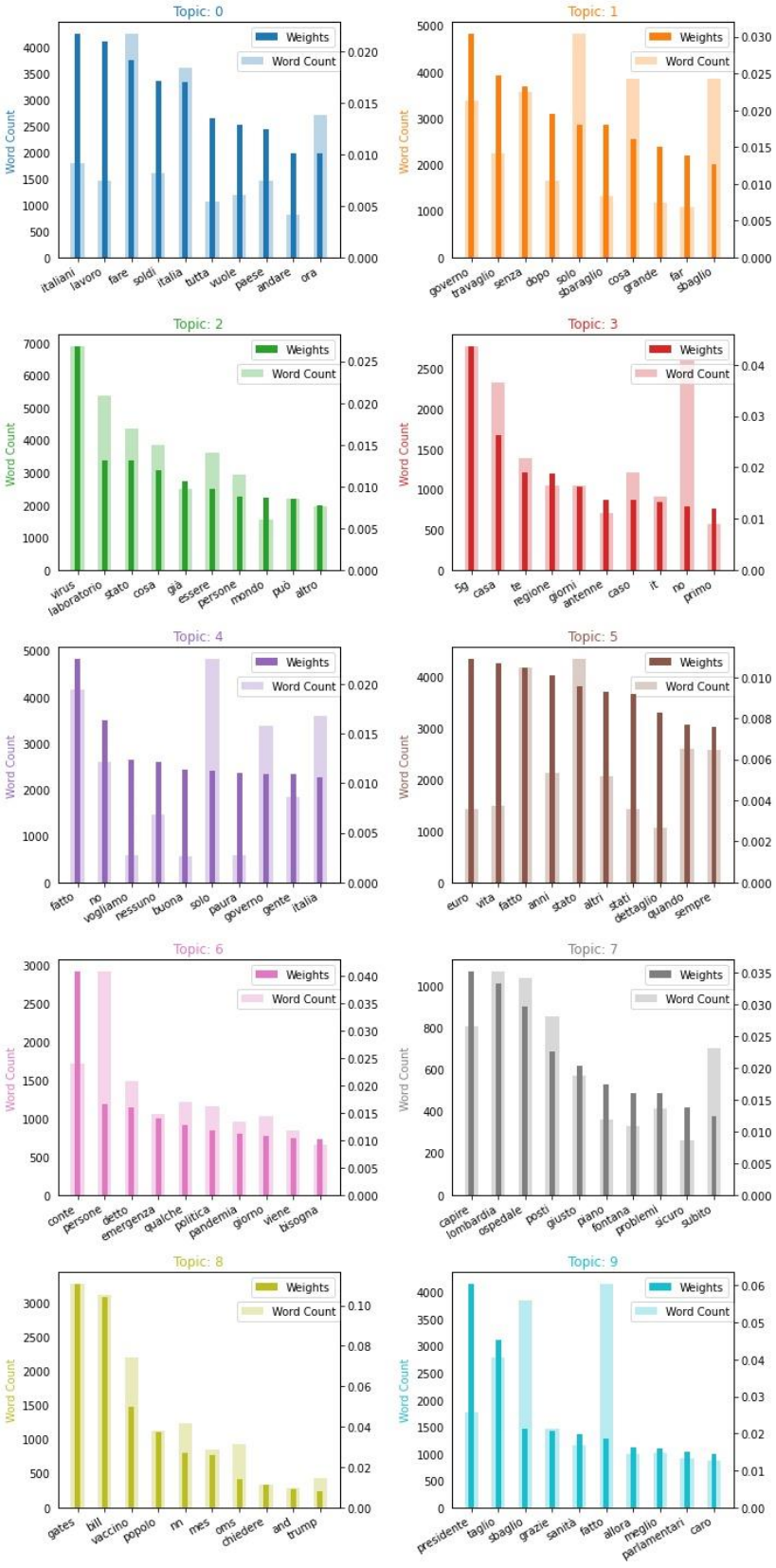

Figura 88 – Peso e frequenza delle parole per ciascun topic Facebook

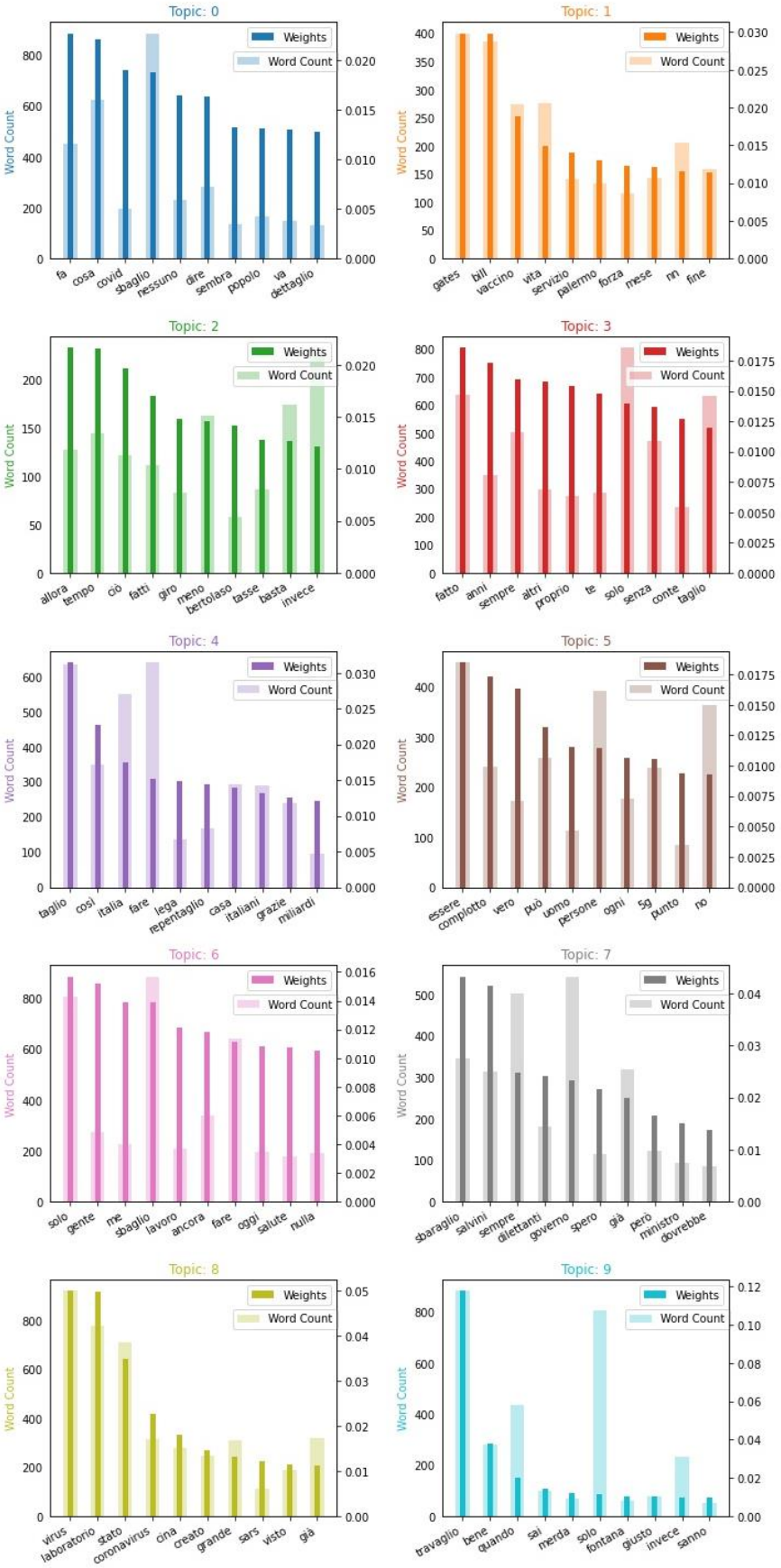

Figura 89 - Peso e frequenza delle parole per ciascun topic Instagram

4.3.1.1. Interpretazione dei topic dei commenti

Osservando le nuvole di parole rappresentate, si possono distinguere delle similarità tra alcuni topic e possono risaltare alcune tematiche particolari. Per quanto riguarda i commenti pubblicati su Facebook, filtrati secondo la lista di parole al 3.2.3, si può riscontrare che i topic che si riferiscono alle tematiche potenzialmente "fake" sono il topic 2, 3 e 8. Questi tre topic segnalano la presenza di un'alta frequenza di parole come, rispettivamente, "laboratorio", 5g", "bill gates", "complotto", che sembrano riferirsi ai tre topic manualmente individuati nella prima parte di analisi. Il topic 2 associa la parola "laboratorio" a parole come "cina" e "coronavirus", il topic 3 associa il termine "5g" a parole come "coronavirus", "antenne", " salute", il topic 4 viene descritto da parole come "complotto" e "paura"; infine, il topic 8 associa le parole "bill" e "gates" alla parola "vaccino". I restanti 6 topic ottenuti si riferiscono a tematiche di cui la popolazione ha discusso nelle argomentazioni nate al di sotto dei post dei politici. Sono presenti, infatti argomenti come il topic 0 e il topic 1 in cui vengono raggruppate parole come "lavoro", "italiani" e "sbaglio", e "governo" e "sbaraglio" che denotano la presenza di lamentele nei confronti della gestione della pandemia da parte della classe politica. Altre argomentazioni vedono come protagonisti la situazione emergenziale in Europa (topic 5), la gestione della pandemia in Italia, con riferimenti al Presidente del Consiglio Giuseppe Conte (topic 6), la situazione emergenziale negli ospedali (topic 7) e riferimenti ai tagli alla sanità e ai parlamentari, come è possibile notare nel topic 9.

Anche per quanto riguarda i topic estratti dai commenti di Instagram ne sono stati individuati tre che, separatamente, sembrano riferirsi alle tematiche già citate: il topic 1 viene descritto con peso maggiore dalle parole "bill", "gates" e "vaccino", il topic 5 dalle parole "5g" e "complotto" e il topic 8 dai termini "laboratorio", "virus", "creato". Gli altri topic, in maniera simile a quanto osservato nei risultati ottenuti su Facebook, si riferiscono a tematiche politiche ed economiche riferite alla crisi dovuta alla pandemia.

4.3.1.2. Confronto tra topic dei due Social Network

Nelle tabelle sottostanti sono state inserite le 25 parole che appartengono ai 10 topic individuati dal modello, evidenziando le similarità che caratterizzano i topic estratti da entrambi i Social Network. Sono state evidenziate utilizzando lo stesso colore parole che compaiono insieme nella descrizione dei topic di Facebook e Instagram, mentre sono stati evidenziati in grassetto i termini che, seppur in topic diversi, compaiono nei dieci topic in entrambe le piattaforme.

Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
italiani	governo	virus	5g	fatto
lavoro	travaglio	laboratorio	casa	no
fare	senza	stato	te	vogliamo
soldi	dopo	cosa	regione	nessuno
italiani	solo	già	giorni	buona
tutta	sbaraglio	essere	antenne	solo
vuole	cosa	persone	caso	paura
paese	grande	mondo	it	governo
andare	far	può	no	gente
ora	sbaglio	altro	primo	italia
visto	quando	vaccini	salute	sa
mentre	gente	cina	capito	famiglia
state	dare	fa	2020	sicuramente
sbaglio	fare	quindi	coronavirus	morti
secondo	mesi	medici	dico	esiste
possono	politici	covid	pubblica	popolazione
tutte	vedere	solo	pare	guerra
regioni	essere	coronavirus	cosa	mondiale
no	così	fare	coraggio	anno
mettere	dovrebbe	ancora	zero	complotto
grazie	legge	dire	bertolaso	serve
buon	dire	credo	post	milioni
senza	dilettanti	così	basta	voglio
prima	va	ciò	dietro	video
tasse	nessun	19	contrario	vogliono

Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
euro	conte	capire	gates	presidente
vita	persone	lombardia	bill	taglio
fatto	fetto	ospedale	vaccino	sbaglio
anni	emergenza	posti	popolo	grazie
stato	qualche	giusto	nn	sanità
altri	politica	piano	mes	fatto
stati	pandemia	fontana	oms	allora
dettaglio	giorno	problemi	chiedere	meglio
quando	viene	sicuro	and	parlamentari
sempre	bisogna	subito	trump	caro
solo	molti	votato	difficile	adesso
mai	ora	servono	futuro	parte
modo	male	pazienti	soros	ospedali
europa	meno	quei	tramite	cassa
altre	ecco	leggere	governi	ministro
momento	volta	responsabili	media	decreto
soldi	ieri	onde	uscito	economia
oltre	pubblici	lega	attuale	fare
aver	pd	pagamento	dovete	integrazione
però	italia	strumenti	sangue	milioni
stipendi	ecc	voto	cavie	infatti
bene	marco	centinaia	complimenti	fase
fine	nulla	risorse	microchip	sappiamo
invece	scuola	assumere	chiedere	sembra
qualcuno	giuseppe	tenere	http	parlamento

Tabella 13 – Topic dei commenti Facebook

Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
fa	gates	allora	fatto	taglio
cosa	bill	tempo	anni	così
covid	vaccino	ciò	sempre	italia
sbaglio	vita	fatti	altri	fare
nessuno	servizio	giro	proprio	lega
dire	palermo	meno	te	repentaglio
sembra	forza	bertolaso	solo	casa
popolo	mese	tasse	senza	italiani
va	nn	basta	conte	grazie
dettaglio	fine	invece	taglio	miliardi
parla	aver	qualche	detto	parlamentari
fuori	fatto	perché	sbaglio	soldi
italiano	ospedale	lavorare	cosa	ora
persona	marzo	signor	essere	stesso
meglio	tanti	ben	caso	prima
governo	sindacalista	preso	molto	devono
paese	vaccini	vergogna	mai	mettere
niente	15	letto	governo	state
po	messo	renzi	italia	fate
verità	10	ormai	parte	fatto
schifo	giorno	david	europa	vita
pure	iniziato	brembilla	stati	possono
neanche	polizia	italiani	ex	pagare
tutte	padre	chiusi	sanità	dobbiamo
chiudere	sospensione	decreto	modo	mai

Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
essere	solo	sbaraglio	virus	travaglio
complotto	gente	salvini	laboratorio	bene
vero	me	sempre	stato	quando
può	sbaglio	dilettanti	coronavirus	sai
uomo	lavoro	governo	cina	merda
persone	ancora	spero	creato	solo
ogni	fare	già	grande	fontana
5g	oggi	però	sars	giusto
punto	salute	ministro	visto	invece
no	nulla	dovrebbe	già	sanno
regione	stato	vogliono	altre	qualsiasi
bisogna	presidente	scritto	video	it
dicono	senza	meloni	19	natura
fare	persone	votato	cov	incapaci
molti	momento	post	2015	nemmeno
stare	prima	tv	2020	dittatura
ottobre	fare	amici	cosa	morire
qualcosa	altro	stelle	20	dice
anno	dato	quando	quando	vuoi
possibile	tanto	italia	mondo	assolutamente
avere	deve	gravi	milioni	febbraio
solo	vuole	bravo	sotto	bisogno
portato	situazione	mai	emergenza	stesso
mesi	cazzo	parlamento	così	bambini
comunque	forse	politici	pandemia	puoi

Tabella 14 – Topic dei commenti Instagram

A ciascun commento è stato assegnato un “Topic Dominante”, cioè quel topic che ha una percentuale di contributo maggiore rispetto agli altri all’interno del contenuto testuale. Nelle figure sottostanti è mostrato il numero di documenti che sono stati assegnati al topic dominante e il numero di documenti per ogni topic sommando il peso effettivo del contributo di ciascun topic ai rispettivi documenti.
Su Facebook, i topic dominanti nella maggior parte dei documenti sono il 2 e il 5, su Instagram il 3 e il 6 e lo 0,8 e 9.

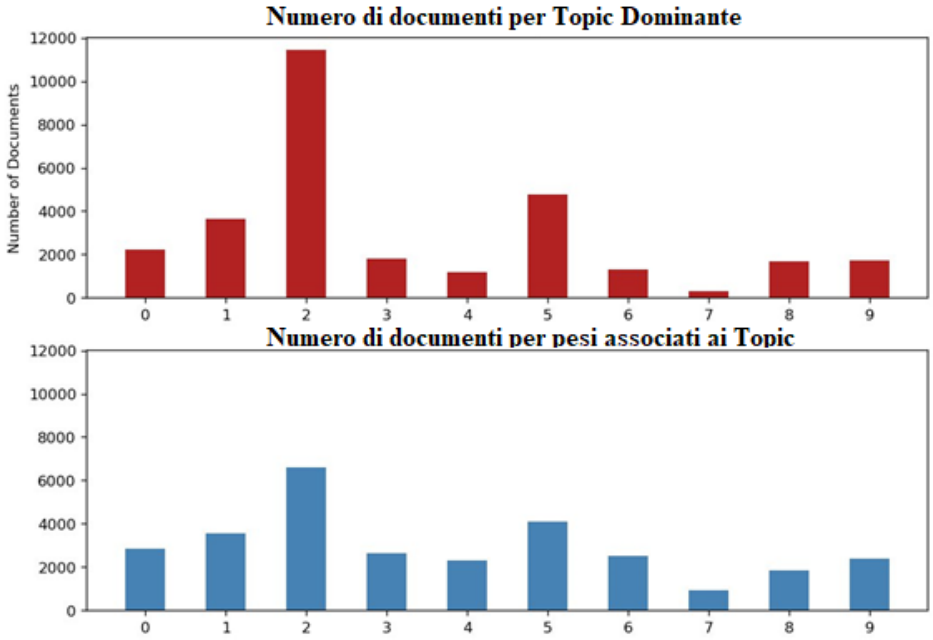

Figura 90 – Distribuzione dei commenti per topic dominante Facebook

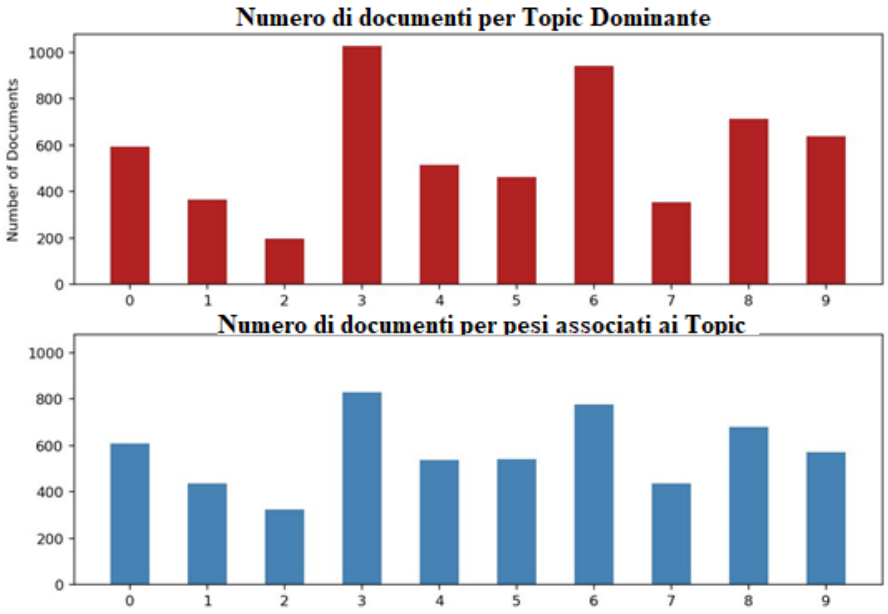

Figura 91 - Distribuzione dei commenti per topic dominante Instagram

4.3.1.3. Visualizzazione dei topic

Nelle figure 92 e 93 vengono mostrate le visualizzazioni dei documenti, ciascuno colorato a seconda del proprio topic dominante, tramite il modello di riduzione della dimensionalità t-SNE.

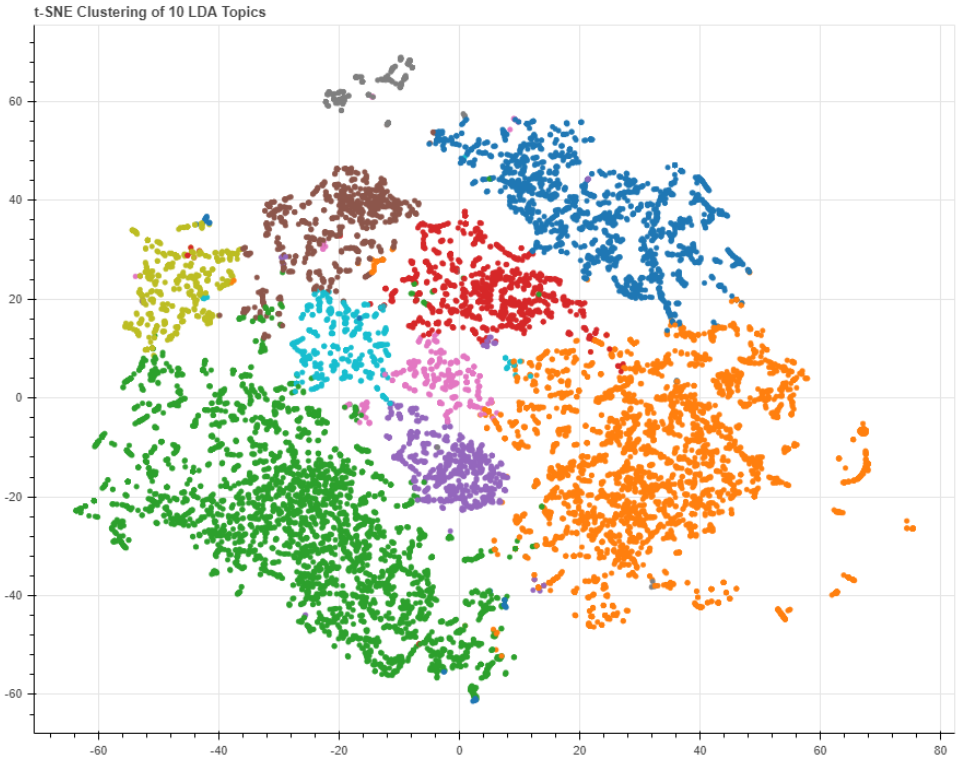

Figura 92 – Visualizzazione t-SNE topic Facebook

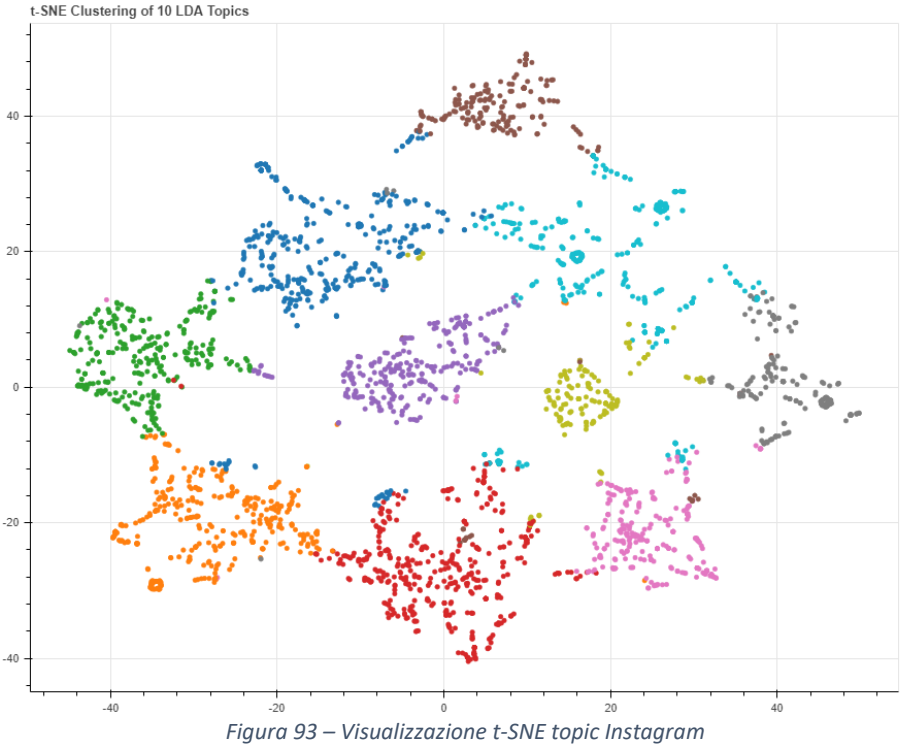

Figura 93 – Visualizzazione t-SNE topic Instagram

Nelle figure 94 e 95 vengono mostrati gli screenshot delle visualizzazioni interattive LDAvis dei topic estratti su Facebook e Instagram. In questo caso non è stato selezionato

alcun topic in modo da mostrare le 30 parole più frequenti nel set di commenti. In questa tipologia di visualizzazione il numero dei topic non corrisponde al numero dei topic assegnato fino a questo momento poiché essi vengono ordinati in ordine di importanza decrescente. Il topic 1 nella figura 94 corrisponde al topic 2 (relativo al coronavirus creato in laboratorio), il topic 1 nella figura 95 corrisponde al topic 3 (relativo alle lamentele sulla gestione politica).

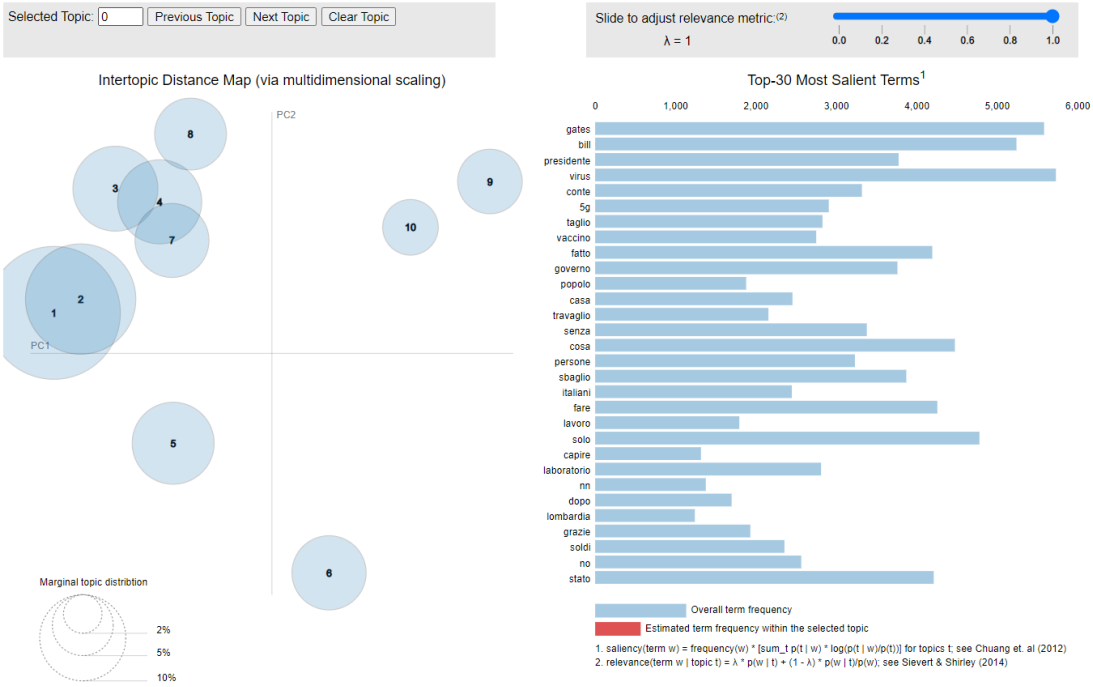

Figura 94 – Visualizzazione LDAvis commenti Facebook

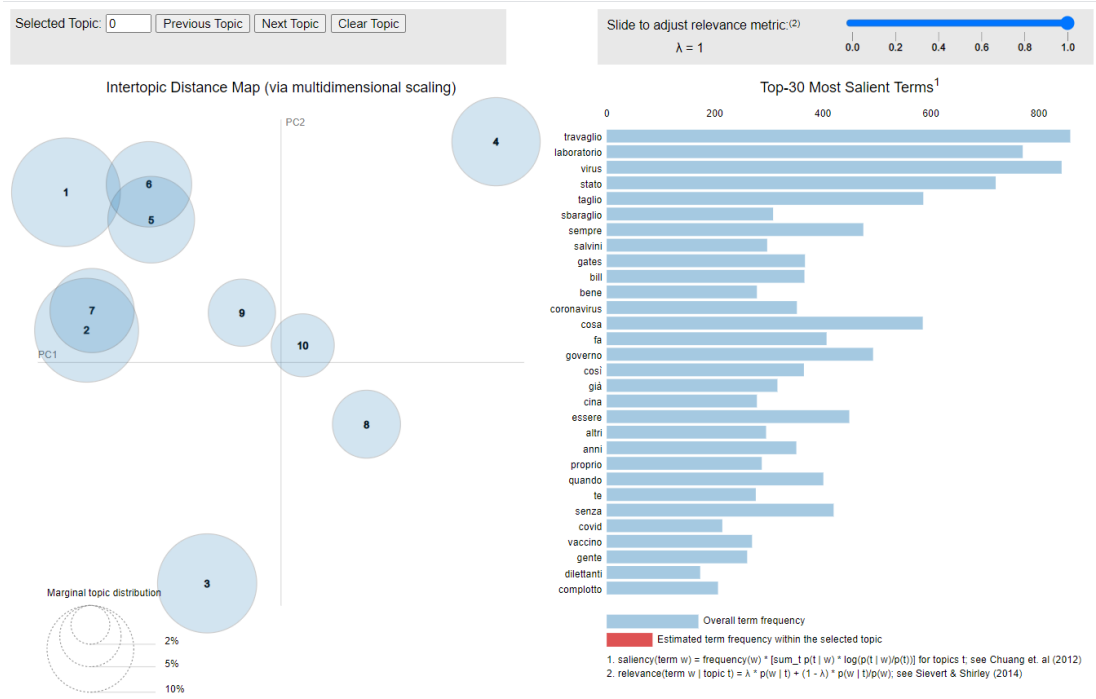

Figura 95 – Visualizzazione LDAvis commenti Facebook

4.3.2. Applicazione del modello sui post

Dopo aver estratto i topic di cui si è maggiormente discusso nei commenti pubblicati dalla popolazione al di sotto dei post pubblicati dalle personalità politiche di cui si è monitorata l'attività durante i primi mesi del 2020, si è scelto di attuare lo stesso modello agli stessi post sotto cui tali commenti sono stati pubblicati. L'intento è stato quello di osservare se ci fossero tematiche simili e, soprattutto, se le tematiche potenzialmente "fake" individuate all'interno dei commenti fossero sollecitate da argomentazioni simili. Poiché il confronto in questo caso è stato tra i topic dei commenti e i post, separatamente per ciascun Social Network, il numero di topic scelto come input del modello non è stato comune.

4.3.2.1. Post di Instagram

Inizialmente, anche per i post che hanno costituito il *corpus* dato in input al topic model, è stata mostrata la distribuzione del conteggio di caratteri contenuti in ciascun documento.

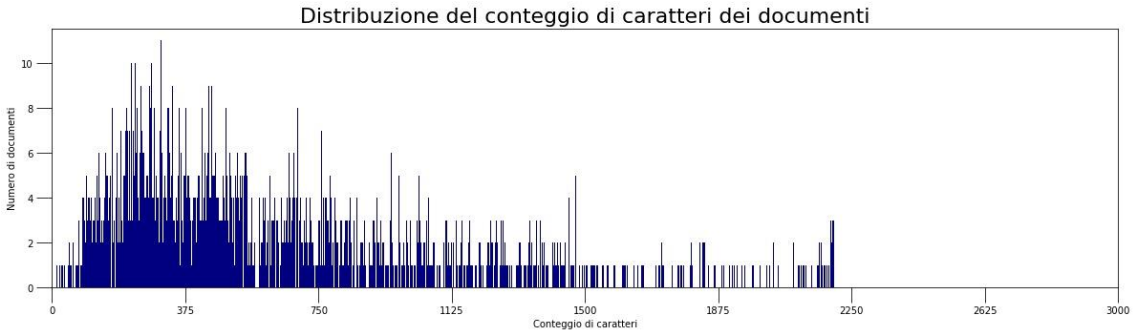

Figura 96 – Distribuzione conteggio di caratteri post Instagram

Per quanto riguarda i post pubblicati dalle personalità politiche su Instagram, dopo aver effettuato diverse prove con parametri differenti, è stato scelto il numero di topic pari a 8, che ha portato ad ottenere dei risultati ottimali dal punto di vista interpretativo e ben distinti dal punto di vista grafico, come si vedrà nelle due visualizzazioni riportate nel 4.3.2.3. Nelle seguenti figure sono riportate le Word Cloud e i diagrammi a barre ottenuti per ciascun topic estratto dal topic model, utilizzando la stessa metodologia descritta per i commenti.

Figura 97 – Word Cloud post Instagram

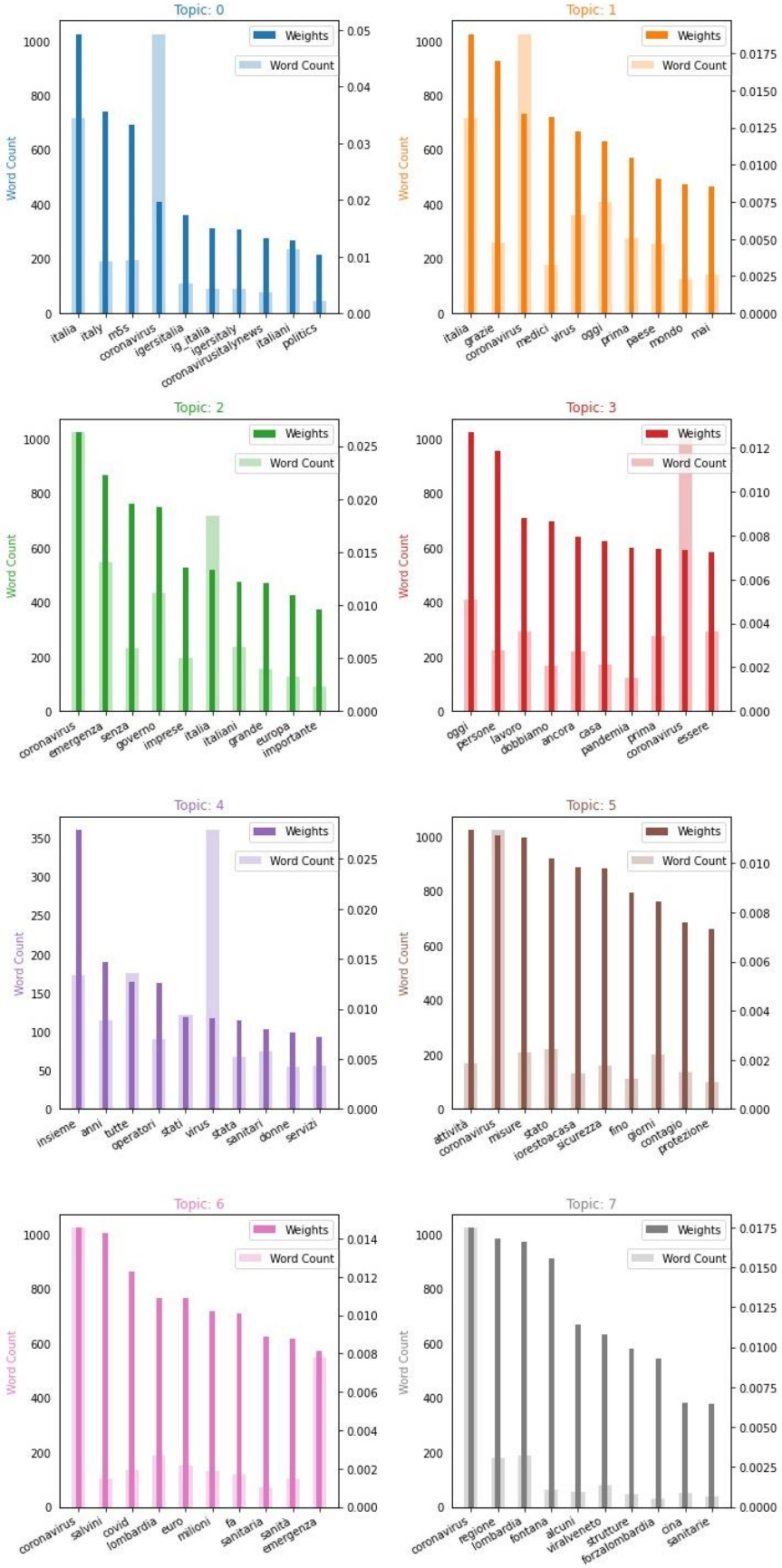

Figura 98 – Peso e frequenza parole topic dei post Instagram

4.3.2.2. Interpretazione dei topic dei post di Instagram

Gli otto topic estratti dal Latent Dirichlet Allocation sembrano definire esattamente otto tematiche differenti trattate dai personaggi politici, relativamente al tema "Coronavirus". A parte il topic 0 che sembra aver raggruppato tutti gli "hashtag" riferiti al Covid pubblicati come didascalia delle foto o dei video condivisi sulla piattaforma, il resto degli argomenti si riferiscono a temi discussi durante il periodo della pandemia. Il topic 1, per esempio, sembra riferirsi alle numerose frasi di ringraziamento e di incoraggiamento rivolte al personale sanitario dato che tale topic è descritto da parole come "grazie", "medici", "infermieri", "forza". Il topic 2 sembra essere riferito alle tematiche relative alla situazione economica e agli effetti del Coronavirus sulle imprese. Il topic 4 e 5 sembrano riguardare gli effetti sulla vita quotidiana dati dalle restrizioni utili per garantire la sicurezza, in particolare il topic 5 è rappresentato da termini come "misure", "sicurezza", "contagio". Il topic 3, 6 e 7 si riferiscono all'emergenza Covid da punti di vista differenti: il lavoro, la sanità e la gestione regionale. È interessante notare come nessun topic estratto si riferisce alle tematiche potenzialmente "fake" individuate nei commenti.

Il topic dominante presente nella maggior parte dei post è il numero 3.

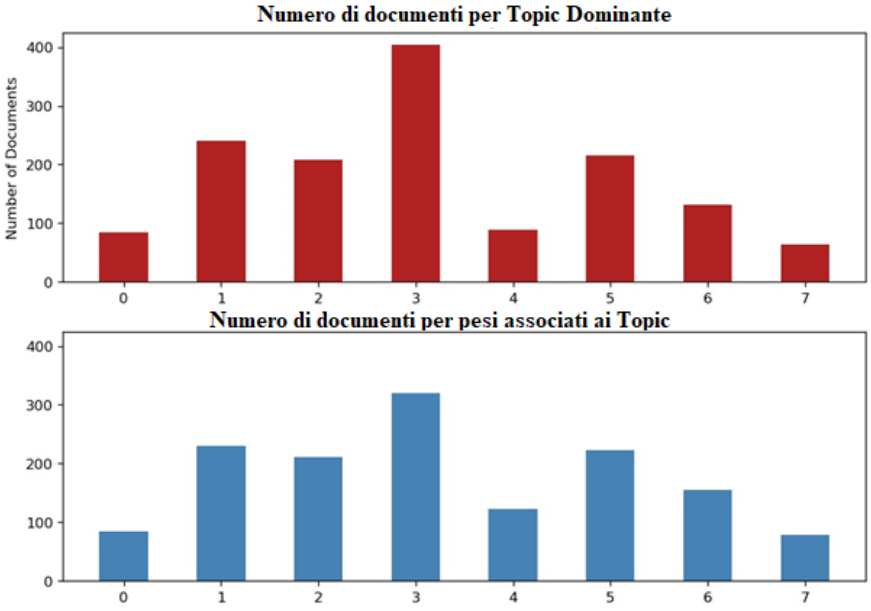

Figura 99 – Distribuzione post per topic dominante Instagram

4.3.2.3. Visualizzazione dei topic dei post di Instagram

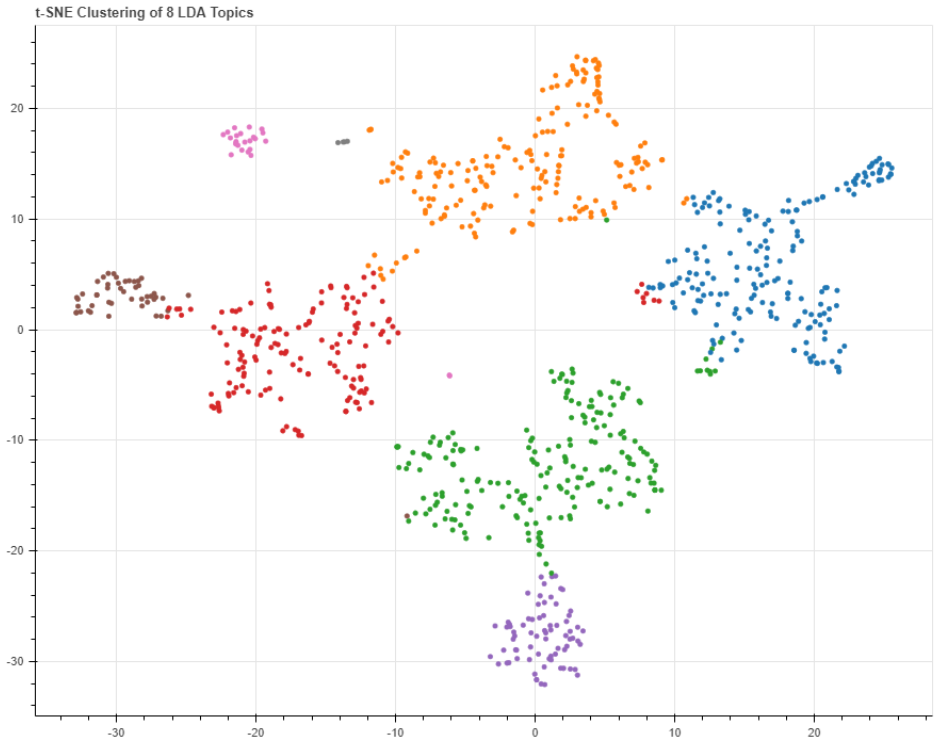

Figura 100 – Visualizzazione t-SNE post Instagram

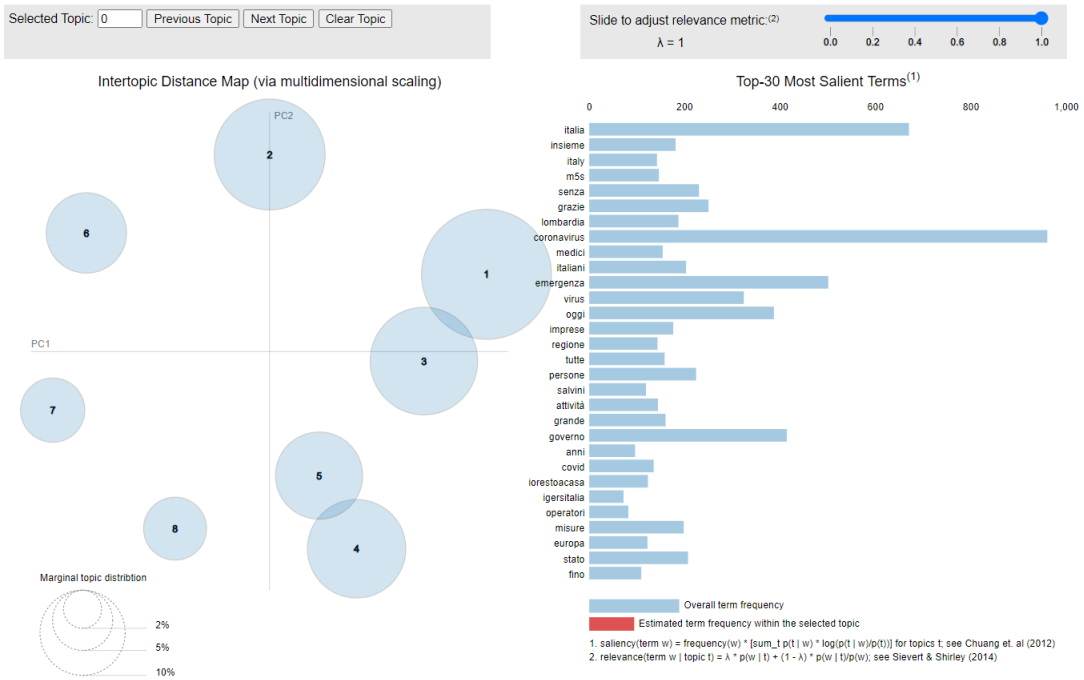

Figura 101 – Visualizzazione LDAvis post Instagram

4.3.2.4. Post di Facebook

La distribuzione del conteggio di parole di ciascun post pubblicato su Facebook è di seguito definita. Come si può notare, anche in questo caso essa sembra essere più spostata verso destra rispetto a quella di Instagram, indice del fatto che su Facebook vengono pubblicati post più lunghi, considerando che il massimo dei caratteri consentiti su Facebook è superiore.

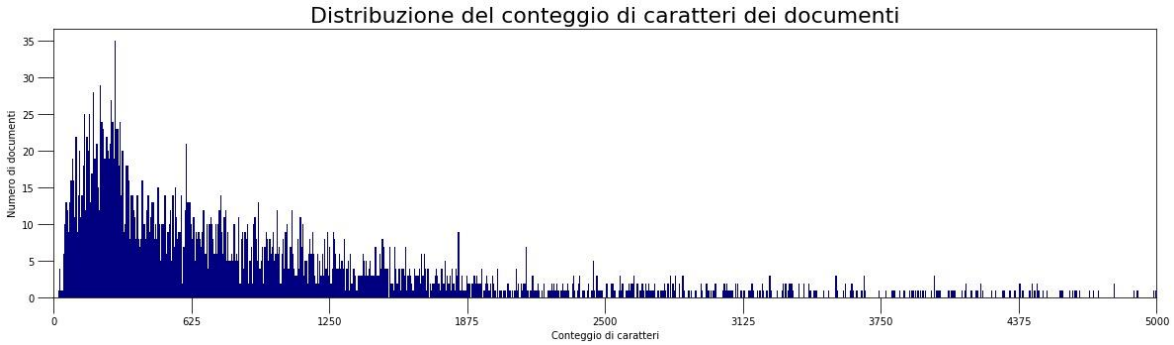

Figura 102 – Distribuzione conteggio di caratteri post Facebook

Seguendo le stesse considerazioni fatte in precedenza, il numero di topic scelto come input dell'LDA è stato 10. Di seguito sono mostrate le parole che meglio rappresentano i dieci topic ottenuti.

Figura 103 – Word Cloud post Facebook

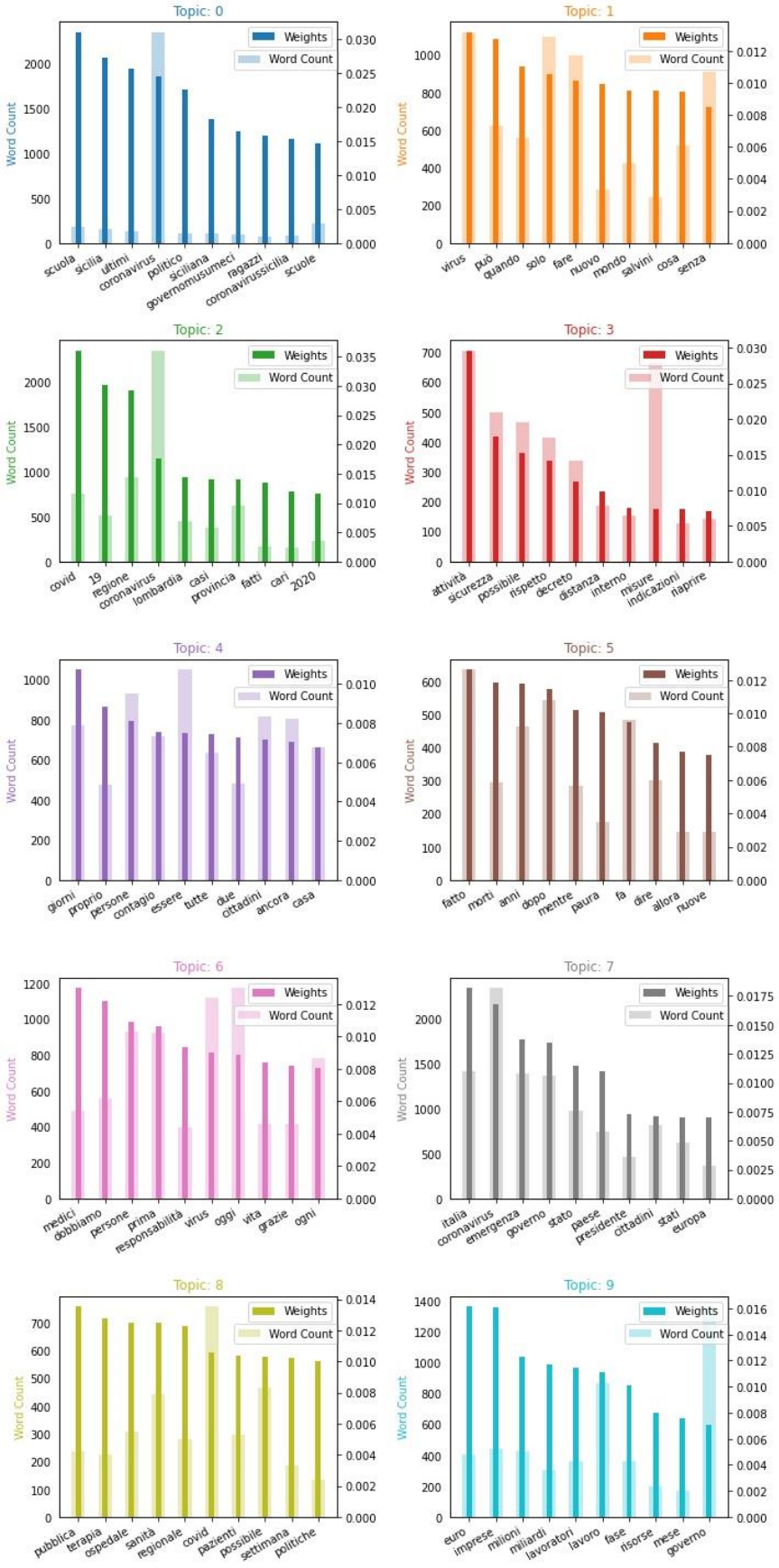

Figura 104 – Peso e frequenza parole per topic post Facebook

4.3.2.5. Interpretazione dei post di Facebook

I topic estratti dall'LDA effettuato sul set di post di Facebook i cui commenti hanno discusso argomentazioni potenzialmente "fake", sono molto simili ai topic riscontrati su Instagram. È da considerare che i profili monitorati su entrambe le piattaforme, soprattutto le personalità più seguite e, quindi, più commentate, sono le stesse, per cui il risultato è molto simile a quello atteso. In particolare, però, emergono delle argomentazioni differenti, come nel caso del topic 0 che sembra essere molto focalizzato sulla questione della didattica e della situazione emergenziale delle scuole italiane. Un altro topic che sembra differenziarsi dalle tematiche già viste sull'altro Social è il numero 2 che sembra rappresentare quel numero di post di aggiornamento sulla situazione dei casi di contagio nelle province e regioni italiane. Il topic 5, le cui parole con maggior peso sono "morti" e "paura", potrebbe essere interpretato come un argomento che raccoglie le sensazioni di paura nei confronti dell'epidemia. Gli altri topic estratti sono rappresentati prevalentemente da parole che riportano a tematiche come le misure restrittive per la sicurezza, le frasi di ringraziamento al personale sanitario, la situazione emergenziale in Europa e le misure economiche per la sanità e le imprese. Anche in questo caso, tra i topic estratti dal topic model, non sono presenti parole riferite alle argomentazioni potenzialmente "fake" riscontrate nei commenti.

La distribuzione dei topic dominanti tra i documenti appare sbilanciata, infatti circa 800 post sono associati al topic 4 e più di 1000 documenti sono associati al topic 7, come è visibile nella figura.

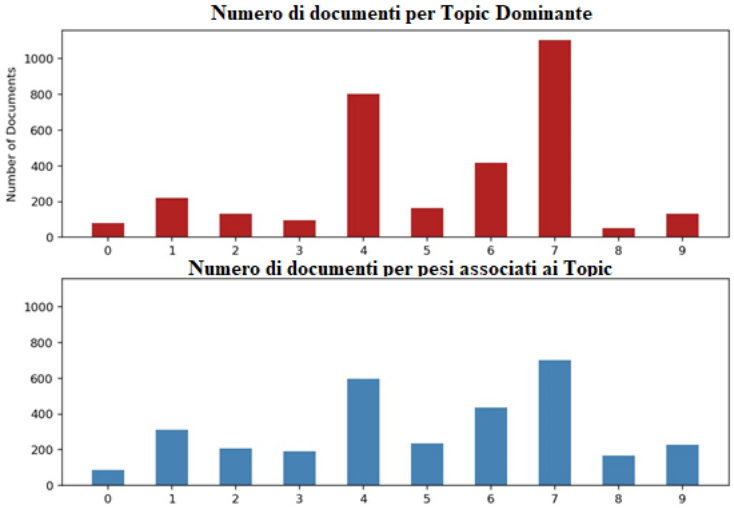

Figura 105 – Distribuzione post per topic dominante Facebook

4.3.2.6. Visualizzazione dei topic dei post di Facebook

Figura 106 – Visualizzazione t-SNE post Facebook

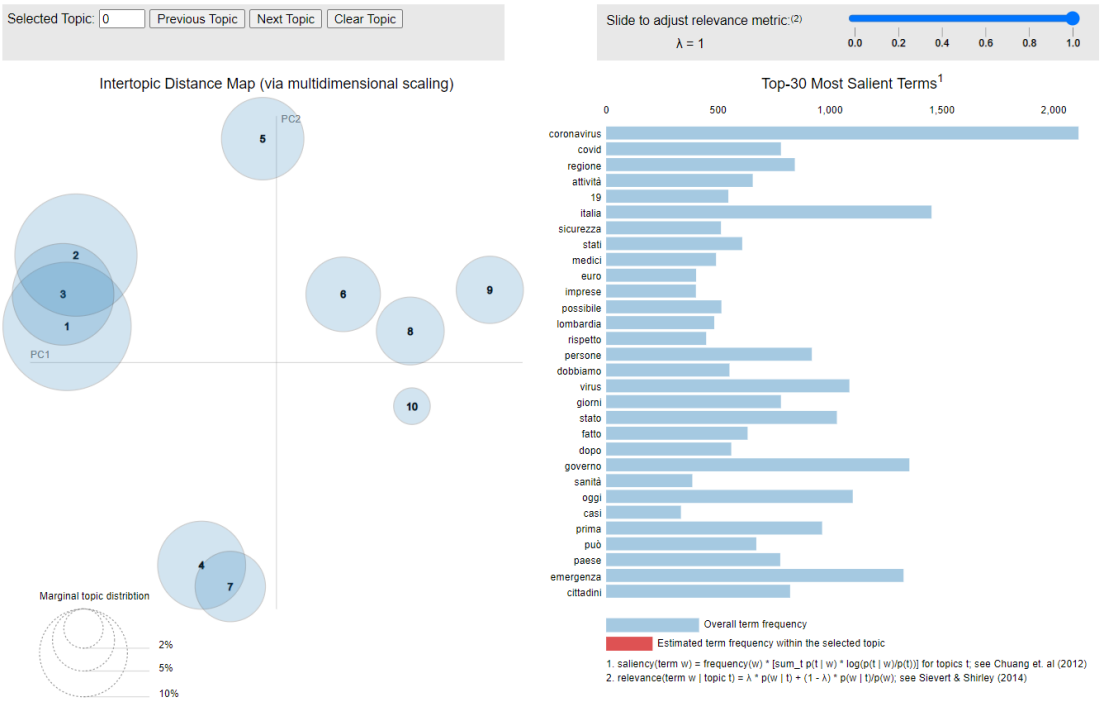

Figura 107 – Visualizzazione LDAvis post Facebook

Capitolo 5

Conclusioni e sviluppi futuri

Gli obiettivi di tale lavoro sono stati molteplici: in primo luogo si è cercato di caratterizzare l'attività e il seguito delle personalità politiche sui Social Network durante il periodo della prima ondata di pandemia Covid-19. Da questo punto di vista, si è riscontrata una generale tendenza dei profili appartenenti alla fazione del Centrodestra e del Movimento 5 Stelle alla pubblicazione frequente di contenuti a cui corrisponde un certo seguito che si traduce, su Facebook, in un alto numero di commenti e su Instagram in apprezzamenti sotto forma di "like". La fazione politica che ottiene un livello di engagement superiore rispetto al numero di follower è quella del Centrosinistra, che è la coalizione che produce però molti meno contenuti rispetto alle altre.

In secondo luogo, l'attenzione si è focalizzata sulla caratterizzazione delle argomentazioni riferite esplicitamente al Coronavirus che si sono diffuse nel periodo preso in considerazione, mettendo a confronto le due piattaforme social. In generale si è constatato che il Social Network in cui si è maggiormente discusso delle tematiche sopra citate è Facebook. L'andamento temporale di pubblicazione di contenuti, invece, è risultato simile nelle due piattaforme: le curve di interazione tra post e commenti hanno presentato infatti comportamenti simili, mostrando i picchi nei periodi in cui gli eventi legati al Coronavirus si sono verificati. In questa analisi si è potuto riscontrare che alcuni personaggi politici rappresentavano quasi degli outliers rispetto al resto delle personalità in quanto a numero di commenti e reazioni ricevute. L'aggregazione partitica che risulta essere stata più attiva sulla tematica Covid-19 e anche più seguita si conferma essere quella del Centrodestra.

In quanto alle tematiche che sono state dichiarate "fake" dall'Autorità per la Garanzia delle Comunicazioni (AGCOM), le cui keywords sono state ricercate all'interno dei contenuti testuali presenti sulle piattaforme, i risultati hanno confermato quanto detto in precedenza. Facebook è la piattaforma in cui tali argomentazioni sono state maggiormente citate; in particolare sono state individuate tre Fake News che per numero di occorrenza di keywords individuate, risultano essere state quelle di cui si è

maggiormente discusso su entrambi i Social. L'andamento della condivisione di contenuti riferiti a tali tematiche risulta essere simile in quanto, data anche la caratteristica propria delle Fake News, si è osservato un picco di pubblicazioni successivo alla diffusione della notizia e un conseguente appiattimento delle curve.

Infine, l'obiettivo ultimo di tale lavoro è stato quello di caratterizzare i contenuti testuali tramite l'applicazione di un modello di estrazione di topic come il Latent Dirichlet Allocation. Tale modello non ha necessitato di una profonda attività di preprocessing dei testi; la rimozione di una lista di stopword e degli elementi caratteristici della comunicazione sui social quali punteggiatura e emoticon, è stata sufficiente per ottenere dei risultati interpretabili. Partendo da un set di commenti con una numerosità medio-alta, il modello ha restituito dei topic interpretabili e riconducibili ad argomentazioni conosciute. Inoltre, tale modello è stato in grado di separare le tematiche relative alle Fake News più diffuse in topic ben definiti. Confrontando le parole che hanno descritto i topic estratti su Instagram e su Facebook, sono state riscontrate numerose similarità. Successivamente è stato applicato lo stesso topic model ai post sotto cui i commenti analizzati nella precedente sezione sono stati condivisi. Tale analisi ha dimostrato che le tematiche relative alle "Fake News" riportate dagli utenti nei commenti non sono state sollecitate esplicitamente nei contenuti testuali pubblicati dalle personalità politiche.

Questo lavoro si propone come punto di partenza per analisi successive che possano approfondire, con l'utilizzo di altri modelli NLP, la caratterizzazione dei testi pubblicati in tale contesto. I lavori futuri che possono essere intrapresi sono molteplici in quanto molteplici sono i punti di vista da cui poter osservare tale tematica. Ad esempio, basandosi sulle Fake News legate al Covid-19 analizzate in questo lavoro, si può considerare l'idea di attuare dei modelli di Sentiment Analysis per identificare ed estrarre le opinioni degli utenti che ne parlano; si può valutare di applicare degli algoritmi di clustering che possano far emergere delle somiglianze nei termini utilizzati nei commenti pubblicati sotto determinate personalità, considerandone il colore politico di appartenenza; inoltre si potrebbe confrontare la diffusione di contenuti "disinformativi" con tematiche informative, analizzandone differenze e similarità.

Riferimenti bibliografici

1. (Agcom), Autorità per le Garanzie nelle Comunicazioni. *Rapporto sul consumo di informazione.* Febbraio 2018.
2. Marco Delmastro, Antonio Nicita. *Big data: come trasformano l'economia e la politica.* Dicembre 2018.
3. (Agcom), Autorità per le Garanzie nelle Comunicazioni. *News vs Fake nel sistema dell'informazione.* Novembre 2018.
4. Comunicazioni, Autorità per le Garanzie nelle. *Le strategie di disinformazione online e la filiera di contenuti fake.* 2017.
5. Arendt, Hannah. *La menzogna in politica. Riflessioni sui Pentagon Papers.* Genova-Milano : Marietti 1820, 1972.
6. [Online] https://en.wikipedia.org/wiki/PageRank.
7. [Online] https://it.wikipedia.org/wiki/Profilazione_dell%27utente.
8. Lippman, Walter. *L'opinione pubblica.* [a cura di] Edizioni di Comunità. [trad.] Cesare Mannucci. Milano : Collana Saggi di Cultura Contemporanea n.26, 1922. Vol. Collana.
9. [Online] https://www.ilfattoquotidiano.it/2016/11/28/bufale-on-line-studio-sui-social-dimostra-che-gli-utenti-non-cambiano-idea-nemmeno-davanti-a-verita-accertate/3204587/.
10. [Online] https://lab24.ilsole24ore.com/storia-coronavirus/.
11. Comunicazioni, Autorità per le Garanzie nelle. *Osservatorio sulla disinformazione online, Speciale Coronavirus.* 2020.
12. *Agcom, fake news record durante l'emergenza covid.* Biondi, Andrea. 25 Novembre 2020, Il sole 24ORE.
13. [Online] https://secure.avaaz.org/campaign/en/facebook_coronavirus_misinformation/.
14. Giungato, Luigi. *La pandemia immateriale. Gli effetti del Covid-19 tra social asintomatici e comunicazione istituzionale.* Aprile 2020.
15. [Online] http://www.report.rai.it/dl/RaiTV/programmi/media/ContentItem-b9a195b5-023c-4366-b365-811022a068da.html.
16. [Online] https://www.repubblica.it/tecnologia/social-network/2020/08/12/news/coronavirus_migliaia_di_ricoveri_per_danni_da_fake_news-264473726/.
17. [Online] https://it.wikipedia.org/wiki/Teoria_degli_atti_linguistici.
18. [Online] http://www.datajournalism.it/tecnologia-fake-news-e-politica-un-incrocio-pericoloso/.
19. [Online] https://www.ilfattoquotidiano.it/2020/05/06/coronavirus-per-la-bbc-matteo-salvini-tra-i-politici-al-mondo-che-hanno-diffuso-fake-news-in-classifica-anche-trump-e-bolsonaro/5793246/.
20. [Online] https://www.repubblica.it/tecnologia/social-network/2020/05/19/news/twitter_i_cinque_profili_italiani_super_diffusori_di_fake_news_sul_coronavirus-257082395/.
21. Zafarani, Xinyi Zhou and Reza. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Computing Survey.* October 2020, p. 40.
22. [Online] https://www.gartner.com/en/information-technology/glossary/big-data.
23. [Online] https://www.cwi.it/big-data.
24. [Online] https://hadoop.apache.org/.
25. [Online] https://spark.apache.org/.

26. [Online] https://www.agendadigitale.eu/cultura-digitale/linguaggio-naturale-e-intelligenza-artificiale-a-che-punto-siamo/.
27. [Online] https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1.
28. [Online] https://it.wikipedia.org/wiki/Modello_della_borsa_di_parole.
29. [Online] https://it.wikipedia.org/wiki/Tf-idf.
30. [Online] https://monkeylearn.com/blog/introduction-to-topic-modeling/.
31. [Online] https://www.researchgate.net/publication/200045222_An_Introduction_to_Latent_Semantic_Analysis.
32. *Latent Dirichlet Allocation.* David M. Blei, Andrew Y. Ng, Micheal I. Jornal. 2003, Journal of Machine Learning Research 3, p. 993-1022.
33. [Online] https://www.researchgate.net/publication/265784473_LDAvis_A_method_for_visualizing_and_interpreting_topics.
34. [Online] [2] http://bl.ocks.org/AlessandraSozzi/raw/ce1ace56e4aed6f2d614ae2243aab5a5/.
35. [Online] https://towardsdatascience.com/an-introduction-to-t-sne-with-python-example-5a3a293108d1.
36. [Online] https://www.datacamp.com/community/tutorials/introduction-t-sne.
37. [Online] https://www.rtinsights.com/why-python-is-essential-for-data-analysis/.
38. [Online] https://it.wikipedia.org/wiki/Python.
39. [Online] https://pandas.pydata.org/.
40. [Online] https://pandas.pydata.org/docs/getting_started/comparison/comparison_with_sql.html.
41. [Online] https://matplotlib.org/.
42. [Online] https://www.nltk.org/.
43. [Online] https://radimrehurek.com/gensim/.
44. [Online] https://spark.apache.org/docs/latest/api/python/pyspark.sql.html.
45. [Online] https://www.javatpoint.com/pyspark-sql#:~:text=PySpark%20SQL%20is%20a%20module,same%20as%20the%20SQL%20language..
46. [Online] https://www.polito.it/ricerca/infrastrutture/hpc4ai/.
47. [Online] https://it.wikipedia.org/wiki/Structured_Query_Language.
48. [Online] https://rapidminer.com/.
49. [Online] https://powerbi.microsoft.com/it-it/.
50. *Towards Understanding Political Interactions on Instagram.* Martino Trevisan, Luca Vassio, Idilio Drago, Marco Mellia, Fabricio Murai, Flavio Figueiredo, Ana Paula Couto da Silva, and Jussara M. Almeida. 2019, In Proceedings of the 30th ACM Conference on Hypertext and Social Media.
51. *Political Polarization on the Digital Sphere: A Cross-platform, Over-time Analysis of Interactional, Positional, and Affective Polarization on Social Media.* Moran Yarchi, Christian Baden & Neta Kligler-Vilenchik. s.l. : Political Communication, 2020.
52. *Fake news and COVID-19: modelling the predictors of fake news sharing among social media users.* Oberiri Destiny Apuke, Bahiyah Omar. s.l. : Telematics and Informatics, 2021, ScienceDirect, Vol. 56.
53. [Online] https://it.wikipedia.org/wiki/Uses_and_gratification.
54. *The COVID-19 social media infodemic.* Cinelli, M., Quattrociocchi, W., Galeazzi, A. et al. 2020, Sci Rep 10.
55. *COVID-19 and the 5G Conspiracy Theory: Social Network Analysis of Twitter Data.* Ahmed W, Vidal-Alaball J, Downing J, López Seguí F. 2020, J Med Internet Res.

56. *Topic Modeling Approaches for Understanding COVID-19 MisinformationSpread in Sub-Saharan Africa.* Nwankwo E, Okolo C, Habonimana C. 2020, in AI for Social Good Workshop.
57. *Fake news agenda in the era of COVID-19: Identifying trends through fact-checking content.* Wilson Ceron, Mathias-Felipe de-Lima-Santos, Marcos G. Quiles. 2021, Online Social Networks and Media, Vol. 21.
58. *Exploring Public Response to COVID-19 on Weibo with LDA Topic Modeling and Sentiment Analysis.* Xie, R., Chu, S. K. W., Chiu, D. K. W., & Wang, Y. 2021, Data and Information Management,, Vol. 86-99, p. 5(1).
59. *Analysis of spatiotemporal characteristics of big data on social media sentiment with COVID-19 epidemic topics.* Zhu B, Zheng X, Liu H, Li J, Wang P. Nov 2020, Chaos Solitons Fractals.
60. [Online] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.
61. [Online] https://it.wikipedia.org/wiki/Crawler.
62. [Online] https://blog.hootsuite.com/social-media-metrics/.
63. [Online] https://machinelearningmastery.com/empirical-distribution-function-in-python/.
64. [Online] https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2.
65. [Online] https://searchstorage.techtarget.com/definition/big-data-storage.

Ringraziamenti

Vorrei innanzitutto ringraziare il Prof. *Carlo Cambini* per avermi proposto un lavoro di tesi che mi ha dato l'opportunità di applicare l'analisi dei dati ad una tematica così attuale. La ringrazio per aver mostrato interesse e cura nei confronti delle tematiche affrontate, ciò mi ha stimolata a dare sempre il massimo. In particolare, la ringrazio per aver trovato sempre parole di conforto e di incoraggiamento per i momenti delicati che hanno caratterizzato questo ultimo anno universitario.

Inoltre, vorrei ringraziare i Proff. *Luca Vassio* e *Luca Cagliero* che ho avuto la possibilità di conoscere grazie a questo progetto di tesi. Vi ringrazio per i numerosi consigli e suggerimenti che mi hanno permesso di portare a compimento il lavoro, per avermi dato l'opportunità di conoscere un mondo così nuovo e interessante e per il tempo che mi avete dedicato in questo periodo così complicato per tutti.

Un particolare ringraziamento va ad *Alessandro Ciociola*, una delle persone che ho avuto la fortuna di "incontrare" grazie a questo progetto. Il tuo contributo professionale e la tua disponibilità mostrata sin dall'inizio sono stati fondamentali per la definizione dell'intero lavoro. Voglio sinceramente ringraziarti per il tempo che mi hai dedicato, per tutti gli spunti di riflessione che mi hai offerto, ma soprattutto per avermi fatto appassionare al "tuo" mondo con la passione e l'entusiasmo che ti contraddistingue e che mostri in tutto ciò di cui ti occupi.