Università degli Studi di Milano-Bicocca
DATA ANALYTICS
2019-2020

Sentiment analysis su tweets riguardanti i vaccini per Covid19

Alessandro Bertolo (808319) - a.bertolo2@campus.unimib.it

Introduzione

Dall'inizio della pandemia di Covid19, uno degli argomenti maggiormente trattato è quello dei vaccini. Social Networks come Twitter, in questi mesi, hanno subito un'accelerazione nella quantità di informazioni prodotte dagli utenti riguardanti la malattia e tutto ciò che la circonda. Risulta quindi interessante analizzare ciò che gli utenti scrivono a riguardo.

Uno degli argomenti di maggior divisione sociale riguarda lo sviluppo del vaccino e la campagna vaccinale.
Tramite l'analisi di questi tweet il progetto mira a rispondere ad alcune domande:

- è possibile identificare una tendenza preponderante nel sentiment dei tweets riferiti ai vaccini?
- in quali paesi l'argomento dei vaccini è più presente? è possibile identificare una polarità preponderante per ciascun paese?
- tra gli utenti con maggiore seguito, vi è una tendenza maggiore al sentiment positivo o negativo?

Non avendo a disposizione dei dati già etichettati con il sentiment, ai fini del progetto è necessario classificare i tweets raccolti. Per farlo vengono utilizzati due metodologie. Le domande correlate sono quindi:

- dai due approcci di sentiment extraction si ottengono risultati simili?
- si riesce ad ottenere una buona astrazione?
- quale dei due è migliore?

Il codice di processing e la dashbord contenente le visualizzazioni sono disponibili nel repository github.

Descrizione del dataset

Ai fini del progetto sono stati utilizzati dati ottenuti dalla combinazione di due datasets:

- Pfizer Vaccine Tweets[1]
- Covid Vaccine Tweets[2]

Entrambi contengono una lista di tweets, estratti utilizzando le API di Twitter, contenenti informazioni inerenti ai vaccini per il Covid19.
Attualmente vengono aggiornati su base quotidiana con nuovi dati.

Pfizer Vaccine Tweets
Contiene tweets riferiti specificatamente al vaccino dalla casa farmaceutica Pfizer & BioNTech. La raccolta è iniziata il 12 Dicembre 2020 ed il numero di tweets presenti, aggiornato al 26 Gennaio 2021, è di **4560**.

Covid Vaccine Tweets
Contiene tweets selezionati in base alla presenza dell'hashtag #CovidVaccine tra gli hashtags. La raccolta giornaliera è iniziata il 1 Agosto 2020 partendo con 6000 tweets, tra questi il più vecchio risale al 9 Agosto 2020.
Il numero di tweet presenti, aggiornato al 26 Gennaio 2021, è di **59396**.
Problemi riscontrati nel dataset:

- il campo text di ciascun tweet non è stato estratto correttamente, di fatto il testo viene troncato al vecchio limite di Twitter di 140 caratteri pur essendo stato scritto successivamente al cambiamento a 280 caratteri
- il campo della data di creazione del tweet non segue la notazione inglese mm/dd/yyyy ma la notazione dd/mm/yyyy

Non condividendo esattamente gli stessi attributi è stato necessario scartare alcune informazioni dal dataset Pfizer:

- id del tweet
- numero di retweets del tweet
- numero di favorites del tweet

Inoltre, essendo il periodo temporale della raccolta parzialmente sovrapposto, è stata eseguita la rimozione dei tweet duplicati sulla base dello username, della data ed del testo del tweet.
Il dataset ottenuto dall'unione dei due dataset descritti contiene **13 campi** e **63695 tweet** unici.

CAMPO	DESCRIZIONE
user_name	Nome del profilo utente
user_location	Location del profilo utente

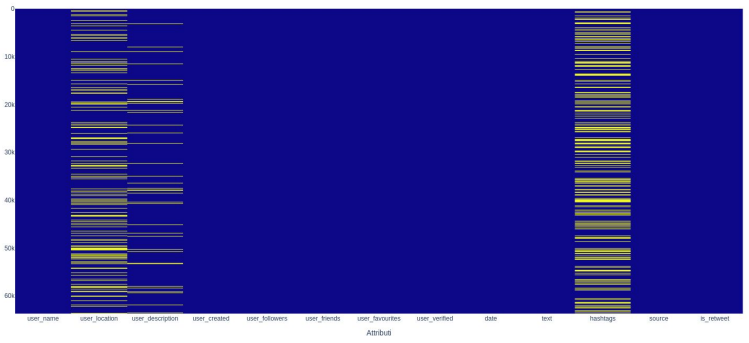
[1] https://www.kaggle.com/gpreda/pfizer-vaccine-tweets
[2] https://www.kaggle.com/kaushiksuresh147/covidvaccine-tweets

user_description	Descrizione dell'utente
user_created	Data di creazione del profilo utente (UTC)
user_followers	Numero di followers del profilo utente
user_friend	Numero di amici del profilo utente
user_favourites	Numero di preferiti del profilo utente
user_verified	Booleano, indica se il profilo utente è verificato o no.
date	Data di creazione del tweet (UTC)
text	Testo del tweet
hashtags	Hashtags contenuti nel testo del tweet
source	Dispositivo o software utilizzato per pubblicare il tweet
is_retweet	Booleano, indica se il tweet estratto è un retweet o un tweet "originale" dell'utente

Analisi preliminare

In questa fase preliminare sono state eseguite le prime verifiche riguardo il contenuto dei dati ottenuti dalla combinazione.
In primo luogo si è cercato il numero di campi vuoti presenti:

```
user_name               1
user_location       13533
user_description     3868
user_created            0
user_followers          0
user_friends            0
user_favourites         0
user_verified           0
date                    0
text                    0
hashtags            19493
source                  4
is_retweet              0
```



(numero di valori nulli presenti e visualizzazione grafica)

- 1 solo tweet non presenta il campo utente
- 19493 tweets non contengono hashtags
- 13533 tweets non contengono una location riguardo il profilo dell'utente, questo perché non è un dato obbligatorio richiesto da Twitter e gran parte degli utenti presenti sulla piattaforma non condividono informazioni sulla loro provenienza
- 3868 tweets non contengono una descrizione riguardo il profilo dell'utente che ha pubblicato il tweet, anche in questo caso la descrizione non è un campo obbligatorio

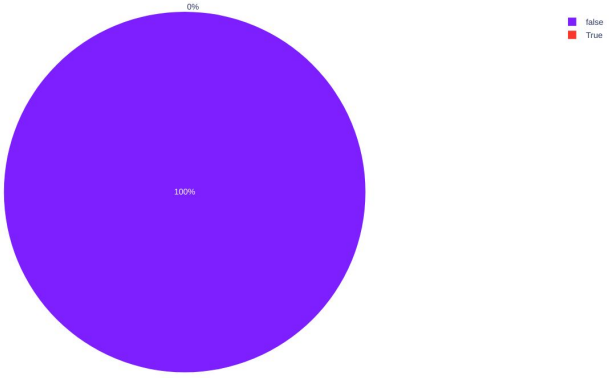
All'interno del campo user_location sono state individuate numerose stringhe non utili ai fini dell'analisi. Essendo un campo opzionale e liberamente gestibile dall'utente, può contenere qualsiasi informazione come: nomi di fantasia ('My bedroom', 'Space', 'The World'), molteplici città o molteplici paesi, singoli stati americani, indirizzi completi o sigle, emoji,

slang di vario genere o semplicemente delle parole non riconducibili ad una qualsivoglia posizione.

12621	San Francisco, CA
15505	Jupiter
13626	London, England
30213	NY
52341	Chennai, India
13019	Salisbury, England
53529	MostlyInTheDogHouseIsTheNorm
11231	Washington, DC| California
26395	NaN

(esempio di informazioni contenute nel campo user_location)

Per quanto riguarda il campo is_retweet, che indica se il tweet è o meno un tweet retweet, esso contiene solamente valori False. Questo significa che nel processo di estrazione dei due dataset sono stati considerati solamente i tweet scritti da ciascun utente. Si tratta quindi di un campo poco utile ai fini della EDA.



(distribuzione valori campo is_retweet)

Preprocessing effettuato

In seguito all'analisi preliminare sono state eseguite delle operazioni di pulizia dei dati:

- gestione dei valori nulli
- standardizzazione del campo user_location
- estrazione di token dal campo text

I valori nulli presenti nel dataset sono stati sostituiti dalla keyword 'None', ad esclusione del tweet senza il campo user_name che è stato scartato.

Standardizzazione locations

Come riportato nell'analisi preliminare, le informazioni contenute in questo campo non possono essere utilizzate direttamente per l'analisi. Per questo motivo è stata condotta la standardizzazione dei valori al codice iso 3-digit[3].

Ciascun campo user_location è stato processato per individuare un iso code a 2 o 3 digit riferito ad una nazione, un nome di nazione (o uno tra i nomi alternativi), uno stato americano (codice o nome) o una tra le città principali del mondo.

[3] https://en.wikipedia.org/wiki/ISO_3166-1_alpha-3

Il processo è stato condotto effettuando prima una pulizia del campo da eventuali elementi inopportuni (come URL, mentions, hashtags, numeri, emoji) e confrontando le parole presenti con dei dataset di nazioni[4], stati americani[5] e città principali globali[6].

```
['home'] -> None
['London', 'UK'] -> GBR
['International'] -> None
['Copenhagen', 'Denmark'] -> DNK
['Philadelphia', 'PA'] -> USA
```

(campione di esempio di standardizzazione effettuata)

In questo passaggio sono stati identificati **36511** nazioni sul totale di **50161** campi user_location presenti.
L'algoritmo di ricerca implementato è molto semplice e grezzo, di conseguenza solamente le occorrenze esatte sono state mantenute.

Estrazione dei token
L'estrazione dei token (tokenization) dal campo text dei tweet è stata implementata in funzione della EDA. Ai fini dell'estrazione del sentiment vengono utilizzate altri passaggi di tokenization, come indicato successivamente per ciascun metodo.

Gli step di tokenization effettuati sono i seguenti:

1. lowercase
2. rimozione di url
3. rimozione di mentions
4. rimozione di hashtags
5. rimozione di separatori e quotes
6. rimozione di numeri
7. rimozione punteggiatura
8. rimozione di tag html
9. rimozione di emoji
10. tokenization tramite WordNet
11. rimozione stopwords inglese e di parole inopportune riscontrate successivamente
12. lemmatizzazione

text	**tokens**
'Just had my second covid vaccine. \n\n#Covid_19 #CovidVaccine https://t.co/vd2SYMgtbv'	['second', 'covid', 'vaccine']
'Soooo I'm noticing that posting the #CovidVaccine card is just like showing the "I voted" sticker... so here \n😂😂\nGe… https://t.co/6DSdEfpvGw'	['soooo', 'noticing', 'posting', 'card', 'like', 'showing', 'voted', 'sticker', 'ge']

(esempio di estrazione di tokens effettuato)

[4] https://www.kaggle.com/gbertou/iso-country-codes-with-alternative-country-names
[5] https://datahub.io/core/country-list
[6] https://datahub.io/core/world-cities

EDA

La parte di analisi esplorativa del dataset ottenuto è stata divisa in due sezioni: la prima riporta le considerazioni sugli utenti mentre la seconda sui tweet.
Il dataset in seguito alle operazioni di preprocessing effettuate contiene **14 campi** e **63694 tweet** unici.

user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_verified	date
Rachel Roh	La Crescenta-Montrose, CA	Aggregator of Asian American news; scanning diverse sources 24/7/365. RT's, Follows and 'Likes' will fuel me 🧑‍💻	2009-04-08T17:52:46	405	1692	3247	false	2020-12-20T06:06:44
Albert Fong	San Francisco, CA	Marketing dude, tech geek, heavy metal & '80s music junkie. Fascinated by meteorology and all things in the cloud. Opinions are my own.	2009-09-21T15:27:30	834	666	178	false	2020-12-13T16:27:13
eli🇱🇹🇪🇺👌	Your Bed	heil, hydra ✋⚙	2020-06-25T23:30:28	10	88	155	false	2020-12-12T20:33:45

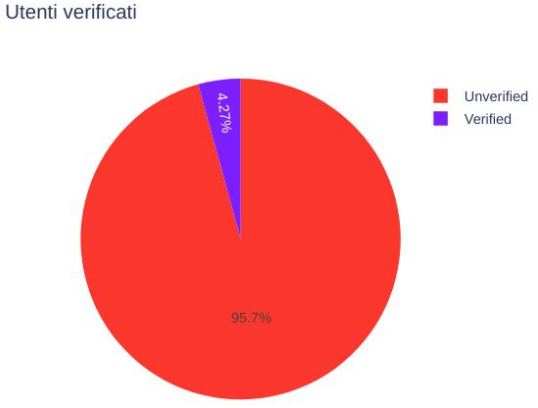
text	hashtags	source	is_retweet	country
Same folks said daikon paste could treat a cytokine storm #PfizerBioNTech https://t.co/xeHhIMg1kF	PfizerBioNTech	Twitter for Android	false	USA
While the world has been on the wrong side of history this year, hopefully, the biggest vaccination effort we've ev… https://t.co/dlCHrZjkhm		Twitter Web App	false	CAN
#coronavirus #SputnikV #AstraZeneca #PfizerBioNTech #Moderna #Covid_19 Russian vaccine is created to last 2-4 years… https://t.co/ieYlCKBr8P	coronavirusSputnikVAstraZenecaPfizerBioNTechModernaCovid_19	Twitter for Android	false	

(esempio di tweets contenuti nel dataset)

Utenti

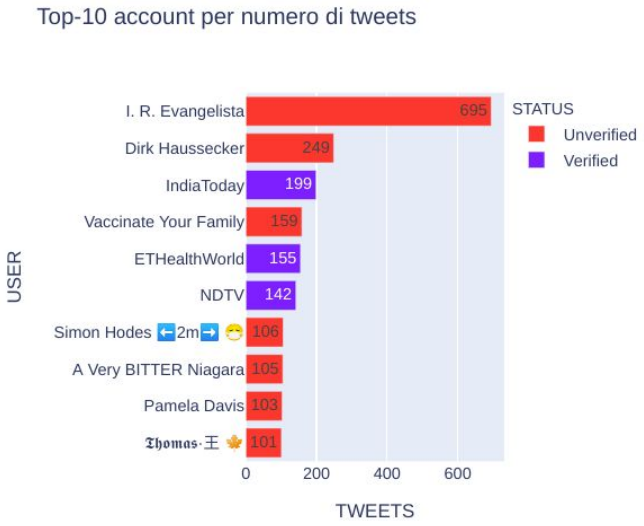
Nel dataset sono presenti **38421 utenti unici**.
Tra i campi presenti riferiti a ciascun utente non sono stati analizzati la data di creazione, la descrizione dell'utente, il numero di amici e di favoriti.

- **Tra questi quanti utenti sono verificati?**



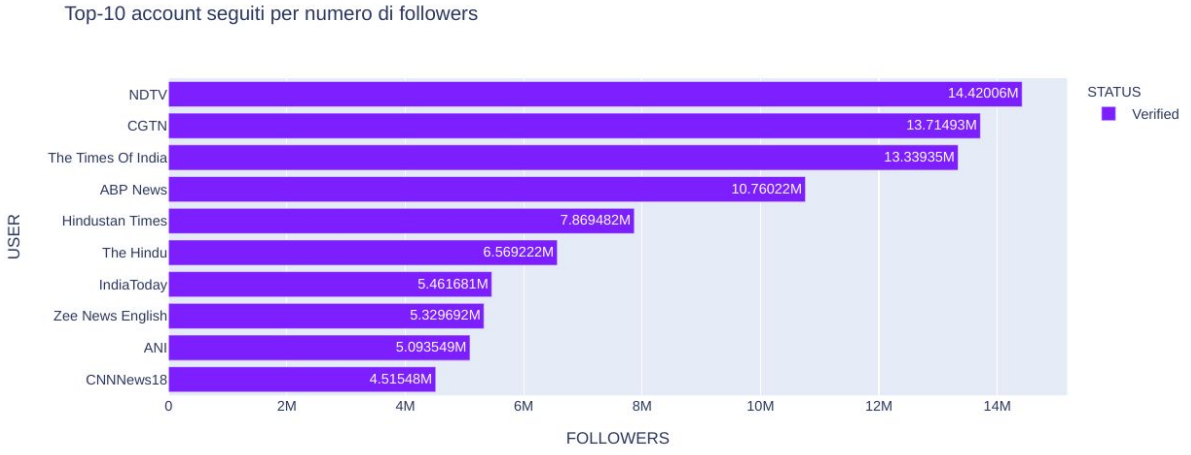
Tra i 3842 utenti unici, 1642 (4.3%) sono verificati e 36779 (95,7%) non sono verificati.

- **Quali sono gli utenti più attivi per numero di tweet?**



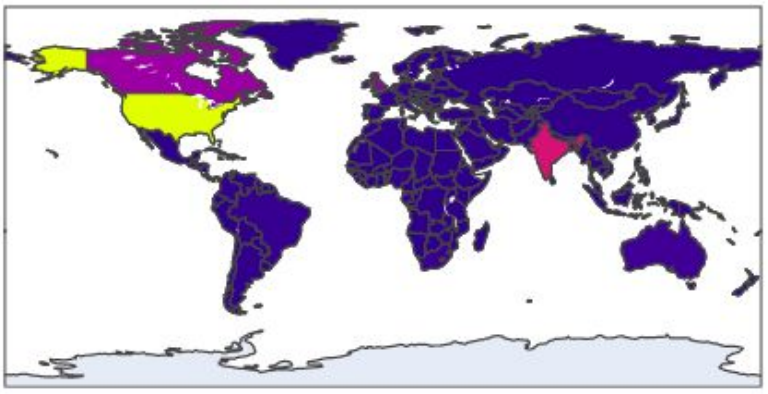
Gli utenti più attivi, per numero di tweets effettuati, sono principalmente utente non verificati. Questo può essere spiegato dal fatto che la gran parte dei dati presenti sono stati raccolti utilizzando come filtro degli hashtags. Di conseguenza, le fonti governative o giornalistiche presenti non sono inclini ad inserire hashtags nei tweets rispetto ai profili utente personali.

- **Quali sono gli utenti con il maggior numero di follower?**

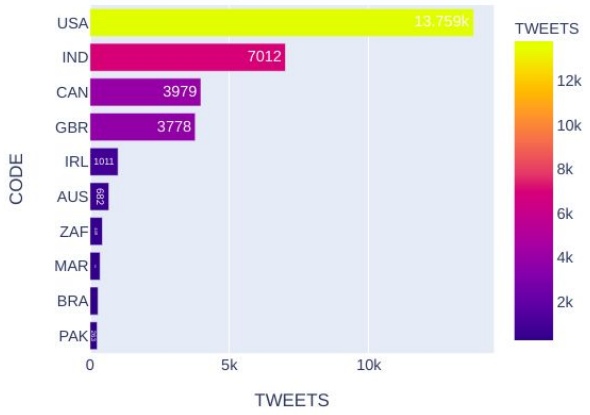


Viceversa, gli utenti con il maggior numero di followers sono utenti verificati e nello specifico delle fonti giornalistiche o governative (NDTV, CGTN, The Times Of India sono network giornalistici molto rilevanti in oriente).

- **Da quali nazioni provengono gli utenti?**



Top-10 nazioni per numero di tweets

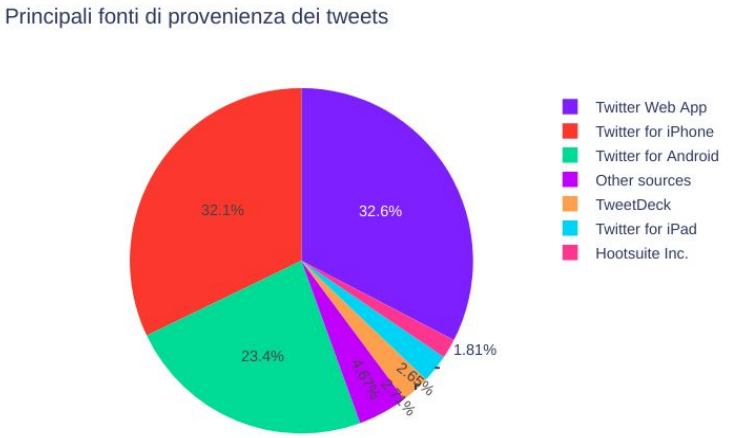


La maggior parte degli utenti presenti dichiara di provenire dagli Stati Uniti e dal Canada mentre una larga parte dall'India.
In generale, tranne per alcune eccezioni, i tweets sono tutti in inglese anche perché l'estrazione è stata fatta utilizzando degli hashtags in inglese.

Tweets

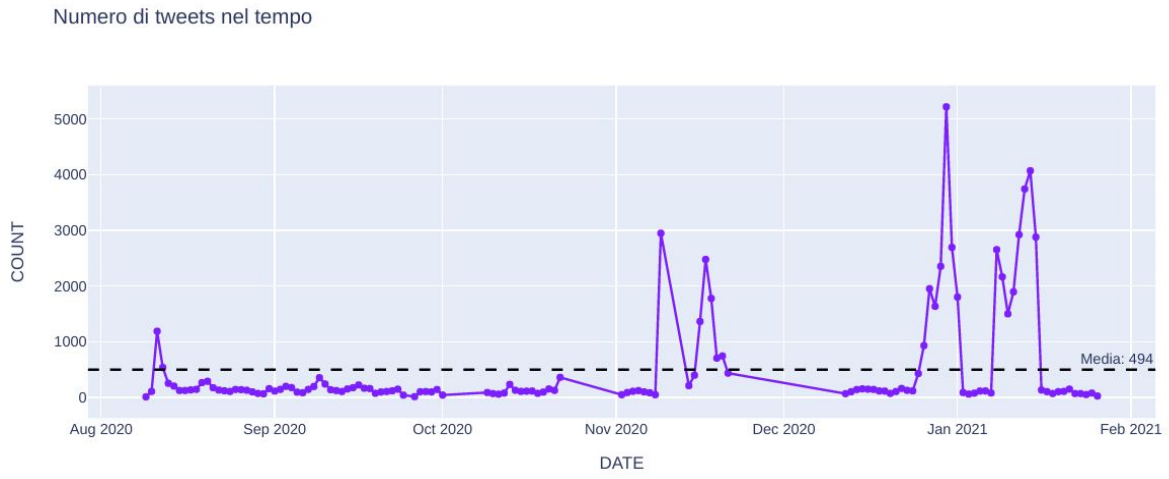
Tra i campi a disposizione riguardanti ciascun tweets, l'analisi è stata focalizzata sulla sorgente (software con cui è stato pubblicato), la data di pubblicazione, il testo e gli hashtags.

- **Da quali sorgenti software provengono i tweets?**



La maggiorparte dei tweets provengono dall'applicazione web di Twitter (desktop) e dall'applicazione per iPhone (mobile).

- **Com'è l'andamento dei tweets nel tempo? Quanti tweets vengono prodotti mediamente al giorno?**



Mediamente sono stati prodotti **494 tweets al giorno** nel periodo temporale che va dal 9 Agosto 2020 al 26 Gennaio 2021.
Dal grafico si può osservare la presenza di notevoli spike nel numero di tweets riguardanti i vaccini. La spiegazione di questi picchi è spiegabile osservando il contenuto dei tweet, in particolare:

- **Agosto 2020** → la Russia inizia a registrare il vaccino Sputnik mentre altre case farmaceutiche iniziano le fasi avanzate di sperimentazione

date	text
2020-08-10	The world's best biopharmaceutical researchers are in a race to develop an effective Covid-19 vaccine. But it's als… https://t.co/KYYXql8o2a
2020-08-10	#Covid_19 #CovidVaccine #coronavirus It's certain this fellow Anthony Fauci is an agent of some US Pharma agent. If… https://t.co/gtH9lPy5In
2020-08-10	#Russia is all set to launch the world's first #coronavirusvaccine on August 12 #COVID19 #coronavirus #Corona… https://t.co/OJqViJ8JJ1
2020-08-10	Atrimed Pharma Plans Lecture Series "An Hour with an Expert" on AtrimedX Platform #COVID19 #CovidVaccine Read… https://t.co/SSJUaDTzYu
2020-08-10	#SerumInstitute will start manufacturing vaccines by the end of August, said Poonawalla. #COVID19 #CovidVaccine… https://t.co/zPcFviOdxZ

- **Novembre 2020** → iniziano i primi ordini e le prime consegne in varie nazioni, vengono inoltre rilasciate le informazioni circa l'efficacia dei vaccini di Pfizer e Moderna

date	text
2020-11-17	Moderna and Pfizer with their 90%+ vaccines #CovidVaccine https://t.co/8qF4ABrBFI
2020-11-17	County by county breakdown. #COVID19 #California https://t.co/1VVsqFIoVe Hopeful for the #CovidVaccine especially n… https://t.co/lMovVdrR0s
2020-11-17	#coronavirus #COVID19 #covid #CovidVaccine #covidvaccines #Pfizer #Pfizervaccine #Moderna #modernavaccine Please… https://t.co/Jey0zMWtBY
2020-11-17	#VaccinesWork but need time to discover the best and safe kind of them! #COVID19 #CovidVaccine https://t.co/fU4fuhTfG3
2020-11-17	#CovidVaccine #Moderna trending.. Corona virus be like :- https://t.co/m7j4ReVteJ

- **Dicembre 2020** → a seguito dell'approvazione dei principali enti del farmaco internazionali, molte nazioni iniziano la campagna vaccinale. Molti utenti iniziano a pubblicare informazioni sul proprio stato di salute dopo aver ricevuto la dose, altri riportano diffidenza. Da notare inoltre, che nel periodo delle festività Natalizie l'attività sul social aumenta

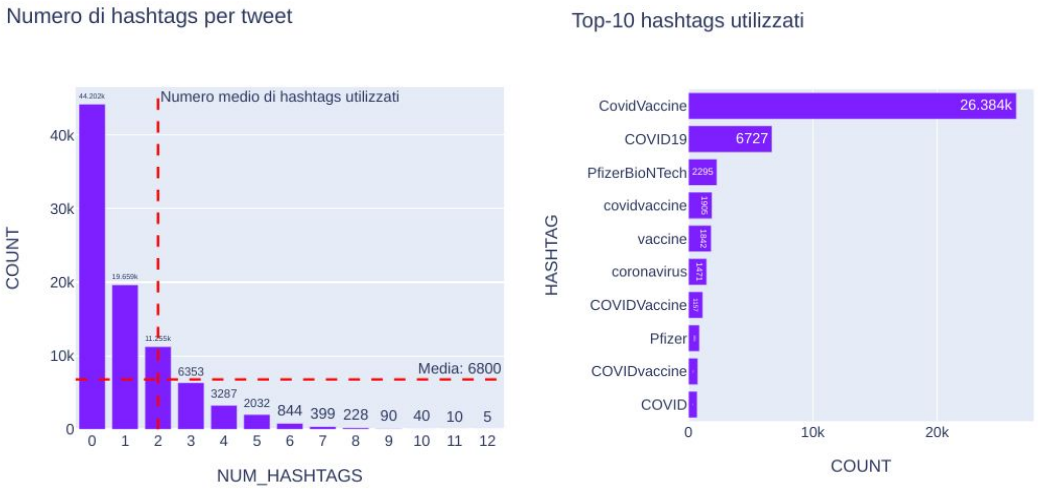
date	text
2020-12-28	After the #CovidVaccine ... via @NYTimes https://t.co/vMxRf8RIPi
2020-12-28	Not only was my results negative I'm considered high risk for Covid because I have chronic bronchitis. #CovidVaccine
2020-12-28	Taking one for the team. #CovidVaccine https://t.co/dxNGFlXNlQ
2020-12-28	Would you risk your health, wellbeing, mind, body and soul just to go on a quick cheap holiday? I reckon loads woul… https://t.co/oTj0s7Szt7
2020-12-28	Vaccinated #CovidVaccine #healthcare https://t.co/K5SMlydy8Z

- **Gennaio 2021** → la campagna vaccinale globale inizia in quasi tutti gli stati che hanno siglato accordi con i principali produttori di vaccino, molti utenti iniziano a pubblicare tweet riguardo l'avvenuta somministrazione delle prime dosi

date	text
2021-01-12	#CovidVaccine Didnt feel a thing and 4 hours later with no side effects Stay Home and Protect the NHS https://t.co/ZqHoksvxcE
2021-01-12	I got vaccination today. Will report back if I mutate, but so far no new superpowers. #CovidVaccine #Nursing
2021-01-12	Hearing elders from northern states flying to AZ to get vaccinated b/c it's available. How widespread is getting o… https://t.co/XiGXkdewwA
2021-01-12	Dose 1 of the #CovidVaccine is in the bag! GFY #COVID19! https://t.co/tQiuUTxrLr
2021-01-12	#COVID19 #VaccineHesitency is prevalent among #PoC because of discrimination in health care settings. Kweku Hazel,… https://t.co/Ijyo5BjSnT

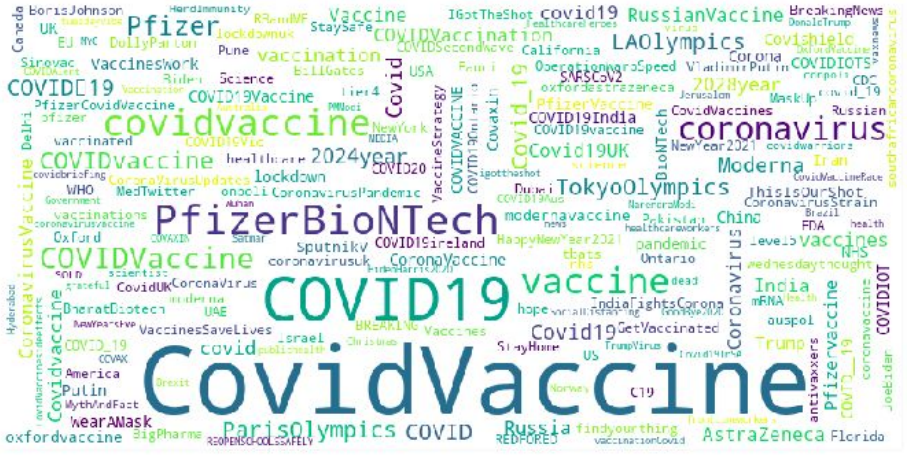
In generale non vi è un trend esplicito, dopo gli spikes di Dicembre e Gennaio il numero di tweets subisce un netto calo. Questo può essere dovuto alla conformazione del dataset, estratto solamente sulla base degli hashtags.

- **Quanti hashtags vengono utilizzati mediamente in ciascun tweet? Quali sono i principali?**



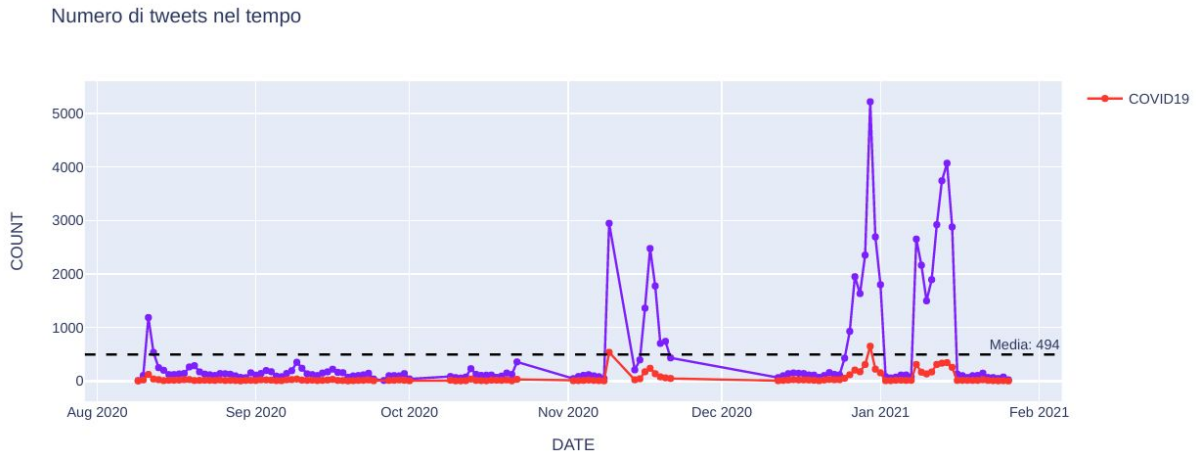
La maggior parte dei tweet presenti, circa 44 mila, non contiene alcun hashtag. Da notare però che in questo numero sono presenti anche parte dei tweets del dataset “Covid Vaccine Tweets”, in quanto l’autore descrive il campo hashtags indicando che contiene gli altri hashtags presenti nel tweet oltre a quello utilizzato per l’estrazione (#CovidVaccine). Non viene però specificato esplicitamente dall’autore se a tutti i tweets è stato rimosso dal campo hashtags questo specifico hashtag.

Mediamente sono stati utilizzati **2 hashtags**. Quelli più utilizzati sono visibili nel grafico sovrastante e nel WordCloud.
Oltre agli hashtags riferiti genericamente al vaccino si possono riconoscere i principali vaccini attualmente disponibili: “PfizerBioNTech”, “AstraZeneca”, “Moderna” e “Sputnik”.
Interessante la presenza di hashtags come “COVIDIOTS” e “stayhome” che indicano chiaramente una polarizzazione sull’argomento Covid19.
Da notare anche la presenza di hashtags quali “borisJohnson”, “Trump”, “Putin” (tema politico di rilievo nel periodo temporale coperto dal dataset).



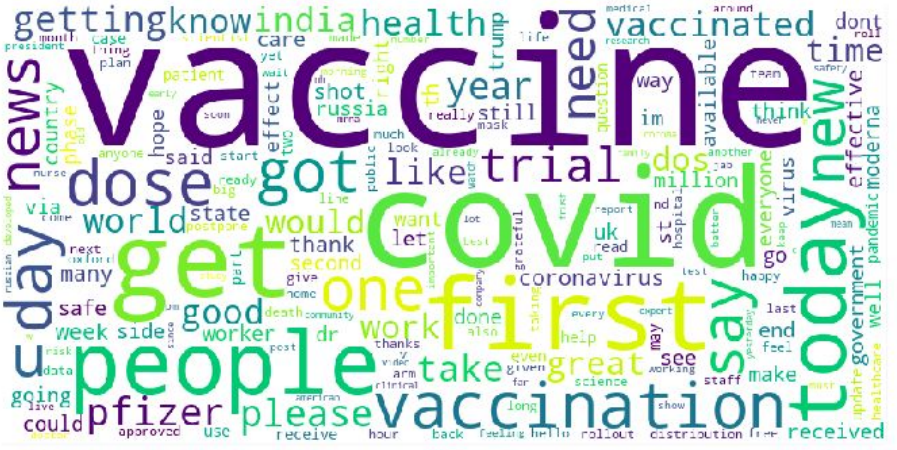
(WordCloud hashtags)

Si può notare anche come i più utilizzati seguano l’andamento temporale del numero di tweets.



(esempio andamento temporale del numero di tweets e del numero di tweets aventi tra gli hashtags #COVID19)

- **Quante parole vengono utilizzate mediamente in ciascun tweet? Quali sono le parole più ricorrenti?**



(WordCloud tokens)

Il numero medio di parole nel campo text non tokenizzato è di **17 parole**, dopo la tokenizzazione il numero di token medio estratto è di **8 tokens**.
Nel WordCloud sono riportate i principali tokens presenti nel dataset. Quelli più rilevanti riguardano sicuramente il vaccino ed il covid.
Interessante la presenza di altre parole come “dont”, “going”, “please”, “received”, “great”, “thank”, “effect” che possono avere un significato preciso per il sentiment di ciascun tweet.
Le parole “one”, “second”, “first”, “dos”, “today”, “getting” possono invece indicare la presenza di molti tweets riferiti alla vaccinazione degli utenti e la loro conoscenza di come la campagna vaccinale deve procedere.

Sentiment Analysis

Non avendo a disposizione dei dati etichettati con il sentiment è stato necessario estrarre questa informazione. Tra le possibilità presenti in letteratura sono state utilizzate due metodologie diverse: una più classica, basata su lessico, ed una facente uso di modelli di NLP più avanzati. Dalla comparazione dei risultati ottenuti è quindi stato scelto uno dei due metodi per etichettare il dataset ed effettuare l'analisi.

Sentiment extraction

Due metodi utilizzati:

- VADER-Sentiment-Analysis [1]
- Twitter-RoBERTa-sentiment [2]

VADER-Sentiment-Analysis

Vader-Sentiment-Analysis è un tool per la sentiment analysis basato su lessico e regole, specificatamente creato estrarre il sentiment (positivo, neutrale o negativo) da testi provenienti da social media.

Per come è stato costruito è in grado di riconoscere elementi specifici di questo tipo di testi come punteggiatura, emoji ed emoticon, negazioni, utilizzo di maiuscole/minuscole, slang, iniziali e acronimi - per la lingua inglese.

Per questo motivo il preprocessing definito sul testo dei tweets va a rimuovere unicamente i links, le menzioni e gli hashtags.

Twitter-RoBERTa-sentiment

Twitter-RoBERTa-sentiment è un modello di NLP avanzato, basato sul modello RoBERTa che a sua volta è costruito sul più conosciuto modello BERT.

Nello specifico, il modello base di RoBERTa è stato addestrato su un dataset contenente quasi 58 milioni di tweets di vario tipo e successivamente è stato specializzato, tramite fine-tuning, per il task di sentiment analysis su un dataset di tweets etichettati manualmente con il sentiment.

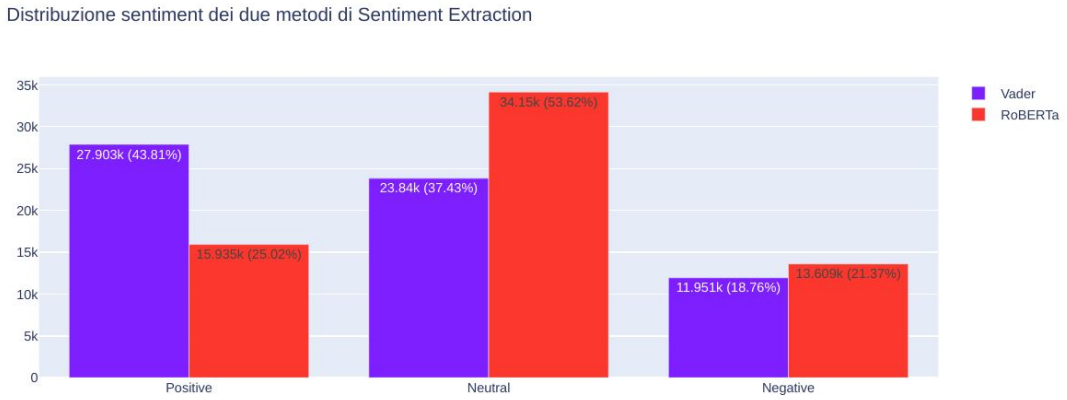
Il modello, fornito insieme ai pesi di addestramento, è utilizzabile direttamente sui testi dei tweets applicando solo delle "maschere" sui links e sulle menzioni presenti.

text	score_vader	sentiment_vader	score_roberta	sentiment_roberta	comparison
Same folks said daikon paste could treat a cytokine storm #PfizerBioNTech https://t.co/xeHhIMg1kF	0.4019	Positive	0.8346	Neutral	Differents
While the world has been on the wrong side of history this year, hopefully, the biggest vaccination effort we've ev… https://t.co/dlCHrZjkhm	-0.1027	Negative	0.4959	Neutral	Differents
#coronavirus #SputnikV #AstraZeneca #PfizerBioNTech #Moderna #Covid_19 Russian vaccine is created to last 2-4 years… https://t.co/ieYlCKBr8P	0.25	Positive	0.8145	Neutral	Differents
Facts are immutable, Senator, even when you're not ethically sturdy enough to acknowledge them. (1) You were born i… https://t.co/jqgV18kch4	0	Neutral	0.6711	Neutral	Equals
Explain to me again why we need a vaccine @BorisJohnson @MattHancock #whereareallthesickpeople #PfizerBioNTech… https://t.co/KxbSRoBEHq	0	Neutral	0.6261	Neutral	Equals

« ‹ 1 / 12739 › »

(esempio di score e sentiment estratti dai due metodi)

Dopo aver estratto il sentiment da ciascuno dei tweets utilizzando entrambi i metodi sono state condotte delle analisi con lo scopo di capire quali tra i due fosse, indicativamente, il migliore per il task affrontato.
Come si può osservare dalla tabella sovrastante e dalla distribuzione del sentiment (positivo, negativo o neutrale) riportata in figura, vi sono delle differenze, specialmente nella interpretazione tra sentiment positivo e neutrale:



- RoBERTa interpreta la maggior parte dei tweets come neutrali
- Vader invece interpreta la maggior parte dei tweets come positivi

Il numero di etichette di sentiment differenti è di 38451. La differenza tra i risultati può essere attribuita essenzialmente alle soglie che delimitano la classificazione del testo tra positivo e neutrale. Di fatto lo score indicato da Vader varia tra 0 ed 1 e l'etichetta (positivo, negativo, neutrale) viene scelta tramite delle soglie consigliate dagli autori del metodo.

Considerando invece le differenze nella polarizzazione, quindi tra tweets classificati come positivi/negativi o viceversa, ve ne sono 3485.

text	score_vader	sentiment_vader	score_roberta	sentiment_roberta
@DrTonyLeachon @ralphrecto why this government insisting the #Sinovac where in fact less efficacy and safety - wors… https://t.co/t5bSPuI7Vt	0.3687	Positive	0.7503	Negative
Today we aimed higher and broke our own record! Jackson administered 1,278 #PfizerBioNTech vaccinations, bringing o… https://t.co/PidTky8n7g	-0.4753	Negative	0.7059	Positive
Yes I got the #COVID19 vaccine tonight and no I'm not flexing! Grateful to receive it. #PfizerBioNTech… https://t.co/LEa0dm3OZe	-0.312	Negative	0.9631	Positive
24 new #Covid19 deaths in #BC as the province reports 640 new cases. 9, 950 active cases in #BC -New type of measu… https://t.co/htz9Xvbjvb	0.4019	Positive	0.5649	Negative
Not a good start. Hope she recovers to full health. But this is one of the concerns of a rushed vaccine. #COVID19… https://t.co/krBN63JZ5D	0.0636	Positive	0.8047	Negative

« ‹ 4 / 697 › »

(campione di tweets classificati in maniera opposta dai due metodi)

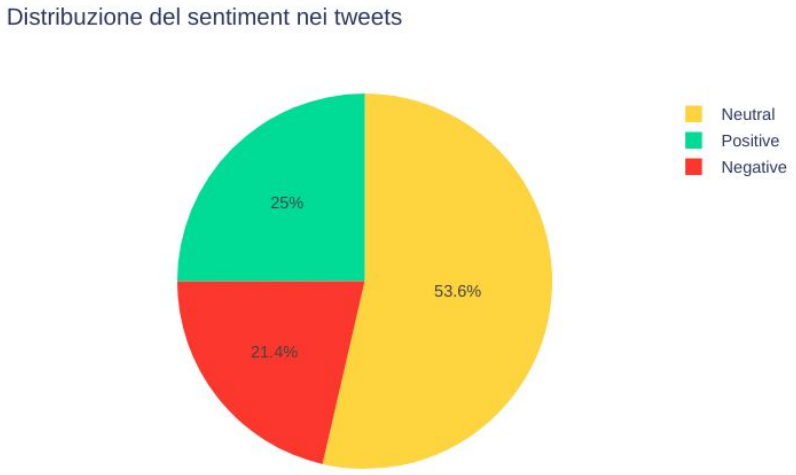
La scelta di quale sentiment estratto utilizzare, rispetto ai metodi utilizzati, è stata presa essenzialmente dall'analisi manuale del contenuto dei tweets aventi una classificazione opposta.
Come si può osservare dal campione riportato il sentiment estratto tramite RoBERTa risulta molto più efficace nel comprendere il sentiment associato ai tweets riferiti ad un argomento così specifico come la vaccinazione.
Per questo motivo, è stato scelto di utilizzare il sentiment estratto tramite RoBERTa nella parte di analisi.

Il punto negativo di questo metodo è la velocità di elaborazione, rispetto a Vader che impiega pochi secondi per estrarre il sentiment, la velocità di classificazione di RoBERTa varia in base all'hardware utilizzato.
Nella configurazione utilizzata (core i7, 8GB ram, SSD) la classificazione del sentiment dell'intero dataset ha richiesto più di 30 minuti di tempo.

Sentiment study

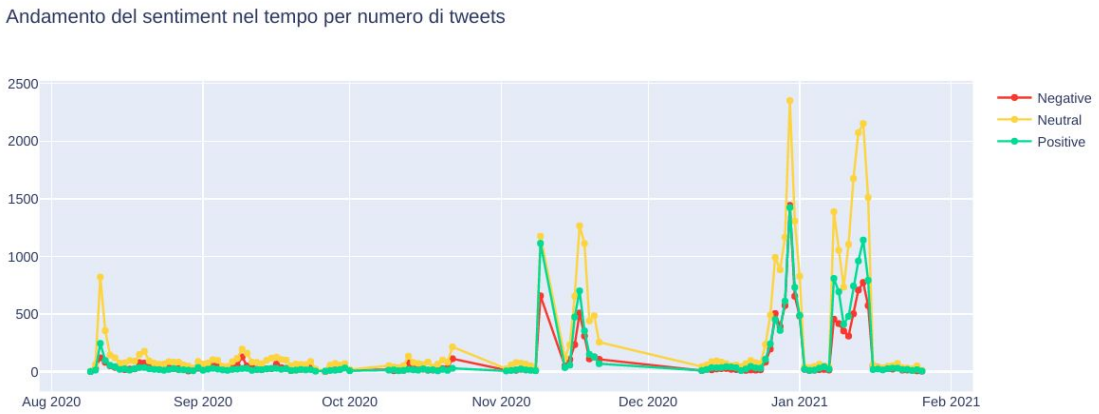
Dopo aver ottenuto il sentiment sono stati condotti una serie di studi, estendendo la parte di EDA con il sentiment.

- **Com'è distribuito il sentiment nei tweets?**

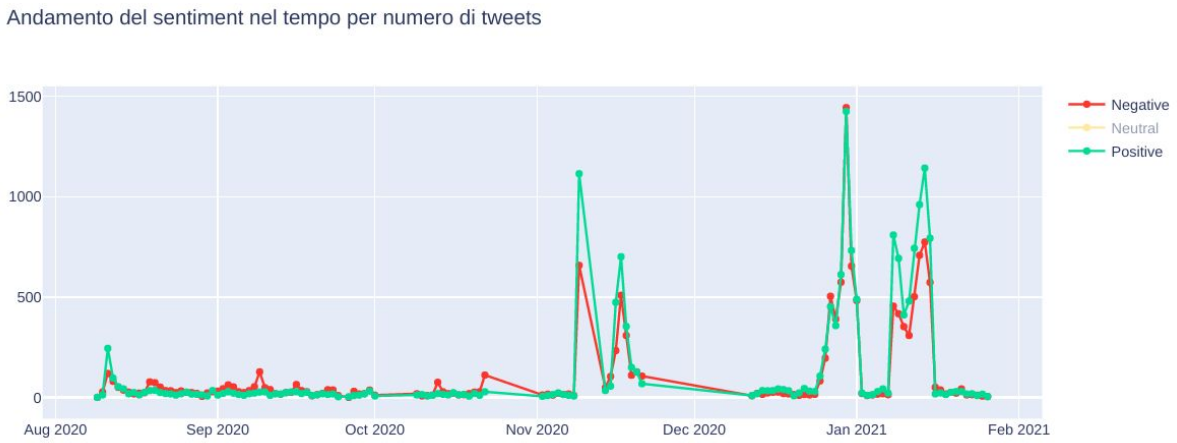


Più della metà dei tweets risulta avere un sentiment neutrale mentre non vi è una sostanziale differenza nel numero di tweets positivi e negativi.

- **Com'è l'andamento del sentiment nel tempo?**



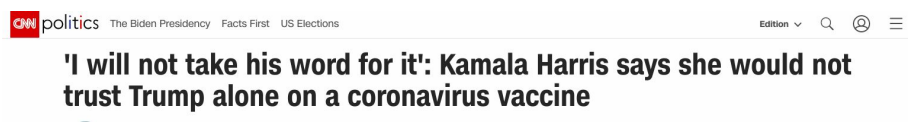
Analizzando il grafico si può notare come il numero di tweets neutrali sia sempre superiore rispetto ai tweet positivi e negativi, ma questo lo si spiega facilmente dato che il numero di tweets neutrali rappresenta la metà del dataset.



Più interessante sono i punti in cui il numero di tweets negativi ha superato il numero di tweets positivi e viceversa.
Riguardo ai picchi di tweets positivi, sono concentrati negli stessi periodi temporali analizzati nella EDA.
Riguardo ai picchi negativi invece, i più importanti sono collocati:

- **Settembre 2020** → molti tweets classificati come negativi riguardano le dichiarazioni della neo eletta vicepresidente degli Stati Uniti d'America (al tempo candidata).

date	text	sentiment
2020-09-06	Worked well with @Boeing self regulating right? #737Max I'm not a fan of the for profit honor system.… https://t.co/9w40XqpVsg	Negative
2020-09-06	@karol @KamalaHarris is a disgusting hack, now against #CovidVaccine	Negative
2020-09-06	#KamalaHarris Says, 'Wouldn't Trust #Trump On Safety Of #CovidVaccine Released Before #USElections… https://t.co/emwl0z0AOi	Negative
2020-09-06	'Wouldn't Trust #Trump On Safety Of #CovidVaccine Released Before #USElections @KamalaHarris #CoronavirusVaccine #USA #America	Negative
2020-09-06	@SenSchumer Sorry you're SO out of it.. GUESS WHAT, Operation WARP SPEED are working NIGHT AND DAY TO COME UP WITH… https://t.co/lQQuHMto6h	Negative



In aggiunta, nello stesso periodo, molti giornali internazionali riportano la notizia di una epidemia di polio scoppiata in Sudan a seguito dell'inizio delle vaccinazioni

- **Ottobre 2020** → la casa farmaceutica Johnson&Johnson annuncia la sospensione della terza fase di sperimentazione del proprio vaccino a causa di alcune anomalie nella salute dei volontari. Gran parte dei tweets negativi sono riferiti a questa notizia, indicando come il vaccino possa essere dannoso per la salute delle persone.

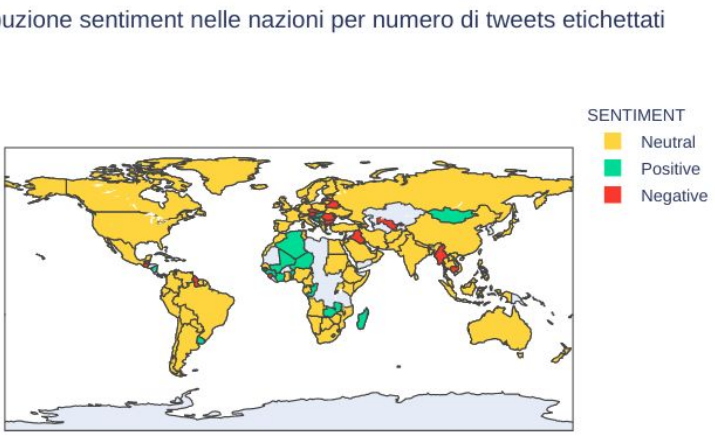
date	text	sentiment
2020-10-13	Johnson and Johnson halts COVID-19 vaccine trial temporarily after participant falls ill https://t.co/cZnUMff7Em… https://t.co/X2ZbZxdVYI	Negative
2020-10-13	#JohnsonAndJohnson pauses #CovidVaccine trial as participant becomes ill https://t.co/CmlBvjwHZH	Negative
2020-10-13	Needless to say, these pharmaceutical companies are racing to develop the vaccine and people will suffer in the pro… https://t.co/gzmaigcVFP	Negative
2020-10-13	So much for #CovidVaccine ...Johnson & Johnson just announce it has suspended 3rd stage trials as someone who was t… https://t.co/92U0nFto57	Negative
2020-10-13	@RebeccaChandle1 They couldn't even make safe talcum powder-forget about a vaccine 😍🙈👏 #johnsonandjohnson #vaccines #CovidVaccine	Negative

- **Gennaio 2021** → i giornali internazionali riportano la notizia della morte di 23 persone in Norvegia in seguito alla somministrazione del vaccino. Oltre alla Norvegia iniziano ad arrivare altre notizie di persone decedute in seguito a reazioni allergiche in varie parti del mondo.

date	text	sentiment
2021-01-16	#Norway has witnessed deaths of 23 elderly #people soon after #CoronaVaccine named #PfizerBioNTech . Norway is… https://t.co/H3jbFVQJaO	Negative
2021-01-16	Covid-19: Norway investigates 23 deaths in frail elderly patients after vaccination #PfizerBioNTech #COVID19 https://t.co/tZHbFzbSsp	Negative
2021-01-16	@SwainITV @piersmorgan @susannareid100 @BorisJohnson needs to get a grip of this now. He should be sorting this ton… https://t.co/CzjJkcQmAF	Negative
2021-01-16	@BorisJohnson 23 dead in Norway #PfizerVaccine 10 people dead in Germany #PfizerVaccine and your going to protect… https://t.co/bS2C7OjBfs	Negative
2021-01-16	55 people in the US have died after receiving a #Vaccine for #COVID19, according to reports submitted to a federal… https://t.co/bxezEVXQX8	Negative

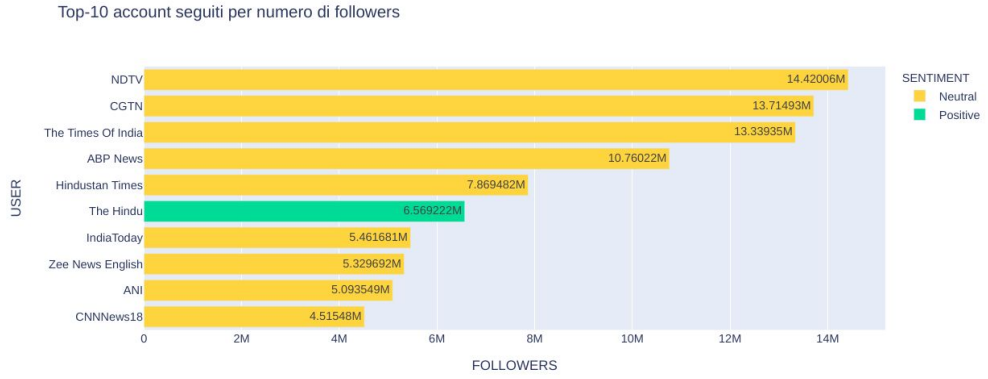
In generale si può notare come determinati eventi abbiano avuto impattato il sentiment dei tweets.

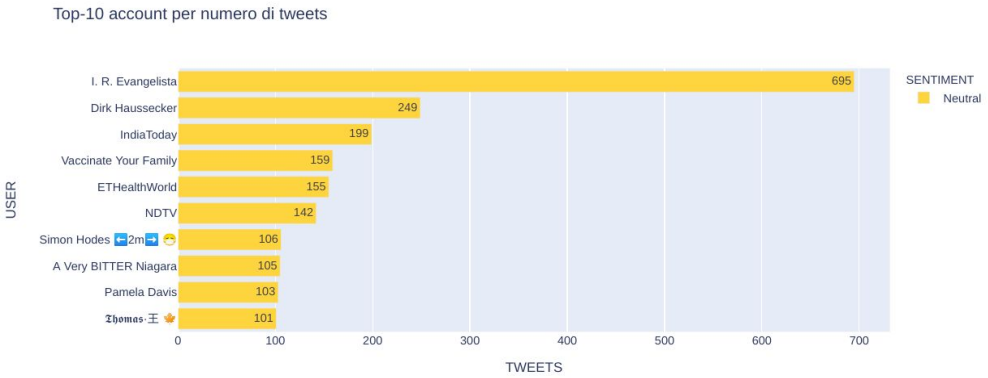
- **Com'è distribuito il sentiment nel mondo? Vi sono nazioni da cui provengono tweets con un sentiment prevalente?**



La maggioranza delle nazioni presenti, che sono state identificate nella fase di preprocessing, presentano un numero maggiore di tweets con sentiment neutrale. Le nazioni più polarizzate rappresentate nel grafico non sono significative in quanto contengono numeri molto bassi di tweets.

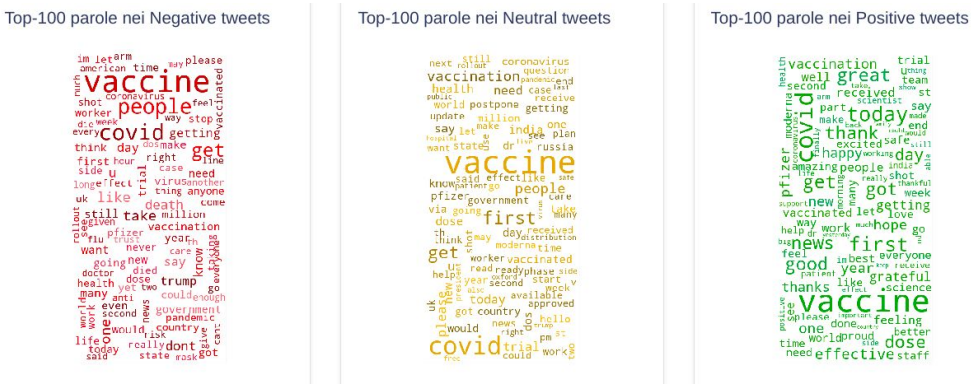
- **Quale sentiment caratterizza, per numero di tweets, gli utenti con più followers? Quale invece caratterizza gli utenti con più tweets?**



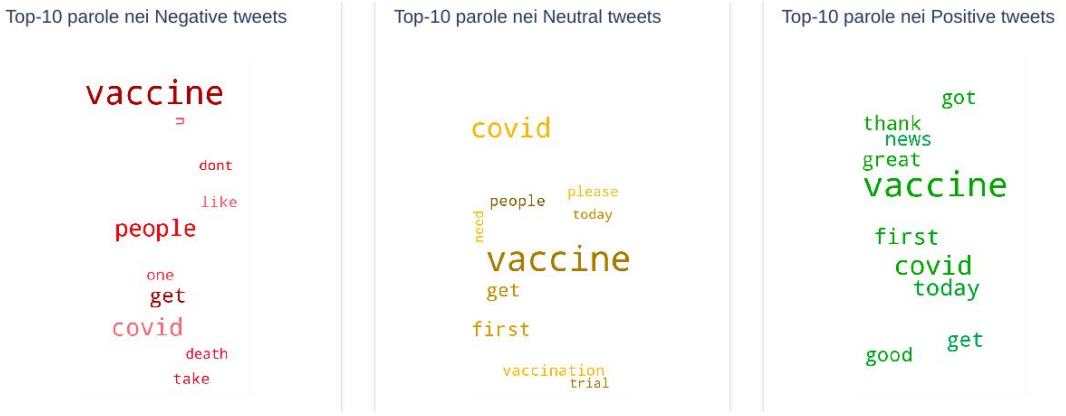


Sia per gli utenti con il più alto numero di followers che per quelli con il più alto numero di tweets, il sentiment maggiormente espresso è quello neutrale.

- **Quali sono le parole e gli hashtags più utilizzati rispetto al sentiment del tweet?**



Visualizzando tramite WordCloud le parole (tokens) più utilizzate in base al sentiment si riesce ad intuire come il processo di sentiment extraction sia stato efficace nell'astrazione del sentimento.



Osservando la top-10 risulta ancora più evidente come il sentiment positivo e negativo sia stato riconosciuto, in particolare nelle parole positive si può intuire come i tweet positivi siano riferiti alla somministrazione del vaccino o della prima dose del vaccino.

date	text	sentiment
2021-01-17	Who'd have thought that I'd one day be vaccinating with my mum? It feels amazing to be doing our bit to help, but I… https://t.co/6PZlaZD3mA	Positive
2021-01-17	Just got vaccinated. 🙏 #CovidVaccine #PfizerBioNTech	Positive
2021-01-17	Many thanks to the vaccination team at York Hospital. #ProtectTheNHS #PfizerBioNTech https://t.co/YTJtON9P3L	Positive
2021-01-17	Have the new jab Hallelujah #PfizerBioNTech #Pfizer https://t.co/FiXV6Qd92K	Positive
2021-01-17	I'm 10 days post my first #PfizerBioNTech vaccine @HullHospitals. Hopefully I'm now developing antibodies.	Positive

date	text	sentiment
2020-12-15T10:49:24	Option C Paul. The vaccine is toxic and unsafe. #pmlive @PMOnAir #auspol #AntiVaccine #Covid_19 #SARSCoV2… https://t.co/yr6YrgVwOM	Negative
2020-12-15T07:51:24	Iran just refused to receive #PfizerBioNTech vaccine for #COVID-19! You know why? Just because does NOT have the fa… https://t.co/DP1o2Mnk8O	Negative
2020-12-15T04:32:13	Now Bhakts all around the world should also boycott #COVID19Vaccine developed by #PfizerBioNTech. @zoo_bear… https://t.co/igtuWw5s2H	Negative
2020-12-15T02:46:35	Vaccines have arrived in #AB . My Mum has been in isolation for 10 days. 18 residents sick. 2 have passed away in… https://t.co/7IzvrleXPc	Negative
2020-12-15T01:57:10	@liberal_party #PfizerBioNTech =NO LIABILITIES #Pfizervaccine = UNTESTED #Pfizer = #CCP #Infiltration… https://t.co/hzvsPm98DW	Negative

Meno evidente risulta invece il sentiment neutrale, probabilmente poiché le parole di maggior rilievo sono presenti nei tweets di enti giornalistici o governativi.

date	text	sentiment
2021-01-18	Update - 30 hours post 2nd shot started to feel chilly and then overheated. Checked my temp and it’s 100.2 - low… https://t.co/qGj5FCE38o	Neutral
2021-01-18	#COVIDVaccination #PfizerBioNTech So far, more than 20mn people worldwide have received the Pfizer vaccine and hav… https://t.co/LoYW1Q401f	Neutral
2021-01-18	Is COVID-19 Vaccine, Developed by Pfizer and BioNTech, Safe? Deaths Among Elderly People After Taking First Shot in… https://t.co/kNb19dCqt9	Neutral
2021-01-18	The Pfizer vaccine was granted emergency use authorization on Dec. 11 while the Moderna vaccine was given the same… https://t.co/CpQELUhcFz	Neutral
2021-01-18	#CovidVaccine #PfizerBioNTech Get the vaccine when its your turn!#Hope2021 https://t.co/Os3kCZ9gyH	Neutral

Top-10 hashtags nei Negative tweets

Top-10 hashtags nei Neutral tweets

Top-10 hashtags nei Positive tweets

Per quanto riguarda gli hashtags non si riesce ad evincere alcuna differenza poiché i tre sentiment classificati condividono praticamente gli stessi hashtags.

Conclusioni

Non avendo a disposizione un dataset etichettato o parzialmente etichettato da persone, l’analisi del sentimento è stata condotta utilizzando degli approcci che, seppur sofisticati, sono soggetti ad errore.
Alcuni esempi:

2020-12-15T19:02:00	The #PfizerBioNTech #CovidVaccine does not contain a live virus. The vaccine cannot and will not cause COVID19 disease. Not even possible.	Negative
2020-12-16T05:59:13	It’ll be interesting to see the post-vaccine birth rates next winter. 😂 #COVID19 #PfizerBioNTech	Positive

Un’altro problema è la presenza di un alto numero di tweets classificati come neutrali, più complicati da analizzare.
Dai risultati ottenuti dalla sentiment extraction si è potuto comunque intuire come un modello di Deep Learning appositamente addestrato per questo tipo di dati sia più efficace.

Sviluppi futuri

- Sentiment extraction utilizzando tecniche non supervisionate (BERT embeddings e clustering), già impostato ma per motivi di tempo non effettuato.
- Sarcasm detection e combinazione con la sentiment analysis, potrebbe essere interessante capire se tra i tweets positivi e negativi sono presenti delle affermazioni sarcastiche riguardo i vaccini.

- Miglioramento interattività e velocità della dashboard, le potenzialità fornite da Dash Plotly sono numerose, sarebbe utile sfruttare al meglio le visualizzazioni e migliorare la velocità della dashboard spostando il carico computazionale “offline”.

Andamento della pandemia

Avendo a disposizione anche dei dati pubblici globali sull’andamento della pandemia e delle campagne vaccinali, in seguito alla sentiment analysis condotta sul dataset dei tweets, uno degli sviluppi futuri molto interessanti sarebbe lo studio di eventuali correlazioni con questi dati.
In particolare:

- rispetto all’andamento della pandemia (casi confermati, ricoveri e decessi) vi è una correlazione con il numero ed il relativo sentiment dei tweets?
- come l’andamento della campagna vaccinale si riflette nei tweets? Rispetto al sentiment vi è qualche variazione?

I dataset descritti sono già stati inseriti nel progetto ma per motivi di tempo non sono stati usati.

Datasets

Due datasets principali:

- COVID-19 World Vaccination Progress[7]
- COVID-19 Data Repository[8]

Entrambi i dataset sono costantemente aggiornati su base giornaliera aggregando molte fonti autorevoli pubbliche, governative o meno, delle principali nazioni del mondo.
Le versioni utilizzate per il progetto risalgono al 26 Gennaio 2021.

COVID-19 World Vaccination Progress
Contiene informazioni sull’andamento delle campagne di vaccinazione delle principali nazioni del mondo che rendono disponibili i propri dati **[5]**.
Tra i campi presenti sono stati utilizzati: data, codice iso della nazione, numero totale di vaccinati e vaccino utilizzato.

COVID-19 Data Repository
Contiene informazioni sull’andamento dell’epidemia delle principali nazioni del mondo. Tra i numerosi dataset a disposizione sono stati utilizzati quelli relativi all’andamento temporale a livello globale (JHU CSSE COVID-19 Dataset **[6]**). Questi dataset contengono informazioni circa: data, nazione, numero di contagi accertati, numero di ricoveri e numero di morti.

[7] https://www.kaggle.com/gpreda/covid-world-vaccination-progress
[8] https://github.com/CSSEGISandData/COVID-19

Citazioni

[1] Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

[2] Rosenthal, Sara, Noura Farra, and Preslav Nakov. "SemEval-2017 task 4: Sentiment analysis in Twitter." Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017). 2017.

[3] Bird, Steven, Edward Loper and Ewan Klein (2009). Natural Language Processing with Python. O'Reilly Media Inc.

[4] Wolf, Thomas, et al. "Transformers: State-of-the-art natural language processing." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020.

[5] Hasell, J., Mathieu, E., Beltekian, D. et al. A cross-country database of COVID-19 testing. Sci Data 7, 345 (2020). https://doi.org/10.1038/s41597-020-00688-8

[6] Dong, Ensheng, Hongru Du, and Lauren Gardner. "An interactive web-based dashboard to track COVID-19 in real time." The Lancet infectious diseases 20.5 (2020): 533-534.