UNIVERSITÀ DEGLI STUDI DI TRIESTE

Natural Language Processing project

sentiment analysis regarding tweets about covid-19 vaccines

Giovanni Pinna

Anno Accademico 2020-2021

Sommario

1 Problema

L'obiettivo e quello fare l'analisi del sentiment che ha la popolazione italiana riguardo i vari vaccini in commercio contro il Covid-19.
In particolare i vaccini che sono stati presi in considerazione in questa analisi sono quelli approvati ad oggi dall'EMA (Pfizer, Astrazeneca, Moderna) più lo Sputnik.

Si è deciso di prendere in considerazione anche il vaccino Sputnik (Russo) per osservare se il sentiment si discosti significativamente in confronto ai vaccini approvati dell'Italia e dell'UE.

I dati sono stati presi dal social network Twitter il quale, con un account developer, permette di scaricare i tweet e i dati ad essi associati abbastanza facilmente.

Per identificare la popolazione italiana si è deciso di assumere come gli utenti italiani siano tutti quelli che hanno scritto un tweet italiano e che hanno come luogo del profilo una città italiana o lo stato italiano.

Da questa analisi ci aspettiamo che i tweet siano la maggior parte contrassegnati da un sentimento negativo. Questo perché normalmente le persone tendono di più a lamentarsi sui social network rispetto che esprimere concetti positivi. In più, le lamentele suscitano sempre più risonanza e colpiscono di più l'utente che, di conseguenza, è più propenso a ricondividerle.

2 Datasets

I dati che sono stati usati provengono tutti dal social network Twitter. Questo perché tale social network permette di scaricare i tweet e altri dati, che vedremo in seguito, facilmente una volta che si è in possesso di un account developer.

Per scaricare i dati è stato necessario entrare in possesso dei vari token e chiavi segrete (accesstoken, accesstokensecret, apikey, apisecretkey) che vengono messe a disposizione una volta ottenuto l'account da developer. Una volta fatto ciò si è usata la libreria Python Tweepy che permette facilmente di scaricare i dati. In particolare per agevolarsi ancora di più il lavoro si è usata la funzione "Cursor" della libreria che scarica in automatico i tweet secondo i criteri di ricerca utilizzati (il programmatore non si deve occupare della paginazione dei tweet).

I criteri di ricerca utilizzati per scaricare i tweet che ci interessano sono:

data_since = (data di inizio della ricerca) settata al 18/04/2021
data_until = (data di fine della ricerca) settata al 28/04/2021
language = settato a 'it'
search_word = (parola chiave) settata dipendentemente dal vaccino in considerazione

Come detto in precedenza non ci siamo limitati a scaricare solo il testo del tweet, ma molti altri dati in modo da rendere, anche le future analisi, molto più complete e consistenti. In particolare le lables sono:

[tweet , is_quoted, tweet_quoted, name_of_who_I_answered , id user_str, lang_user, creation_date_of_tweet, result_type, number_of_retweet, result_type, number_of_retweet, retweeted, source, user_name, user_screen_name, location, user_description, number_of_follower, number_of_friend, is_verified, creation_date_of_account, like]

nella tabella A sono descritte con maggior dettaglio.

Una volta scaricati tutti questi dati si è passati alla pulizia di quest'ultimi.

3 Pre-processing

Il processo di pre-processing del testo è estremamente importante per rendere i documenti meno soggetti al rumore e alla varianza.

Per questo progetto si è deciso di applicare due tipi di pre-processing, uno che chiameremo hard e uno soft. Questo viene fatto perché non tutti gli algoritmi che andremo ad utilizzare avranno bisogno di un hard pre-processing.

Prima ancora di applicare il pre-processing è stato necessario eliminare le righe che presentavano dei tweet uguali. Questo è fatto perché molte persone semplicemente riproponevano un tweet senza alcun commento, quindi nel nostro dataset erano presenti molte righe con lo stesso tweet, retweetato, da due o più utenti diversi. Questa decisione è stata dettata da due fattori: il primo è quello non volevamo vare un assunzione, particolarmente importante, come quella che: se una persona retwitta un tweet assumiamo che pensi in maniera concorde al testo che ha retweetato. La seconda motivazione riguarda il costo computazionale, infatti lavorare ed elaborare una quantità enorme di tweet risultava pesante ed estremamente dispendioso in termini di tempo.

Quindi una volta eliminato tutte le righe che presentavano lo stesso identico tweet si è potuto procedere con il pre-processing vero e proprio.

3.1 Hard preprocessing

Per l'hard pre-processing si è deciso di eliminare tutti i nomi utenti, gli hashtag , link e caratteri speciali e tutti i numeri sono stati trasformati in 0 facendo uso delle regex. Prima di eliminarli però, questi dati sono stati salvati in delle apposite colonne che sono state utilizzate in fase di analisi.

In più la frase e stata anche processata ponendola in lower case e ogni parola si è ridotta al suo lemma. Si è utilizzata l'analisi del POS per tenere le parole che presentavano i seguenti attributi {'NOUN', 'VERB', 'ADJ', 'ADV', 'PROPN'}.

3.2 Soft pre-processing

Per quanto riguarda il soft pre-processing si è deciso di eliminare solo i link e i caratteri speciali. Ogni nome utente invece è stato sostituito con la parola "user".

Si è deciso, al contrario dell'hard pre-processing, di mantenere gli hashtag senza il carattere speciale "#" ipotizzando che in fase di allenamento degli algoritmi questi dati in più potessero aiutare in una buona predizione o una migliore comprensione del testo.

3.3 Altri tipi di pre-processing

Per le analisi che vedremo in seguito è stato necessario anche ripulire la colonna che riguarda la location dei vari profili utente. Questo perché erano presenti dei nomi che indicavano lo stesso luogo, ma scritti in maniera diversa (es. "Milano", "Milan", "Milano, Lombardia", "Milano, Italia" ;tutte queste combinazioni state collassate nella parola "Milano").
In questo caso sono state create delle espressioni regolari apposite per ogni città principale, poiché creare un espressione regolare per ogni città sarebbe risultato troppo oneroso e anche senza un gran senso dato che sono pochissimi i tweet raccolti nelle città non principali (come vedremo nella parte di analisi).

Altro pre-processing che è stato necessario fare è stato sulla colonna creata artificiosamente nel momento in cui abbiamo salvato i vari hashtag. Con essi infatti, sono stati salvati anche i caratteri speciali non utili e non significativi per le analisi. Quindi sempre tramite le regex questi caratteri speciali sono stati eliminati.

4 Sentiment analysis with BERT

Come primo modello ho deciso di utilizzare il modello bastato su BERT costruito dall'università Bocconi (feel-it). Questo modello ha la peculiarità di essere pre-allenato su dati in italiano e quindi utilizzabile sui tweet che ho scaricato.

Questo modello ha avuto il compito di fare la sentiment analysis sul testo dei tweet pulito con la procedura soft pre-processing. Infatti, questa rete neurale è stata creata per capire il linguaggio naturale e quindi non c'è stato bisogno di fare un hard pre-processing per diminuire il rumore e la varianza al massimo.

Il primo modello utilizzato è stato questo non solo per farsi un idea generale del sentiment nei vari tweet, ma anche per utilizzare le sue predizioni come base per altri modelli. Infatti le predizioni fornite dal modello della Bocconi son state usate come "true values" per tutti i modelli che vedremo in seguito.

Per scopi di analisi e per fare una verifica sulle predizioni del sentiment ho utilizzato anche l'analisi dell'emozione messa a disposizione sempre dallo stesso modello.

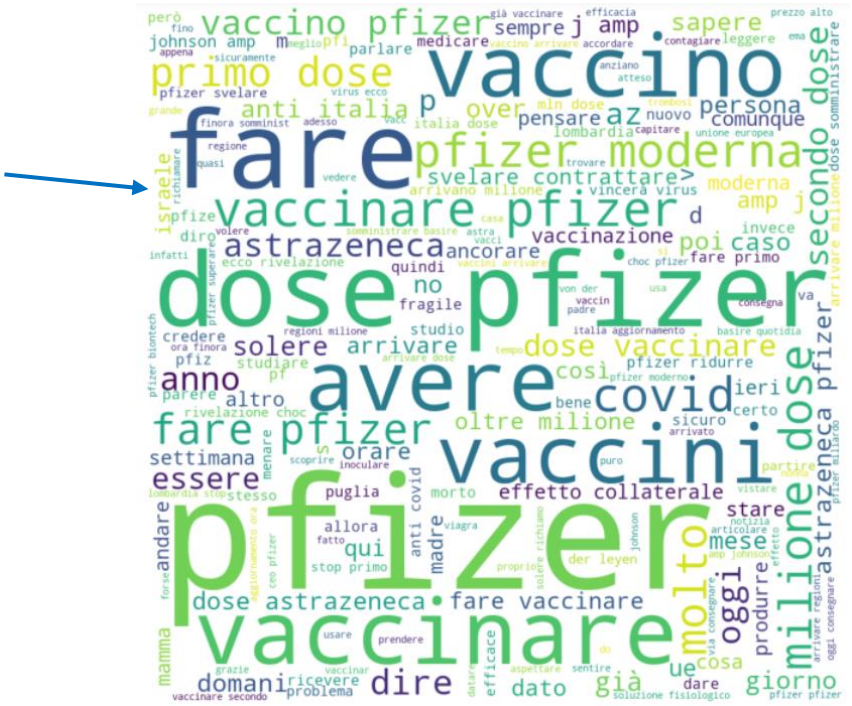
5 Analysis

A questo punto siamo in possesso del dataset che presenta due colonne in più: una per il sentiment e una per l'emozione.

Per prima cosa abbiamo deciso di vedere quali sono le parole che compaiono di più tramite una rappresentazione "WordCloud" . (Nella rappresentazione "WordCloud" non sono state prese in considerazione le stopwords).

5.1 WordCloud

5.1.1 Pfizer

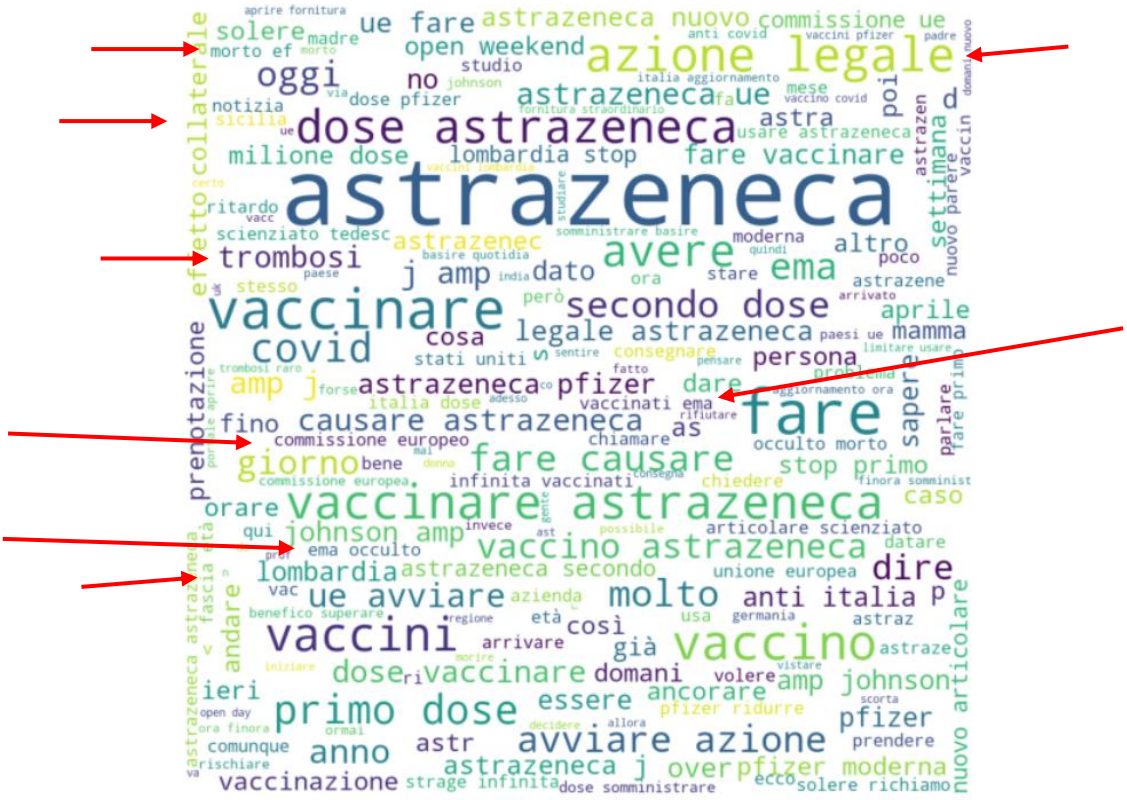


Come possiamo vedere dall'immagine si possono subito notare le parole che sono più grandi come corrispondano alle key word della nostra ricerca. Questo è normale perché sono tutte parole che saranno presenti in quasi ogni tweet. Per quanto riguarda Pfizer si possono notare le parole "milione dose" che fa palese riferimento alle scorte che dovrebbero arrivare all'Italia in questo periodo. Si

possono notare anche i nomi di altri vaccini, come il nome di Astrazeneca, Moderna e j amp j (johnson & johnson). Non ci sono delle parole che spiccano con un connotato palesemente positivo o negativo. Sembra si parli soprattutto: di fare le dosi, degli altri vaccini e dell'arrivo di uno stack di un milione di dosi.

Credo che sia importante notare (freccia blu) come sono state catturate delle parole anche che riguardano l'andamento della situazione pandemica e vaccinale in altri stati. Primo fra tutti Isdraele che ad oggi è lo stato più avanti nella vaccinazione e che ha usato soprattutto dosi del vaccino Pfizer.

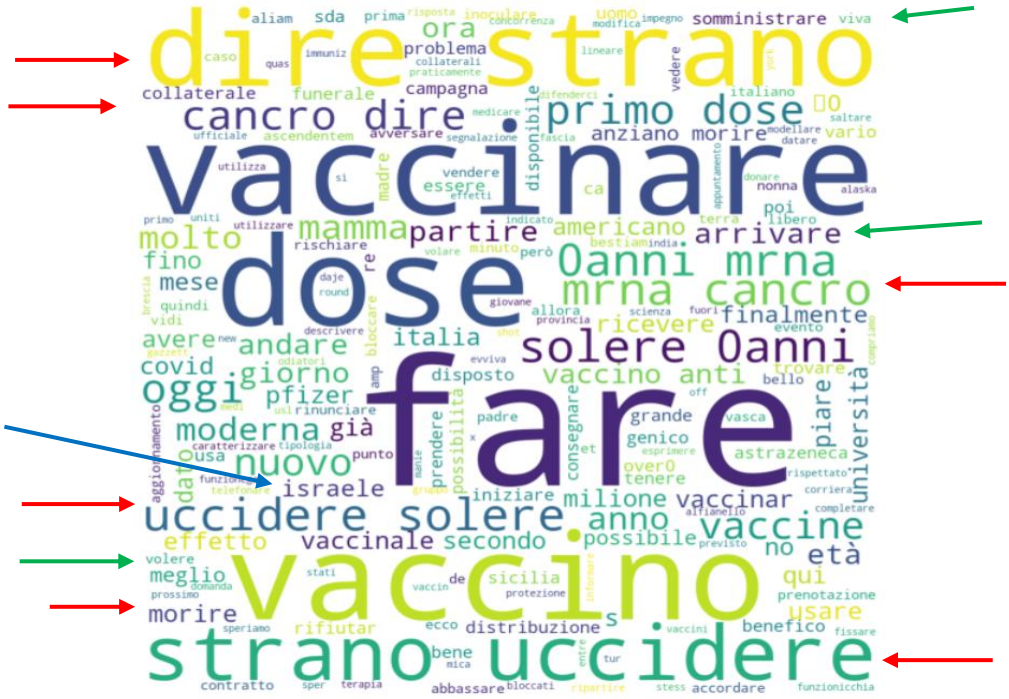
5.1.2 Astrazeneca



Per quanto riguarda il wordcloud di Astrazeneca oltre alle parole ovviamente presenti come "Astrazeneca" e "vaccino" si può notare che la grandezza, e di conseguenza l'importanza delle parole, è sommariamente tutta uguale. Quindi per quanto riguarda Astrazeneca non ci sono dei possibili argomenti più importanti di altri che possiamo subito individuare.

In ogni caso ad un analisi più approfondita possiamo trovare delle parole che richiamo a degli argomenti negativi come: i casi di trombosi, gli effetti collaterali e l'azione legale contro l'azienda svedese. Alcune parole riguardano l'EMA e le sue approvazioni.

5.1.3 Moderna



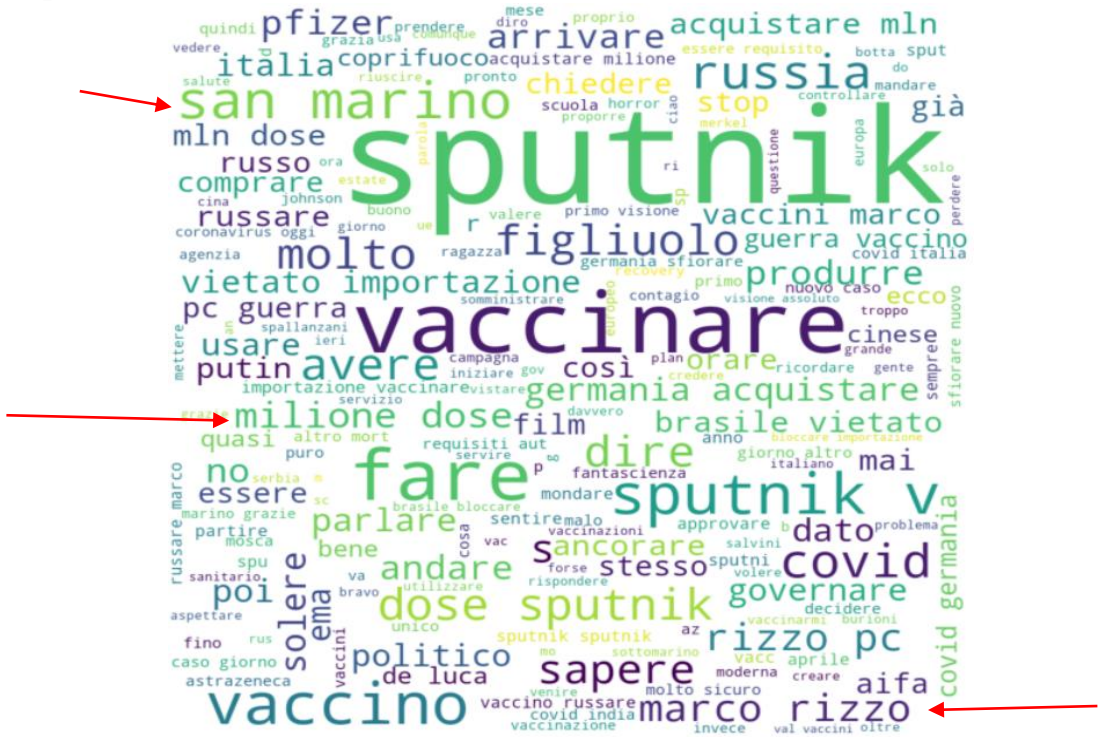
Da questa immagine si possono già notare quali sono le parole più importanti a cui possiamo intuitivamente ricondurre a un connotato positivo o negativo.

In particolare, le parole indicate dalle frecce rosse hanno un significato negativo poiché parlano di morte o di malattie, probabilmente riconducibili a dei tweet che vogliono sottolineare la possibilità che con la vaccinazione si possa incappare in complicazioni anche molto gravi.
Le frecce verdi, invece indicano delle parole che sono probabilmente legate a dei tweet di sentimento positivo.

Credo che sia importante notare (freccia blu), come nel caso di Pfizer, come sono state catturate delle parole anche che riguardano anche l'andamento della situazione pandemica e vaccinale in altri stati. Questo era aspettato anche perché la maggior parte delle volte si parla di Pfizer e Moderna insieme (come vedremo in seguito sulla parte di analisi delle frequenze delle parole).

Dalla figura si può già immaginare che saranno presenti più tweet con un sentiment negativo essendo riusciti a individuare più parole che ci portano a pensare che esse derivino da tweet negativi.

5.1.4 Sputnik



Da questo word cloud mi aspettavo di vedere più parole che suscitassero un sentimento negativo e diffidenza, ma sorprendentemente non è accaduto. In queta immagine si vedono le parole “russia” e “putin” le quali sono alquanto scontate dato che il vaccino è russo e la figura di Putin e legata a tale nazione. Ci sono delle parole come “comprare” che possono far pensare che molti dei tweet riguardino il fatto di comprare anche questo vaccino per riuscire ad avere più dosi possibili.

Una tra le parole che spicca sulle altre è “San Marino”. Infatti questa nazione è stata una tra le prime nella zona Europea a vaccinare con il vaccino Sputnik.

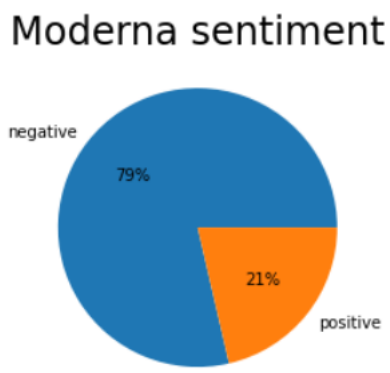
Un'altra coppia di parole che salta all’occhio è “Marco Rizzo”(segretario del partito comunista) il quale è un politico italiano che afferma che il vaccino Sputnik non è possibile utilizzarlo in Europa perché: “c’è un blocco atlantico che non vuole”. Egli si è espresso in maniera analoga anche per altri vaccini come quello cinese. Queste sue affermazioni devono aver generato una consistente risposta sui social network ed è per questo che troviamo il suo nome nell’immagine.

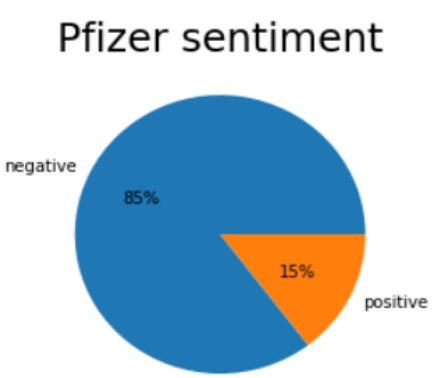
Si può vedere anche che nonostante tutte le pulizie nel testo e la selezione che è stata fatta ci siano delle parole che non centrano tanto come “russare marco”.

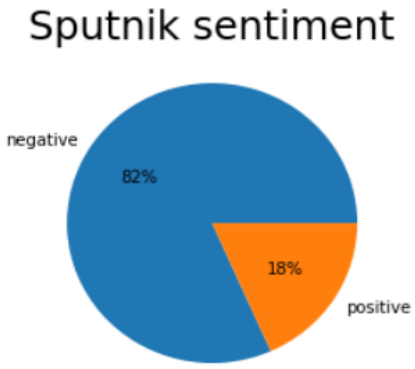
5.2 Grafici sul sentiment

Adiamo adesso ad analizzare il sentiment che è stato predetto dalla rete neurale basata su BERT.

Ci aspettiamo che ci siano più negativi che positivi, questo perché le parole viste nel word cloud ci fanno sospettare così. In più è da tenere in considerazione la tendenza a lamentarsi delle persone sui social network.







Da questi grafici si nota subito quanto i tweet categorizzati come negativi siano molti di più rispetto a quelli positivi, ben oltre le nostre previsioni iniziali.
Ci si aspettava che Astrazeneca avesse una predominanza di tweet con sentimento negativo. Questo sicuramente a causa di tutti gli scetticismi riguardanti i vaccini e, rafforzati dal fatto, che i primi casi di trombosi e di morte si siano verificati proprio con il vaccino Astrazeneca. Il fatto che l'EMA e l'AIFA si siano espresse in merito più volete su questo vaccino hanno alimentato il sentimento di diffidenza e paura. Anche il cambio di nome dell'azienda e del bugiardino del vaccino devono aver influenzato negativamente l'opinione pubblica.

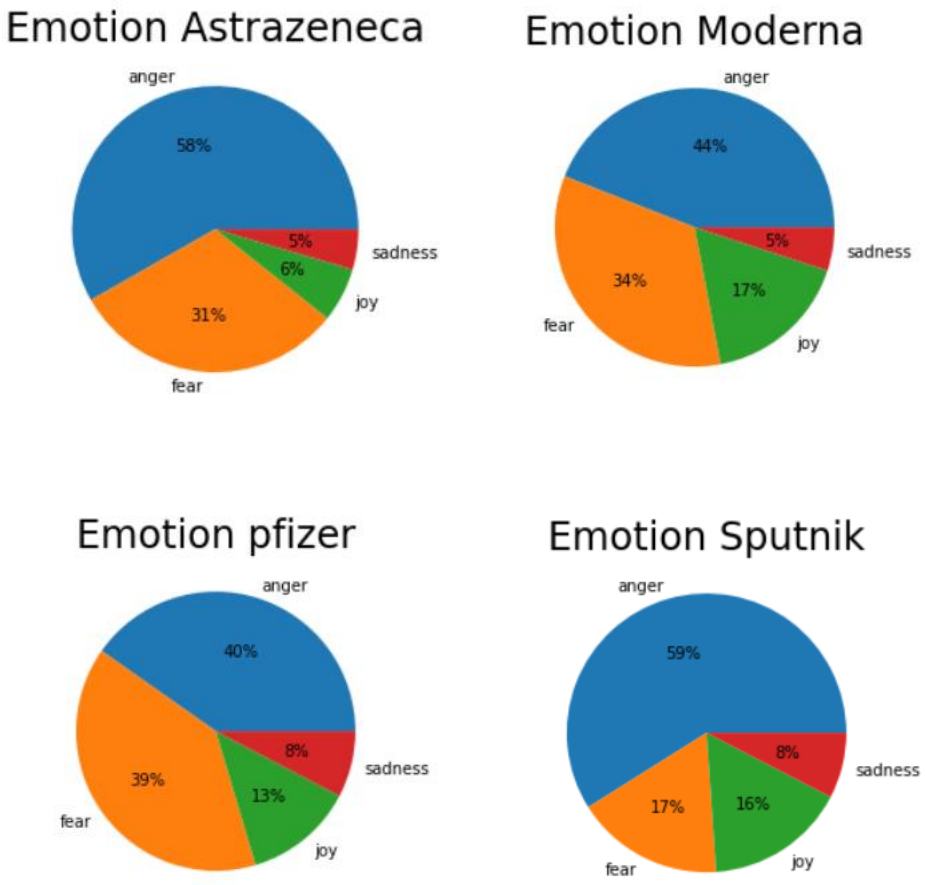
Per quanto riguarda Pfizer ci aspettavamo che i tweet positivi fossero in numero maggiore rispetto ad esempio a quelli di Astrazeneca, ma non ci aspettavamo così tanti negativi. In particolare dopo il successo avuto nella campagna vaccinale di Isdraele credevo che il sentimento positivo fosse maggiore, non di molto, rispetto a quello attualmente trovato. Forse questo potrebbe essere dovuto al condizionamento dell'opinione pubblica dopo i casi di trombosi di Astrazeneca. Questo scetticismo potrebbe essere anche dovuto al fatto che la maggior parte delle persone non si fidino a priori dei vaccini in analisi poiché creati e distribuiti in pochissimo tempo. Questo perché normalmente per un farmaco ci vogliano anni per l'approvazione prima di essere messo in commercio.

Per quanto riguarda Moderna, il sentiment di tale vaccino e in linea con le aspettative e, forse giova anche del fatto che non se ne parli tanto rispetto agli altri due approvati dall'EMA.

Per quanto riguarda lo Sputnik mi aspettavo di trovare più negativi non essendo un vaccino non approvato e proveniente dalla Russia, uno stato che nell'opinione comune non è trasparente e non gode di molta fiducia in Europa. Forse il fatto di avere un sentiment in linea con gli altri vaccini potrebbe essere dovuto al fatto che nei giorni di scaricamento dei tweet si parlava e si spingeva per il fatto di farlo approvare. Negli stessi giorni allo Spallanzani si stavano facendo i primi test con tale vaccino. Altro evento che potrebbe aver fatto cambiare un po' l'opinione delle persone e che alcuni

governatori di regione popolari tra i cittadini prendessero in considerazione l'acquisto di dosi del vaccino russo, alimentando così la fiducia della popolazione tu tale vaccino.

5.3 Emotion Analysis

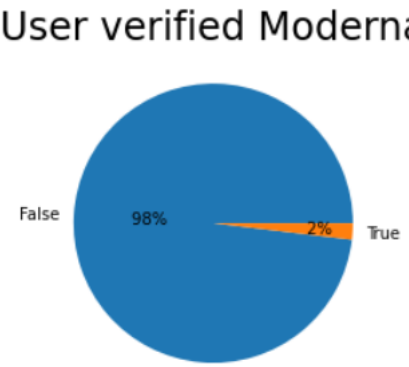


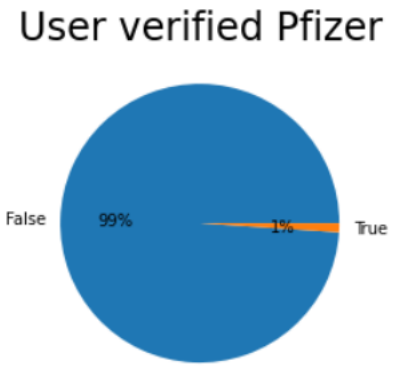
Ho utilizzato l'emotion analysis anche come prova per verificare realmente che i tweet negativi fossero caratterizzati da un emozione "negativa". Infatti, questa affermazione si può provare subito dai grafici dove l'unica emozione positiva "joy" è in netta minoranza rispetto alle altre emozioni che hanno una connotazione più negativa. In particolare è interessante sempre prendere in considerazione il caso di Astrazeneca in cui l'emozione predominante e la "anger" e "fear". Questo probabilmente dovuti ai casi di trombosi e al fatto che l'azienda non ha guadagnato la fiducia dei cittadini (es. cambio del bugiardino, varie ripensamenti degli enti di approvazione dei farmaci e al fatto che l'UE ora voglia procedere legalmente conto l'azienda per non aver soddisfatto gli accordi). Per quanto riguarda Pfizer e Moderna il sentimento di "fear" è più alto rispetto ad Astrazeneca questo forse perché gli effetti collaterali (che probabilmente generano l'emozione "anger") in seguito alla somministrazione di questi vaccini non hanno avuto la stessa risonanza mediatica di quelli di Astrazeneca, ma rimane il sentimento di sospetto riservato a tutti i vaccini.

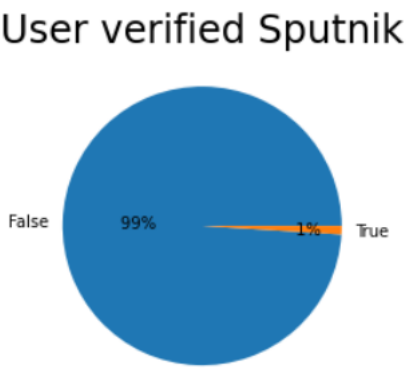
Lo Sputnik presenta il sentimento positivo "joy" in quantità maggiore rispetto agli altri vaccini, può essere forse perché, come detto anche nella sezione precedete, il vaccino stava venendo preso in considerazione per una possibile approvazione.

5.4 User verified analysis

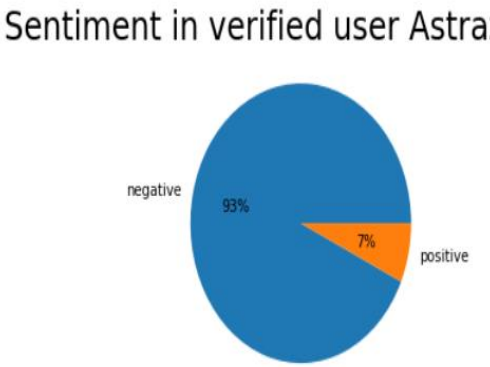
Per provare a capire meglio da dove venissero questi sentimenti negativi e se fossero indotti da qualcosa; si sono provate a fare varie ipotesi. Una di queste è che gli utenti verificati, quindi quelli che hanno più risonanza e più importanza rispetto ad un utente medio del social network potessero aver condizionato gli altri utenti con le loro idee.

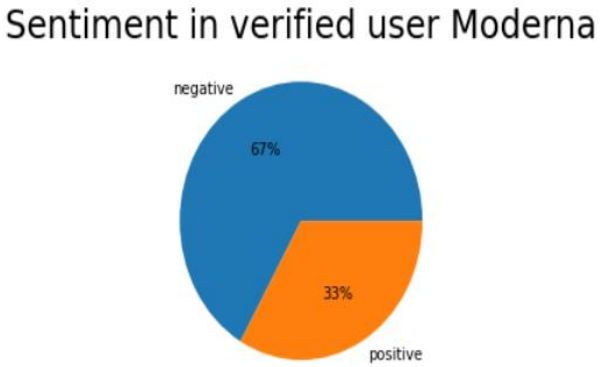


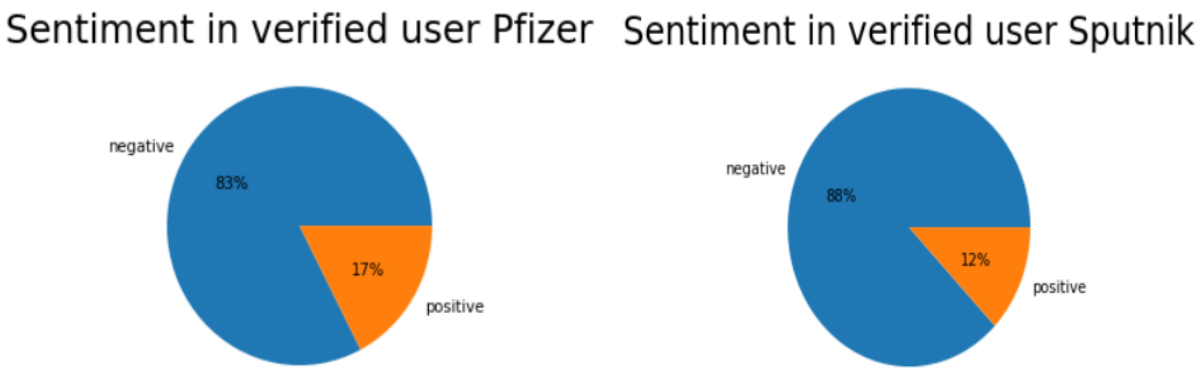




Vedendo questi grafici si nota subito come la percentuale di utenti verificati nel nostro dataset sia talmente esigua da poter rigettare subito l'ipotesi fatta precedentemente. Soprattutto perché come si può vedere dall'immagine (sotto) gli utenti verificati rispecchiano sommariamente l'andamento del sentiment globale.



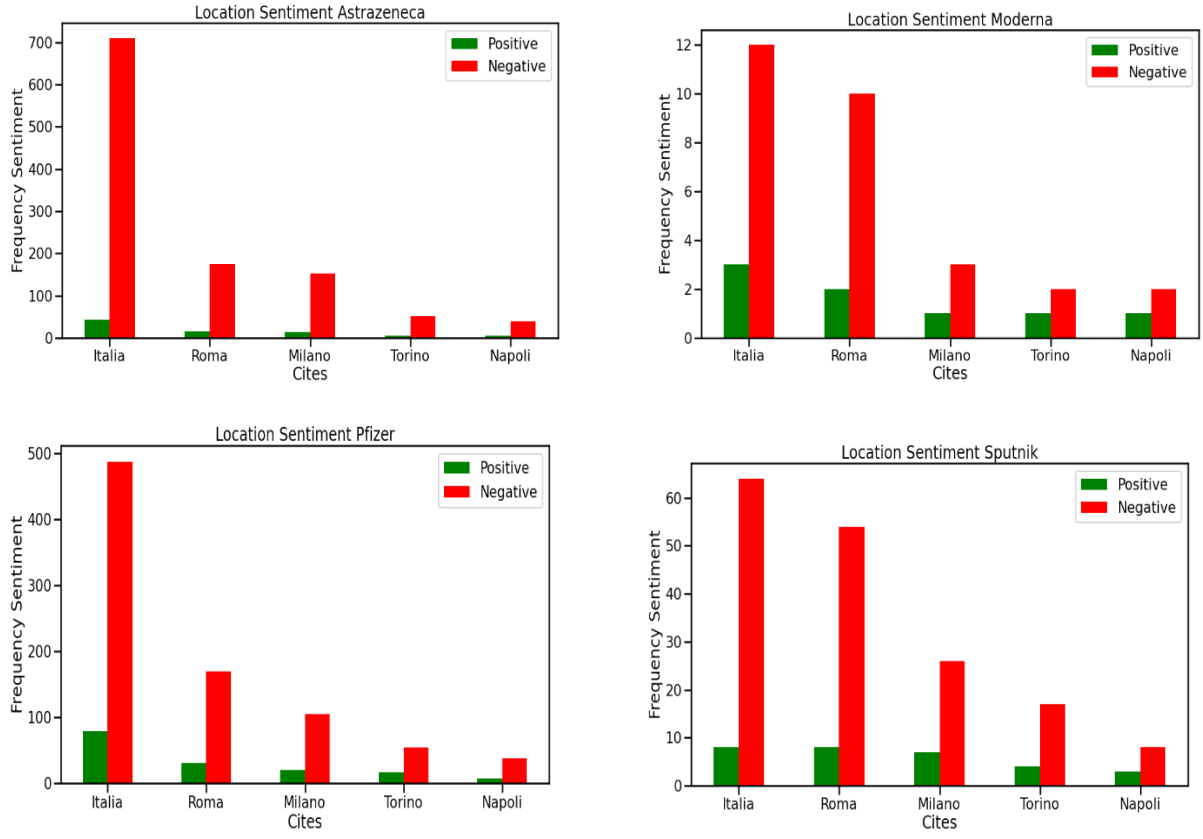




Solo per il vaccino Moderna si vede come la percentuale di tweet degli utenti verificati siano un terzo del complessivo. Potrebbe essere forse per questo che Moderna ha il sentiment positivo più presente rispetto agli altri vaccini. (Si veda appendice A per sapere cosa pensano gli utenti più influenti.)

5.5 Location Analysis

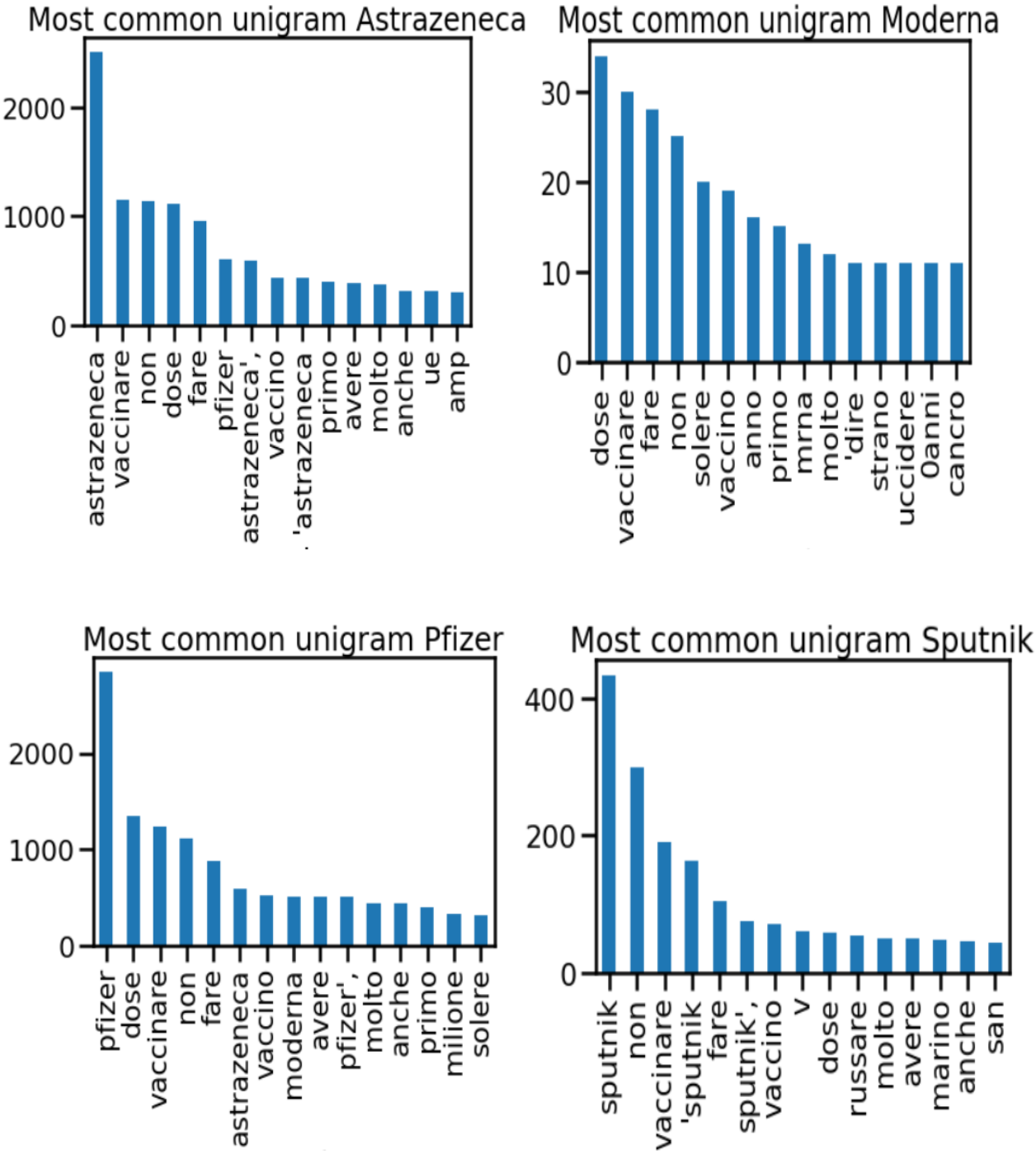
In questa sotto sezione l'intenzione è di andare a vedere se la distribuzione dei tweet negativi e positivi cambia in base alla locazione dell'utente. Infatti, potrebbe capitare che delle città che sono state molto colpite dalla situazione pandemica siano più propense a farsi vaccinare e, di conseguenza, ci siano dei sentimenti più positivi riguardanti i vaccini.



Da questi istogrammi possiamo vedere come anche nelle città principali tra cui c'è Milano, che è una tra le città più colpite, la distribuzione del sentiment rimanga sostanzialmente la stessa vista in precedenza. Non è possibile di conseguenza affermare che in base al luogo dell'utente esso la pensi più positivamente o più negativamente rispetto ai vari vaccini.

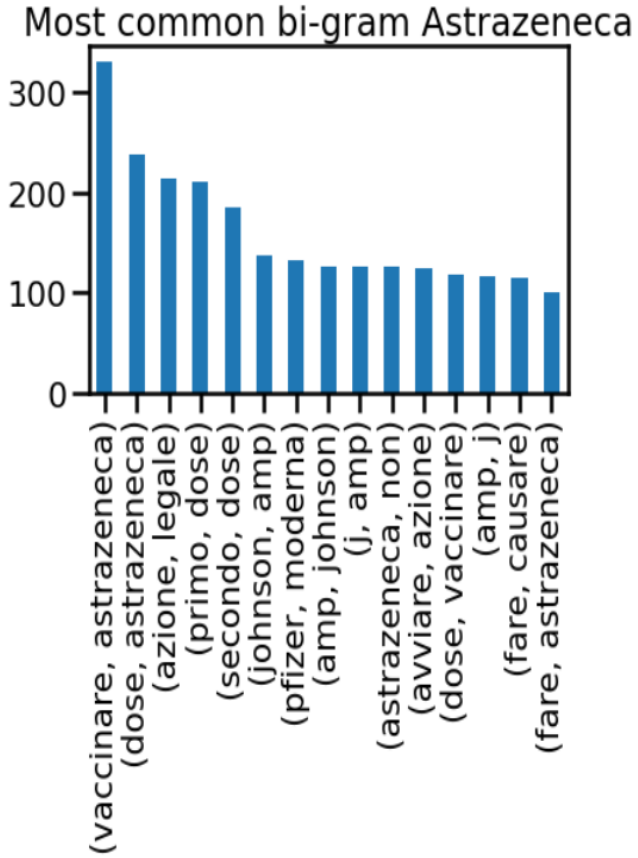
5.6 Most frequent word

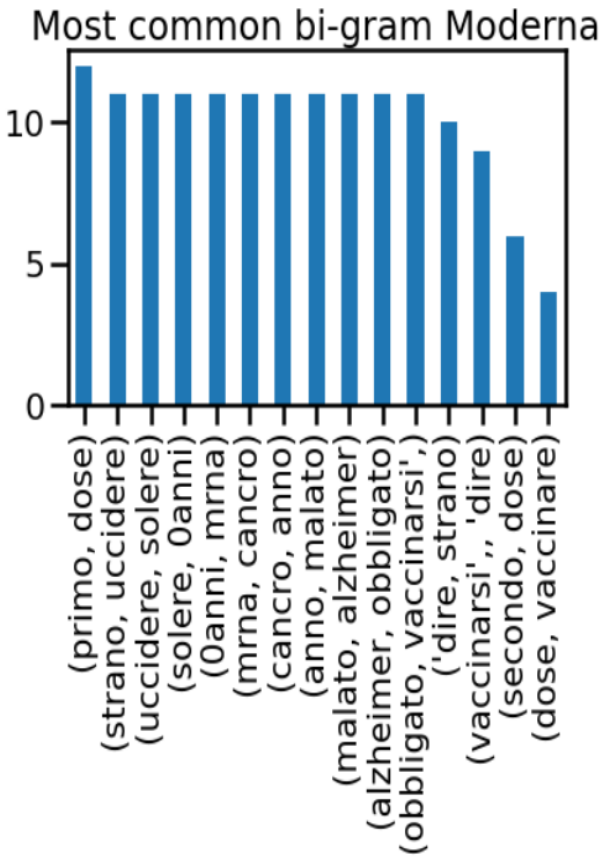
In questa sezione volgiamo analizzare le parole cercando di capire quali sono le più frequenti e che significato danno alle frasi. Un po' come si è già fatto nella sezione riguardante il word cloud, ma adesso prendendo in considerazione meno parole, ma conoscendo esattamente la loro frequenza.

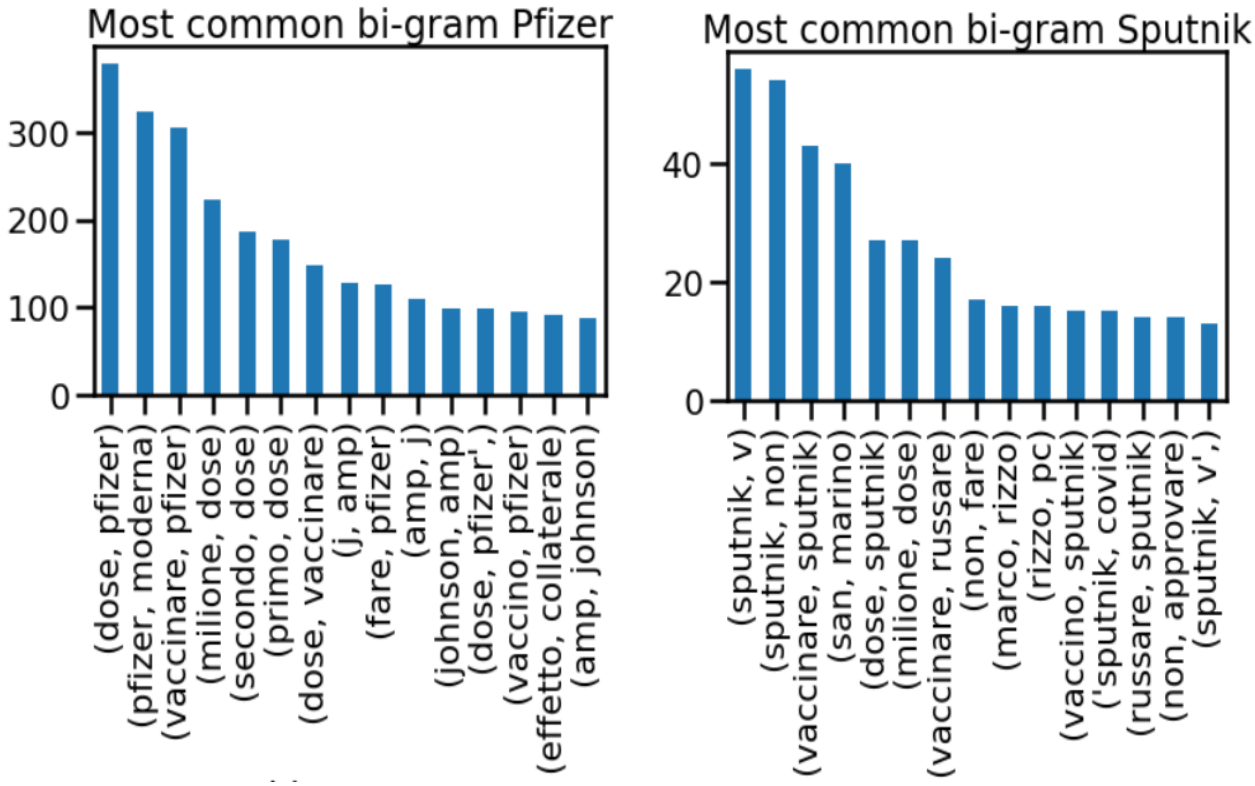


Vediamo, come già notato nel word cloud che le parole più frequenti sono i nomi stessi dei vaccini e ovviamente la parola "vaccino". Molto interessante notare, cosa che non si è riusciti a cogliere nel word cloud, come la parola "non" compaia tra le prime posizioni in ogni grafico e con essa anche la parola "fare". Potrebbe essere che entrambe le parole compaiano insieme dando così un connotato negativo alla maggior parte dei tweet (concorde a quanto analizzato prima).

Si vedere come analizzando solo le parole singole non si riesca a catturare tutto il loro significato. Infatti, nell'ultimo grafico vediamo che le parole "san" e "marino" sono analizzate separatamente quando sappiamo benissimo che dovrebbe essere insieme. Di conseguenza si è deciso di analizzare anche la frequenza dei bi-grams per cercare di estrarre più informazioni possibili dalla frequenza delle parole nel corpus.







Vediamo come la parola “non” si verifichi spesso vicino al nome di un vaccino, o vicino alla parola “fare” che conferma la tesi sopra descritta. L’unico grafico in cui la parola “non” non si presenta è il grafico di Pfizer. Questo accade perché nei bi-grams è presente del rumore dovuto ad esempio alle combinazioni con il vaccino J&J che prendono più posizioni in questa “classifica”. In particolare, se andassimo a prendere più posizioni potremmo osservare la combinazione “pfizer , non” (80 comparizioni) e “non avere” (80 comparizioni) rispettivamente nelle posizioni 17 e 18 (nel grafico sono riportate solo le prime 15).

6 Predizione

In questa parte abbiamo provato ad implementare dei modelli di machine learning bastati sulla Logistic Regression e a valutare le loro prestazioni. In Si è notato che tali modelli non presentavano delle prestazioni nettamente migliori rispetto alla loro baseline basata sul “most_frequence”. I modelli di apprendimento automatico presentano delle performance paragonabili, se non uguali, a quelli della baseline. Questo fatto potrebbe essere dovuto al grande sbilanciamento che presenta il dataset. Infatti, come ricordiamo più del 80% delle osservazioni presentano un sentiment negativo. Appurato ciò si è provato a migliorare i vari modelli facendo features selection e regualrization. Dopo aver applicato queste due tecniche il modello comunque non presentava dei miglioramenti evidenti.

Di conseguenza ho ipotizzato che il problema derivasse dal fatto che non avevamo abbastanza dati per allenare il modello al meglio. Per cercare di risolvere questo problema ho deciso di unire tutti i dataset creandone uno con più di 10.000 osservazioni e di fittare le Linear regression su questo. Anche con questa metodologia però l’accurancy (parametro che ho tenuto in considerazione per valutare quando il modello fosse buono) non si è alzata di molto rimanendo sempre paragonabile alla baseline.

Adesso andremo a commentare i risultanti derivanti dall’applicazione della Logistic Regression al dataset totale. Per i dataset singoli (quindi contenenti informazioni riguardanti un solo vaccino) si possono fare delle constatazioni simili.

```
BASRELINE
0.8803118168629357
```

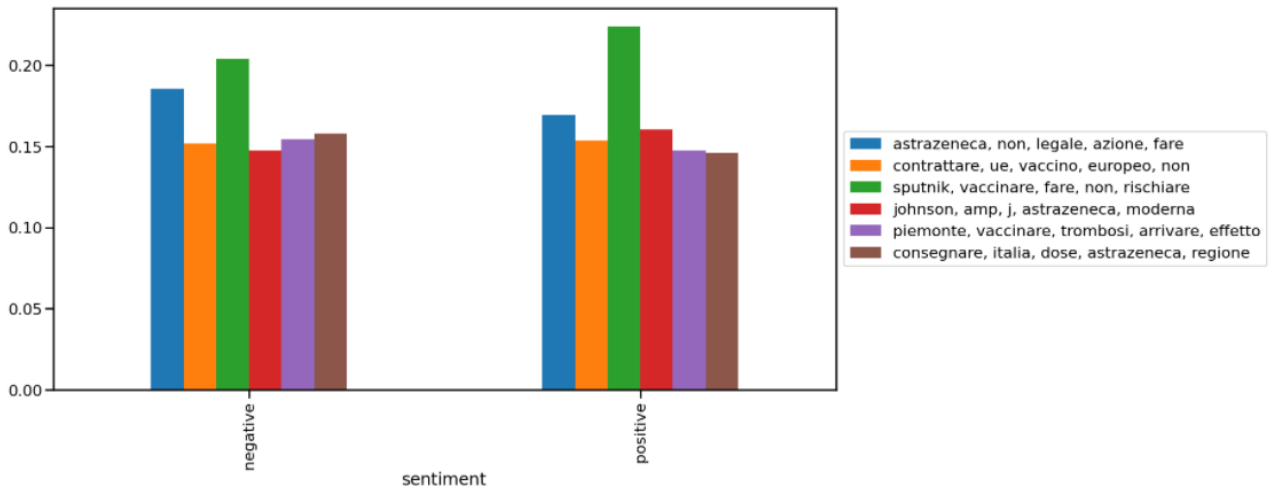
	precision	recall	f1-score	support
0	0.89	1.00	0.94	1770
1	0.00	0.00	0.00	219
accuracy			0.89	1989
macro avg	0.44	0.50	0.47	1989
weighted avg	0.79	0.89	0.84	1989

	feature	coefficient
800	mrna grazia	-5.911083
803	muore seconda	-5.594234
1397	vaccinare molto	-5.272544
778	moglie	-5.246938
1104	regioni milione	-5.137886
...	...	...
1141	rimanere	4.632228
406	direttore	4.712447
530	fatta primo	4.795051
1318	tempo	5.536607
875	pario consegnare	5.550169

Nelle immagini sopra abbiamo riportato i risultati del modello dopo aver applicato la selezione e la regolarizzazione dei parametri del modello. Vediamo come la baseline e l'accurancy siano particolarmente vicine come valori e quindi di prestazioni comparabili. Abbiamo anche voluto riportare i coefficienti che hanno le features nel nostro modello. Vediamo che la feature "vaccinare molto" e "mrna grazia" possiedono dei coefficienti molto negativi, mentre "tempo" e "pario consegnare" dei coefficienti molto positivi. Tramite questa semplice tabella non possiamo dedurre molto più di quello già fatto in fase di analisi. Di conseguenza si è ritenuto necessario approfondire con delle tecniche specifiche i topics dei vari documenti e capire se hanno senso i nostri risultati.

7 Topic modelling

Per analizzare gli argomenti dei vari tweet abbiamo utilizzato la tecnica dell'LDA. L'abbiamo applicata prima solo sui testi negativi e poi solo sui positivi. In modo da capire per ogni sentimento quali risultavano essere gli argomenti latenti nei nostri dati. Infine abbiamo applicato la tecnica a tutto il dataset per avere una visione complessiva degli argomenti.
Per scegliere il numero di topic migliore abbiamo usato le tecniche di coerence scores (UMass and CV) con le quali abbiamo scoperto che il numero di topic migliore sta tra 5 e 8 (noi abbiamo scelto 6). Dopo di che abbiamo estratto i nostri descriptros. Per vedere se questi topic risultavano più correlati ai positivi o ai negativi e abbiamo fatto l'aggregation in base al sentiment come si può vedere dal grafico sotto.



Questa figura ci suggerisce qualche osservazione interessante.

In particolare vediamo come il topic di colore verde sia maggiormente presente nelle osservazioni positive. Le parole che compongono questo topic ci suggeriscono che si parla del vaccino Russo e di come forse non sia una buona idea vaccinare con questo vaccino. Sembra corretto che questo topic sia leggermente più presente tra i positivi poiché probabilmente i tweet riguarderanno il fatto che le persone pensino che il vaccino russo sia rischioso e che sia più opportuno vaccinarsi con un altro vaccino.

Il topic identificato dal colore blu è maggiormente presente tra i negativi. Leggendo le parole che descrivono il topic si può dire che si parla quasi inequivocabilmente di Astrazeneca e dell'azione legale che l'UE ha intenzione di fare contro il colosso svedese.

Anche il topic di colore viola è leggermente più presente tra i negativi. Le parole che lo compongono hanno un connotato che mediamente è negativo, in particolare la parola "trombosi". Probabilmente questo topic si riferisce agli effetti collaterali dei vaccini.
Per la parola "piemonte" non si riesce a trovare un grande nesso con le altre parole che compongono il topic.

Infine, l'ultimo topic che sembra interessante commentare è quello di colore marrone. Infatti, questo topic è più presente tra i negativi e sembra che parli del fatto che non arrivino abbastanza vaccini all'Italia da parte delle case farmaceutiche. Questo molto probabilmente genera dei tweet arrabbiati e indignati e, di conseguenza, categorizzati come negativi.

8 Conclusioni

Dalle analisi fatte si può concludere che la maggior parte dei tweet esprimono un sentimento negativo. Questo è dovuto a molteplici fattori, sia psicologici che di fatti di cronaca. Infatti, tra i fattori psicologici/comportamentali possiamo trovare il fatto che le persone tendono di più a lamentarsi sui social generando così più tweet con sentiment negativo. Oltre a questo negli ultimi anni si sta pensando sempre peggio dei vaccini e quindi si risulta più diffidenti su questo argomento, soprattutto per i vaccini riguardanti il Covid-19. Infatti, le persone hanno molta diffidenza soprattutto per il fatto che tali vaccini siano stati creati e commercializzati in pochissimo tempo rispetto agli altri farmaci (anche se la cronaca ci insegna che in realtà questo è dovuto a delle velocizzazioni burocratiche, e degli investimenti sia in termini di denaro che di menti fuori dal comune). Le persone esprimono spesso negatività su questi argomenti anche perché: gli effetti collaterali, azioni legali e altri eventi di cronaca generano molta risonanza mediatica e un enorme flusso, nel nostro caso, di tweet negativi. Infatti, gli eventi di cronaca negativi hanno molto più scalpore nell'opinione pubblica, rispetto a quelli positivi.

In seguito a tutte queste deduzioni possiamo affermare che sia normale avere una grandissima percentuale di tweet negativi.

Per quanto riguarda la parte della predizione e inutile dire che non siamo riusciti ad avere delle buone predizioni e che, almeno con il modello di regressione logistica, non riusciamo a superare nettamente la baseline. Questo ci porta a dover pensare di implementare degli altri modelli più complessi.

8.1 Possibili sviluppi futuri

Gli sviluppi futuri potrebbero essere quelli di provare a allenare altri tipi di algoritmi predittivi come ad esempio le SVM e provare a utilizzare anche i dati uscenti dall'emotion analysis. In questa maniere si potrebbero allenare degli algoritmi più complessi, ma che probabilmente presentino un'accuratezza più alta.

Un'altra possibile implementazione potrebbe essere quella di implementare un algoritmo di clustering e vedere come le osservazioni vengono raggruppate tra loro. In questo modo forse si potrebbero scoprire altre informazioni latenti nei dati.

9 References

Bocconi BERT model for sentiment:
https://huggingface.co/MilaNLProc/feel-it-italian-sentiment

Bocconi BERT model for emotion:
https://huggingface.co/MilaNLProc/feel-it-italian-emotion

Twitter developer account and data dictionary:
https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet

BERT original paper:
https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

LDA :
https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

Appendice A

In seguito all'analisi sugli utenti verificati abbiamo provato a vedere anche l'utente/utenti che hanno più follower, il che non significa per forza che siano verificati. È giusto analizzare anche loro dato che i loro tweet potrebbero arrivare a molte persone.

	Pfizer	**Astrazeneca**	**Sputnik**	**Moderna**
user_name	La Stampa	Tgcom24	iodonna	ilGiornale
user_screen_name	LaStampa	MediasetTgcom24	IOdonna	ilgiornale
tweet	'Accordo Ue-Pfizer per 1,8 miliardi di dosi, Von der Layen: "Entro luglio vaccinato il 70% dei cittadini europei" '	"Vaccino AstraZeneca , Ema ribadisce: no limite uso per fasce d'età #astrazeneca "	Amanti dell fantascienza, onestamente, come si può perdere un film con un titolo cosi Guerra fredda, così vintage? https://t.co/JROypbrYe2	"Non sono previste modifiche al piano, impegni rispettati". #Fontana spiega la decisione di destinare #Pfizer e #Moderna alle prime dosi e rassicura chi deve fare il richiamo con #AstraZeneca https://t.co/Aqip8kqM0X
number_of_follower	1195115	1155901	117191	537196
is_verified	True	True	True	False
sentiment	negative	Positive	negative	Negative
emotion	fear	Fear	sadness	anger

Da questa tabella si nota subito come la maggior parte degli "user" con più follower siano delle testate giornalistiche che degli enti già verificati, quindi possiamo attenerci alla categoria di prima. Purtroppo dalla tabella si vedere anche come il tweet riguardante sputnik non centri in realtà nulla con i vaccini, ma parli di un film sulla guerra fredda. Purtroppo è inevitabile non riuscire a scremare queste informazioni a priori e quindi i nostri dati saranno sempre in presenza di rumore o "outlayers". Probabilmente questa è solo un errore casuale dato che le parole analizzate nella sezione dedicata al wordcloud risultano congrue con l'ambito dei vaccini.

Tabella A

Label	type	signficato
tweet	string	testo del tweet o del commento ad un tweet
is_quoted	boolean	This field only surfaces when the Tweet is a quote Tweet. This field contains the integer value Tweet ID of the quoted Tweet.
tweet_quoted	string	This field only surfaces when the Tweet is a quote Tweet. This attribute contains the Tweet object of the original Tweet that was quoted.
name_of_who_I_answered	string	*Nullable.* If the represented Tweet is a reply, this field will contain the screen name of the original Tweet's author.
id user_str	string	The string representation of the unique identifier for this Tweet. Implementations should use this rather than the large integer in id
lang_user	string	*Nullable.* When present, indicates a BCP 47 language identifier corresponding to the machine-detected language of the Tweet text, or und if no language could be detected.
creation_date_of_tweet	string	UTC time when this Tweet was created.
result_type	string	Metadata.resulte_type
number_of_retweet	Int	Number of times this Tweet has been retweeted.
retweeted	boolean	Indicates whether this Tweet has been Retweeted by the authenticating user.
source	string	Utility used to post the Tweet, as an HTML-formatted string. Tweets from the Twitter website have a source value of web.
user_name	string	The user who posted this Tweet. See User data dictionary for complete list of attributes. We selected the name of user
user_screen_name	string	The user who posted this Tweet. See User data dictionary for complete list of attributes. We selected the twitter name of the user
location	string	The user who posted this Tweet. See User data dictionary for complete list of attributes. We selected the location of the user if he/she put it in his/her twitter profile
user_description	string	The user who posted this Tweet. See User data dictionary for complete list of attributes. In particular we extraxt the user bio
number_of_follower	int	Number of follower of the user
number_of_friend	Int	Number of follow
is_verified	boolean	If the user is verified from twitter
creation_date_of_account		UTC time when the user profile was created.
like	Int	Number of like to the tweet