Università di Pisa

Laurea Magistrale in Informatica Umanistica

Seminario di cultura digitale
a.a 2021-22

Educazione e COVID-19
Analisi delle opinioni sulla DAD degli utenti di Twitter

Eleonora Rossi
Matricola *562337*

Sommario

Quali sono le opinioni maggiormente diffuse sulla DAD all'interno della piattaforma *social Twitter*? L'obiettivo di questa relazione consiste nell'individuare in maniera automatica emozioni, sentimenti e temi che trapelano dai tweet pubblicati dagli utenti italiani in merito alla DAD, didattica a distanza, attraverso l'utilizzo di strumenti di *Natural Language Processig*.

Indice

1 Introduzione

Con lo scoppio della pandemia COVID-19, si è reso necessario reinventare il modo di educare ed istruire la popolazione, in quanto è dimostrato come anche brevi interruzioni scolastiche possano causare una significativa perdita formativa[1].

Infatti, con il sempre più utilizzato metodo preventivo del *lockdown* soprattutto in Italia, primo Paese in Europa ad attuare tale pratica a livello nazionale, è stato fondamentale il repentino passaggio da didattica in presenza a didattica a distanza, reso possibile grazie all'ausilio di strumenti elettronici come PC e cellulari nonché di applicazioni online volte alla comunicazione audio e video come *Teams* o *Google meet*. A partire dal 10 marzo 2020, tale fenomeno ha riguardato tutti i campi della didattica, dalle scuole elementari, alle scuole medie e superiori, fino all'università dove in alcuni casi tale metodo è stato adottato già a partire dal giorno dopo la promulgazione dello stato di *lockdown*. Attraverso lo studio dei risultati ottenuti a partire da un'indagine del Ministero dell'istruzione, alcune testate giornalistiche hanno riferito come dal 18 marzo 2020 il 67% delle scuole aveva spostato tutte le proprie attività didattiche online, raggiungendo quindi 6,7 su 8,3 milioni di studenti in Italia[2].

La didattica a distanza, o come viene definita sinteticamente **DAD**, è stata quindi una soluzione immediata al problema che ha permesso milioni di italiani di proseguire con i propri studi in un momento di così grande bisogno. Tuttavia l'opinione pubblica sull'argomento sembra essere cambiata notevolmente nel contesto di questa pandemia: con il diffondersi sempre più pervasivo e capillare del COVID-19, tanto che l'intera popolazione ha dovuto vivere in quasi totale isolamento per circa due mesi fino al 3 maggio 2020, tale soluzione ha iniziato a dividere l'opinione pubblica. Ad esempio, come spiegato in (Mascheroni et al., 2021), un indagine sull'esperienza formativa da remoto vissuta da bambini e ragazzi tra i 10 e i 18 anni coordinata da *Joint Research Center* della Commissione Europea sottolinea come tale metodo sia di per sé pervaso da evidenti disuguaglianze a causa della mancanza di un accesso a internet o di dispositivi digitali adeguate e ciò potrebbero "compromettere le opportunità di apprendimento a distanza". Tale indagine mostra anche come un numero considerevole di genitori ha riscontrato invece un miglioramento delle capacità dei propri figli proprio grazie all'utilizzo della didattica a distanza.

É all'interno di tale contesto e prendendo spunto dal seminario del professor Paolo Monella sul tema della DAD che si è svolto tale progetto. In particolare attraverso l'ausilio di tecniche di NLP e *machine learning* per l'analisi dei testi si andrà a studiare un corpus costituito da tweet prodotti da utenti italiani dal primo marzo 2020 al primo marzo 2022 in merito alla DAD al fine di carpire le opinioni maggiormente diffuse. In un primo momento si farà uso di strumenti volti all'analisi linguistica computazionale per poter studiare alcuni aspetti di base dei tweet estratti, così da poter dare un primo sguardo sulle preoccupazioni che maggiormente interessano e accendono l'opinione pubblica. Si procederà poi con un'indagine più profonda attraverso l'ausilio di tecniche di *deep learning* per estrarre in maniera automatica opinioni ed emozioni che trapelano dai tweet prodotti.

Nella prima sezione di tale relazione si tratterà in maniera dettagliata della creazione del dataset, contenente i dati ottenuti mediante la tecnica di *scraping* dal *social network* Twitter. A partire da tale dataset verrà così realizzato il corpus oggetto di analisi a cui verranno applicate dei preliminari metodi di indagine matematici e statistico-descrittivi così da comprendere al meglio la natura dei dati linguistici raccolti e le tematiche affrontante.

[1] (Alban Conto et al., 2020)
[2] (Mascheroni et al., 2021)

Successivamente si analizzerà in maniera automatica le emozioni e i sentimenti polarizzanti (positivi o negativi) che scaturiscono da tali tweet attraverso tecniche di Sentiment-Analisys con lo scopo finale di individuare le opinioni maggiormente diffuse in merito alla DAD.

Come penultima analisi si effettuerà un'ulteriore indagine sugli argomenti maggiormente discussi dal popolo di Twitter al fine di carpire quali sono le problematiche relative a tale metodo educativo.

L'ultima sezione sarà dedicata a riassumere quanto studiato finora, andando a riportare i risultati più importanti che sono emersi nel corso dell'indagine, nonché menzionare possibili futuri sviluppi della ricerca.

2 Creazione corpus di analisi

In prima istanza si è resa necessaria la fase di ***Data Collection*** durante la quale sono state raccolte le risorse occorrenti per procedere poi alla creazione del corpus linguistico a partire dal quale effettuare le nostre indagini. Si è deciso di utilizzare come fonte di dati il *social network* **Twitter**, in quanto risulta essere molto più dinamico e immediato nell'esprime opinioni su argomenti di tendenza. Ciò favorisce l'interazione tra utenti, la discussione e la diffusione di opinioni anche tra utenti molto polarizzati. Il dinamismo e la velocità con cui le informazioni vengono diffuse sono favorite anche dalla quasi totalità dei profili pubblici, aspetto che incarna perfettamente la logica dietro questo *social network* (lo stesso non si può dire, invece, per *social* più comuni come Instagram e Facebook). Questa decisione è stata ulteriormente avvalorata da una maggiore facilità nell'effettuare la tecnica di *scraping* per cui è stato adottato come strumento la libreria ***Twint***[3] e il linguaggio di programmazione Python.

2.1 *Scraping* dei dati

L'arco temporale durante il quale sono stati raccolti i dati ha come data di inizio il primo marzo 2020 e si conclude il primo marzo 2022. In questo modo l'indagine sarà effettuata in un arco temporale di 2 anni esatti a partire da 3 giorni prima dell'entrata in vigore in Italia del primo lockdown. Sono stati raccolti tweet pubblicati sia da account verificati che da account non verificati, contenenti i due seguenti hashtags: ***#dad***, ***#didatticaadistanza***. Al fine di raggiungere un sufficiente numero di dati non è stato applicato alcun vincolo sulla ricerca sul numero di tweets ammessi.

I dati sono stati raccolti in due file *csv* ciascuno relativo ad uno degli hashtag appena menzionato. Alla fine del processo sono stati scaricati in totale 17367 tweets in lingua italiana. Abbiamo così ottenuto un dataset composto da **37 *features*** e **17367 records**. Segue una lista del numero di tweets ottenuti per ciascun hashtag:

1. **#dad:** 15049
2. **#didatticaadistanza:** 2420

2.2 *Data preparation* e *data understanding* dei dati

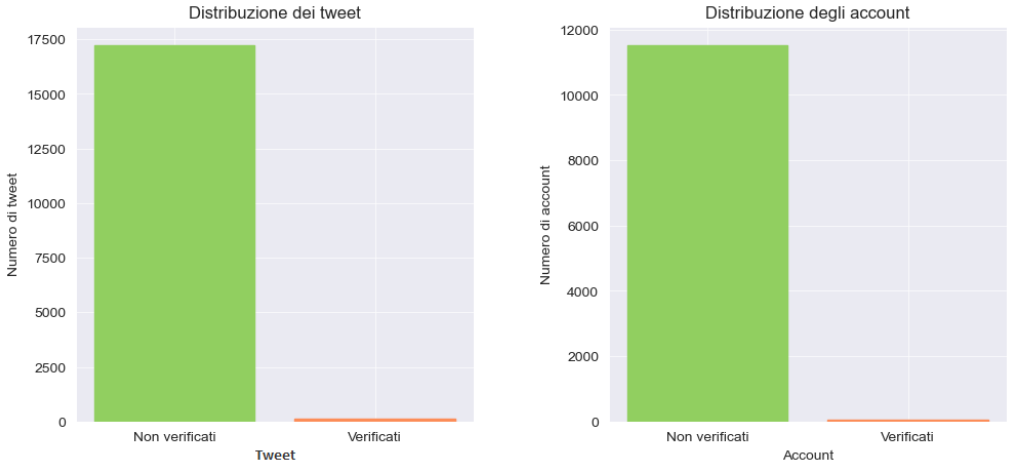
Prima di procedere con l'analisi del dataset ottenuto, è stato necessario operare una prima fase di ***data cleaning*** sia sulle features che sui records presenti. Difatti, data l'incapacità da parte di Twint di reperire alcune informazioni, come il luogo di pubblicazione del tweet, al fine di evitare un dataset composto da molti valori mancanti si è reso necessario eliminare molte colonne. Si è inoltre provveduto a scartare tutte quelle features costituite da valori pressoché uguali o non

[3]Progetto disponibile e consultabile all'interno della seguente Repository su GitHub: `https://github.com/twintproject`

funzionali alla nostra analisi, come *timezone* o *language*, e records contenenti valori duplicati. In seguito a questa fase di pulizia, il dataset risulta ora composto da **15 colonne** (Tabella 1) e **17334 righe**. Come è possibile vedere dalla Figura 1 il dataset in questione è caratterizzato da un numero sproporzionalmente sbilanciato di tweets provenienti da account non verificati.

Tabella 1: Features eliminate (valori nulli e inutili) dal dataset durante la fase di pulizia e features mantenute

Feat. valori nulli	**Feat. inutili**	**Feat. mantenute**
near	conversation_id	id
geo _id	thumbnail	place
source	language	date
user_rt_id	video	time
retweet_id	photos	user_id
user_rt	created_at	username
retweet_date	timezone	name
translate	cashtags	tweet
trans_src	urls	mentions
	retweet	replies_count
	quote_url	retweets_count
	link	likes_count
		hashtags
		reply_to
		verified



(a) **(b)**

Figura 1: Distribuzione degli account verificati e non (1b) e dei loro tweet (1a)

In questa prima fase di analisi è stato ritenuto opportuno studiare l'andamento del numero dei tweet pubblicati giorno per giorno nell'intervallo temporale in cui l'indagine si è concentrata. L'andamento nel grafico (Figura 2) mostra 8 fasi principali: 4 picchi e 4 valli. Entrati nel periodo del primo lockdown, è possibile assistere ad una prima fase di crescita molto ripida della curva, a testimonianza di un aumento esponenziale nel numero di tweets fino a raggiungere un primo picco nel mese di aprile (picco relativo al periodo marzo-maggio). La seconda fase corrisponde a una marcata discesa della curva che vede il suo minimo nel mese di agosto, ciò è causato probabilmente dall'arrivo delle vacanze estive e dunque da una chiusura delle scuole. Una terza fase è invece in prossimità della ripresa delle scuole che conferma un numero di tweets molto alto

e poco stabile. Il secondo picco quindi è da attribuire ad un periodo che va da ottobre a gennaio con massima frequenza a novembre. Questo stesso trend si ripete poi per l'anno 2021-2022, fino ad una totale discesa nel marzo del 2022 a indicare una ripresa della normale didattica in presenza. Questi aspetti sono perfettamente in linea con le aspettative circa il comportamento degli individui sui *social*: l'interesse pubblico è più alto, quanto più cruciale è l'avvenimento in discussione.

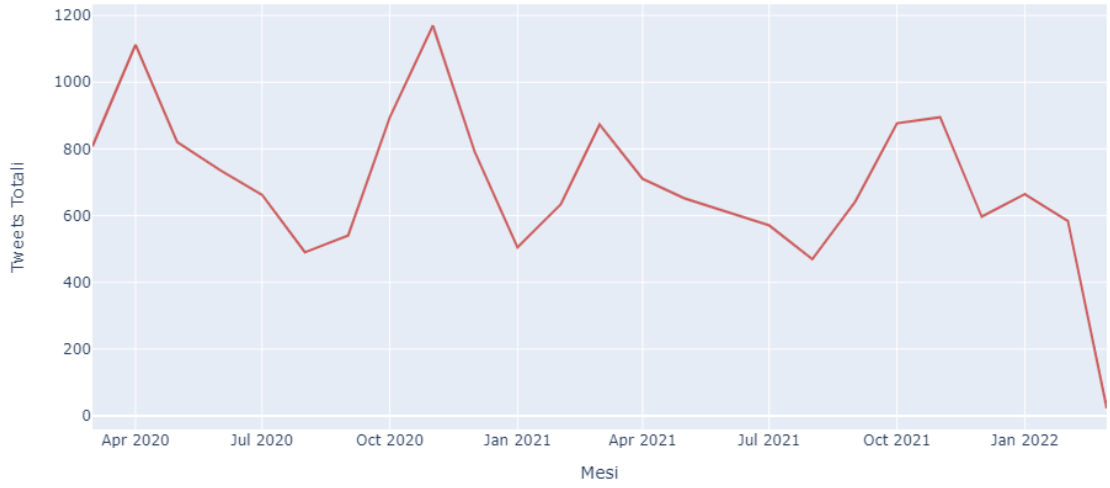


Figura 2: Andamento temporale dei tweets

Il corpus è stato infine costruito a partire dall'estrazione dei tweet presenti nel dataset finora descritto. Prima di poter utilizzare le metodologie inizialmente presentate è stato necessario eseguire alcune **operazioni di *preprocessing*** sui dati linguistici, ripulendoli quindi da tutti quegli elementi che possono in qualche modo alterare le successive elaborazioni computazionali:

- **rimozione dei simboli superflui come hashtags e menzioni** tramite l'utilizzo delle espressioni regolari (re);
- **rimozione di eventuali doppi spazi, links, emoticons e segni di punteggiatura** non fondamentali.
- **normalizzazione del testo** trasformandolo in minuscolo.

3 Elaborazione computazionale base

La prima fase del progetto ha richiesto l'utilizzo della libreria Python **NLTK**[4] (Natural Language Toolkit) per *Natural Language Processing* e della libreria **spacCy**[5] necessaria per poter applicare alcuni task su corpus in lingua italiana, tra cui il POS tagging e il NER.

Le informazioni base sono state estratte seguendo la pipeline tradizionale proposta per NLP: si è provveduto innanzitutto a segmentare il testo in frasi (***sentence splitting***) così da poter studiare il numero medio di frasi per ciascun tweet, successivamente, al fine di ottenere per

[4]Libreria open source disponibile e consultabile ai seguenti link: `https://www.nltk.org/`, `https://www.nltk.org/book/`

[5]Libreria open source disponibile e consultabile sul sito web: `https://spacy.io/`

ciascuna frase i token in esse contenuti, ovvero le unità atomiche dell'analisi linguistica, è stata applicata la fase di **tokenizzazione**; da tenere presente che per token non si intende solo le parole ma anche i segni di punteggiatura, date, numeri ecc.[6]. Per poter poi effettuare alcune analisi statistiche base, come la frequenza delle parole, è stata applicata la tecnica di **lemmatizzazione** in cui a ciascun token viene associato il suo lemma, ovvero la categoria lessicale astratta che accomuna tutte le forme flesse a esso riconducibili.

Alla fine di questa preliminare fase il corpus è stato sottoposto ad **analisi morfosintattica** dei tweet nonché semantica, in particolare attraverso l'utilizzo della tecnica *Named-Entity Recognition* (*NER*), che vuole estrarre e associare ai nomi di entità presenti nel testo una categoria semantica predefinita come, ad esempio, nomi di persona, di organizzazioni o di luoghi.

3.1 Analisi di base

In prima istanza, una volta effettuato il *sentence splitting*, si è deciso di studiare il numero medio, massimo e minimo di frasi presenti in ciascun tweet, effettuando poi un'ulteriore distinzione tra i tweet provenienti da account verificati e non.

Osservando i risultati ottenuti (Tabella 2) è possibile evincere quanto, mentre il **numero minimo di frasi** è lo stesso (pari a **1**), il **numero medio di frasi** risulta essere maggiore per i tweet provenienti da account verificati (pari a 2.4) rispetto a quelli non verificati (pari a 1.8); risultato non affatto sospendente se si pensa quanto tali utenti, essendo spesso personaggi di spicco o comunque conosciuti dal grande pubblico, abbiamo maggior remore nel presentare tweet completi e sufficientemente argomentati.

	N° Medio	**N° Minimo**	**N° Massimo**	**N° Totale**
Account verificati	2.4	1	7	264
Account non verificati	1.84	1	18	31722
Corpus Totale	1.85	1	18	31984

Tabella 2: Analisi statistiche sul numero medio, minimo, massimo e totale di **frasi** nel corpus

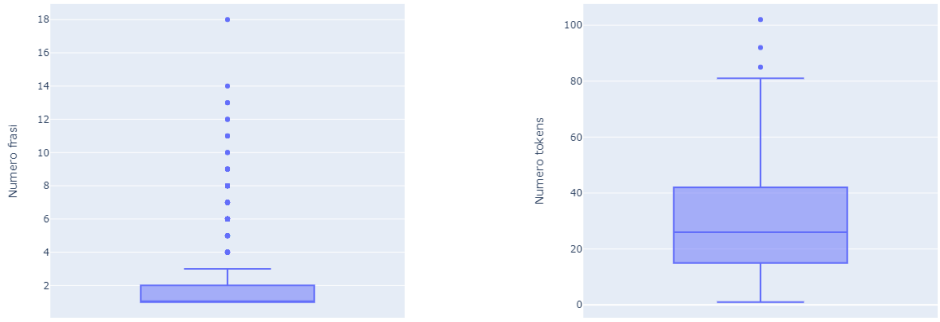
Questo stesso trend è infatti confermato se, dopo aver effettuato la tokenizzazione, andiamo ad analizzare il numero medio di token per tweet (Tabella 3). Anche in questo caso si ha un **numero medio di token** più alto per gli account verificati (pari a 30.12) rispetto a quelli non verificati (28.36). Lo stesso non si può dire se si effettua un confronto tra il comportamento mostrato dal numero minimo di frasi e il **numero minimo di token**, dove si può notare un valore nettamente più alto per gli account verificati (pari a 6) rispetto a quelli non verificati (pari a 1).

Una variazione alquanto interessante si ha se ci volgiamo verso il **numero massimo di frasi e tokens**: in entrambi i casi si hanno valori nettamente differenti tra verificati e non. Ciò potrebbe indicare la presenza di *outliers*, supposizione che viene confermata osservando i boxplot della distribuzione del numero di frasi e tokens nel corpus (Figura 3)

[6](Lenci et al., 2016)

	N° Medio	N° Minimo	N° Massimo	N° Totale
Account verificati	30.12	6	57	3313
Account non verificati	28.36	1	102	488515
Corpus Totale	28.37	1	104	491828

Tabella 3: Analisi statistiche sul numero medio, minimo, massimo e totale di **token** nel corpus



(a) **(b)**

Figura 3: Distribuzione delle frasi (3a) dei tokens (3b) nel corpus

3.2 Analisi statistica

In questa sezione ci concentreremo sul fornire un'elaborazione statistica del corpus oggetto dell'analisi. In particolare è possibile stabilire le parole più frequenti e quantificare la ricchezza lessicale, ovvero il rapporto tra vocabolario (parole tipo) e numero di tokens.

La rappresentazione *World Cloud*[7] in figura 4 mostra come, considerando l'arco temporale totale, la maggior parte delle parole più frequenti siano maggiormente correlate con la scuola e la didattica a distanza (*scuola*, *studenti*, *dad*, *docenti* ecc.).

Figura 4: Word Cloud delle parole più frequenti, maggiore è la dimensione più frequenti sono le parole.

[7]La Word Cloud è una tecnica di visualizzazione grafica delle parole contenute in un testo, classificate secondo il numero di occorrenze.

É tuttavia più interessante osservare come le parole varino se si considerano sottoinsiemi differenti del corpus (Figura 5). In particolare, nonostante in tutti i casi le prime tre parole si ripresentino con frequenze proporzionalmente simili, è possibile notare come lo stesso non accade per le successive. Come si può vedere in figura 5a, se si considera l'arco temporale relativo al periodo di primo lockdown (marzo-maggio 2020), all'interno del social si è discusso maggiormente del dovere di rimanere a casa (***iorestoacasa***) e di proseguire con gli studi anche se in dad (***lascuolanonsiferma***), confermando un generale un senso positivo nei confronti della didattica a distanza. Analizzando il periodo immediatamente successivo in figura 5b, si può notare come tale solidarietà vada sempre più a scemare lasciando spazio alla necessità di tornare in presenza, come mostrato dalla parola ***presenza*** in settima posizione. Si può dedurre che questo sempre minor interesse per la didattica a distanza, che stava lasciando sempre più spazio alla didattica in presenza, sia dovuto all'introduzione della Didattica Digitale Integrata. Da marzo 2021 a marzo 2022 questo trend si conferma, tanto che nel grafico in figura 5c la parola *didatticaaditanza* lascia il posto alla parola ***presenza*** dopo la decisione da parte del governo di ripristinare ad aprile 2021 le lezioni in presenza per il 60% .

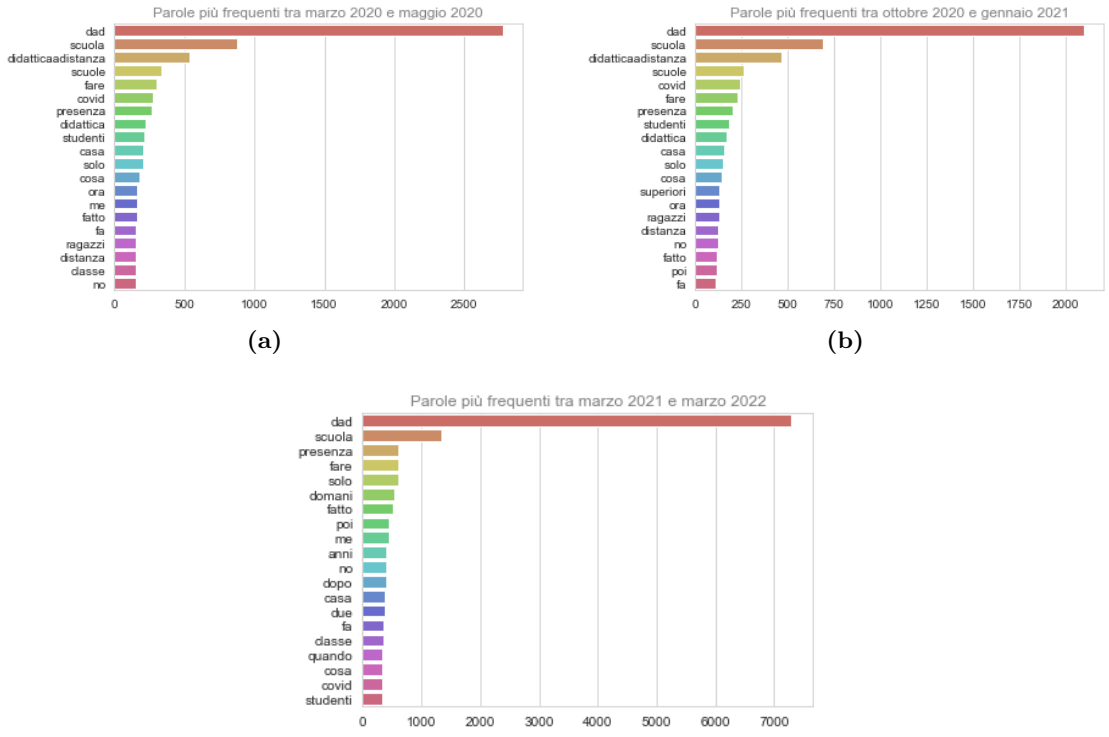


Figura 5: Distribuzione delle 20 parole più frequenti durante il primo lockdown, (5a) il periodo immediatamente successivo ottobre-gennaio (5b) e i restanti mesi (5c)

Al fine di analizzare graficamente la distribuzione di alcune delle parole più frequenti all'interno del corpus di tweet, si è fatto uso del grafico di **dispersione lessicale** in figura 6. Si è deciso di inserire le parole più interessanti viste finora, in quanto argomenti di spicco nel corso delle varie fasi. Dal grafico di Dispersione lessicale è visibile quanto siano stati utilizzati contemporaneamente in stessi tweet hashtag più legati ad una prima fase del *lockdown* come *didatticaadistannza iorestoacasa* e *lascuolanonsifera*, mentre il termine *presenza*, nonostante sia quasi sempre onnipresente in tutti i tweet, risulta essere speculare ai primi, a indicare un maggior utilizzo solo nelle fasi successive, senza eccessive sovrapposizioni.

Un ulteriore indice statistico del testo è la **ricchezza lessicale**, misurabile grazie al metodo

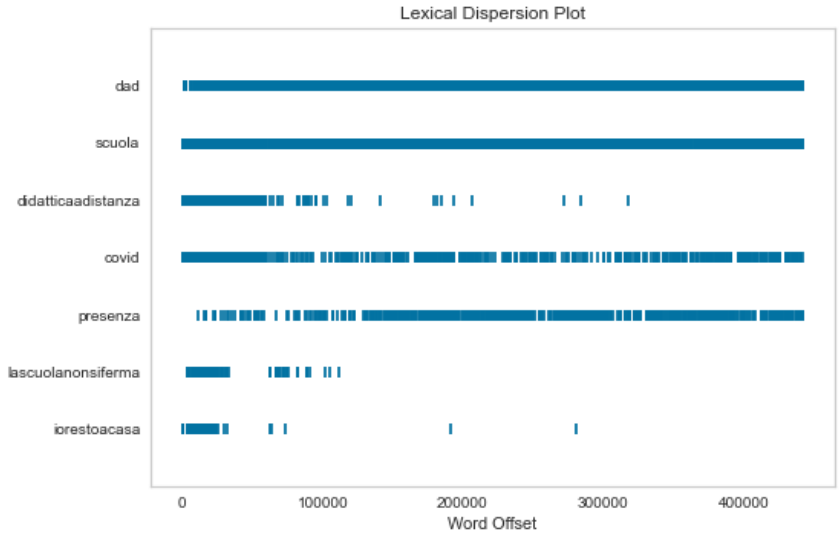


Figura 6: Dispersion plot delle parole *dad*, *scuola*, *didatticaadistanza*, *covid*, *presenza*, *lascuolanonsiferma*, *iorestoacasa*

Type Token Ratio, ovvero il rapporto tra *tipo-unità*, che permette di individuare la varietà del vocabolario del corpus. Come è possibile vedere dal grafico in figura 7 tale indice si concentra intorno al valore **0.9-1.0** a indicare che il vocabolario di ciascun tweet tende ad essere vario. Esistono tuttavia tweet aventi minor ricchezza lessicale, in cui il numero di parole distinte è il **65%** del numero totale delle parole, ma, essendo nella coda della distribuzione, rappresentano solo una minoranza. Successivamente si è andato ad osservare questo stesso indice su tweet provenienti dalle due diverse tipologie di account: in entrambi i casi si registra un valore molto alto, **0.92** per i verificati e **0.91** per i non.

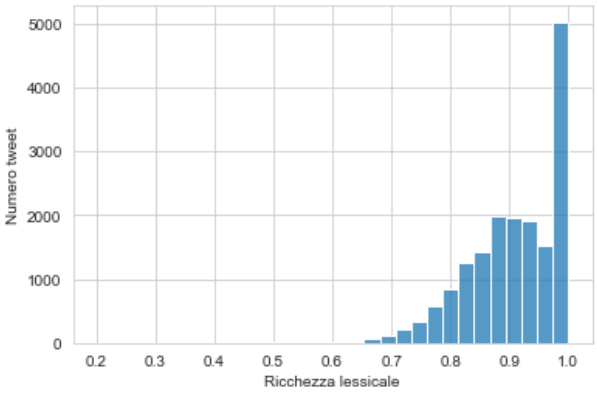


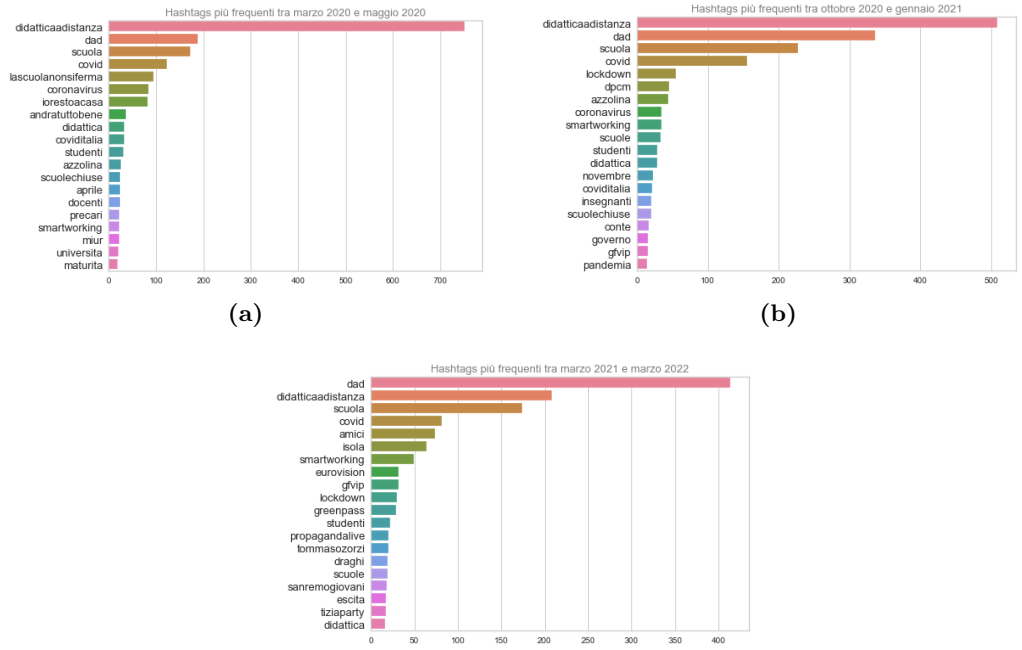
Figura 7: Distribuzione della ricchezza lessicale nei tweet

3.3 Analisi degli hashtags

Per approfondire ulteriormente la questione ed avere una prospettiva ancor più completa, abbiamo infine affiancato alle analisi finora condotte lo studio della frequenza sugli hashtags (Figura 8) dal quale è possibile evincere gli argomenti più discussi, anche in questo caso introducendo il confronto tra le varie fasi.

Interessante è stato notare come il comportamento precedentemente descritto (sottosezione 3.2) si ripresenti qui in maniera quasi del tutto analoga, tanto che, se nella prima fase si ha un atteggiamento più positivo nei confronti della DAD con hashtag come *lascuolanonsiferma* e

nella terza fase, si nota invece una **perdita di interesse** da parte degli utenti nei confronti della didattica a distanza. Infatti agli hashtag, in linea con tale argomento come *dad* o *didatticaadistanza*, parole chiave della ricerca dei tweet, si associano hashtag completamente discordanti come *eurovision*, *gfvip* ecc. Altro aspetto degno di nota è la presenza dell'hashtag *greenpass*, argomento di forte dibattito probabilmente a causa dell'imposizione in quel periodo da parte del governo di tale certificazione per poter frequentare diversi luoghi, come appunto le scuole.



(a) **(b)** **(c)**

Figura 8: Distribuzione dei 20 hashtags più frequenti durante il primo lockdown, (8a) il periodo immediatamente successivo ottobre-gennaio (8b) e i restanti mesi (8c)

3.4 Part-of-Speech Tagging

Questa fase del lavoro si concentrerà sul profilo morfo-sintattico del testo. Per fare ciò è stato necessario effettuare un'annotazione morfo-sintattica, al fine di ottenere alcuni elementi statistici di base come il numero medio di parole semanticamente piene come nomi, aggettivi e verbi, in quanto indici del grado di informatività del testo: come definito in (Montemagni, 2013), in generale testi costituiti da un maggior numero di nomi saranno più esplicativi. Nel corpus, si registra una media di **6.4 sostantivi** per tweet a discapito di aggettivi (**1.9**) e verbi (**3.4**). Considerando anche account verificati e non è possibile constatare anche qui un'alta discrepanza, con in media **8.5 sostantivi** per i primi rispetto ai **6.4** per i secondi.

Si è infine analizzata la **densità lessicale**, ovvero la proporzione di parole semanticamente piene (ovvero portatrici di significato come nomi, aggettivi, verbi e avverbi) rispetto alla totalità delle parole, che, come definito in (Johansson, 2008), fornisce un ulteriore indice sul grado si informatività del testo. In genere un testo con una maggiore quantità di parole piene sarà più informativo di un testo con maggiore proporzione di parole funzionali. Dal grafico in figura 9 risulta subito evidente una distribuzione normale della densità lessicale nei tweet, con la maggior parte di essi avente un valore intorno alla media pari a **0.4**. Tenendo conto che una densità lessicale bilanciata è di circa il 50%, quindi con metà di ogni frase composta da parole lessicali e metà da parole funzionali, tale valore risulta essere in linea con quanto è stato affermato in (Johansson, 2008) secondo cui di norma il parlato tende ad avere una più bassa densità rispetto

alla scritto. Difatti è possibile considerare i tweet, in quanto più colloquiali e informali, alla stregua del parlato.

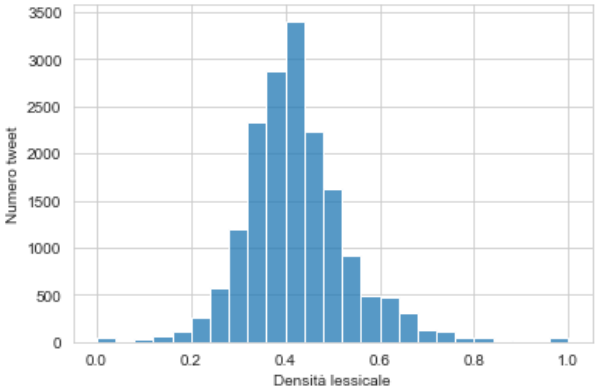


Figura 9: Distribuzione della densità lessicale nei tweet

3.5 Named-Entity Recognition

Vista la difficoltà nell'annotare un corpus con il significato specifico tratto da un dizionario, per questa fase del lavoro si è deciso di adottare la tecnica *Named-Entity Recognition* (NER) che vuole localizzare e identificare certe occorrenze di parole o espressioni come appartenenti a predeterminate categorie *Named Entities*. Come definito in (Marrero et al., 2013) tale task rappresenta la base per molte altre aree fondamentali nel campo dell'*Information Management* come appunto l'*Opinion Mining*.

Per poter effettuare NER è stata utilizzata la libreria spaCy che va ad assegnare categorie semantiche ai tokens di un corpus. Una volta processato il corpus oggetto di analisi, spaCy è riuscito a identificare in totale 28240 entità tra persone PER, località LOC, organizzazioni ORG ed entità miste MISC (*eventi*, *nazionalità*, *prodotti* o *opere d'arte*).

Osservando la tabella 4 è possibile vedere tuttavia delle forti difficoltà da parte del classificatore di identificare in maniera corretta persone, le quali vengono spesso riconosciute erroneamente come località o aziende, e anche termini legati al COVID-19 e alla DAD i quali vengono talvolta etichettati come persone o località. Analizzando invece i luoghi identificati dall'algoritmo possiamo osservare una forte presenza di diverse **regioni e città italiane soprattutto del nord e sud Italia**, realtà spesso al centro del dibattito politico e pubblico in particolare in merito al grande divario economico, sociale e strutturale che le viene attribuito. Come affermato in (Bazzoli et al, 2021) tale diseguaglianza è stata nuovamente tirato in ballo con l'imposizione forzata della didattica a distanza; tra le cause di tale problematica si annoverano l'impossibilità di molti studenti, soprattutto provenienti dal Centro-Sud, di accedere ad una buona connessione internet, lo scarso possesso di strumenti tecnologici e la generale bassa partecipazione alle lezioni a distanza.

Oltre al nome **DAD**, risultano degne di nota anche alcune entità classificate come MISC come **DPCM** (*Decreto del Presidente del Consiglio dei ministri*) e **DDI** (*Didattica Digitale Integrata*)

LOC	Occorenza	MISC	Occorrenza
COVID	196	DAD	1880
Italia	192	PC	71
Roma	60	Natale	55
Milano	52	YouTube	53
Campania	43	MiurSocial	46
Puglia	39	CottarelliCPI	41
Lombardia	37	Riforma	30
Europa	34	DPCM	29
Napoli	25	DDI	19

Tabella 4: Località ed entità miste ottenute dal classificatore NER

Interessante è anche la possibilità data da spaCy di visualizzare graficamente le etichette assegnate ai token dal classificatore, come è possibile vedere nell'esempio di seguito riportato (Figura 10)

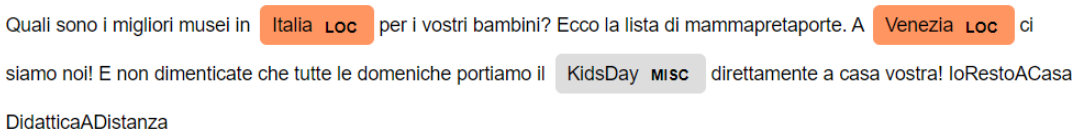


Figura 10: Esempio di frase con associate etichette di entità

4 Sentiment Analysis

4.1 Il modello *FEEL-IT*

La seguente analisi consta dell'utilizzo della tecnica di **Sentiment Analysis**, disciplina a volte definita anche come *Opinion mining*. Tale task si occupa di studiare, identificare ed infine estrarre a partire da risorse testuali opinioni, sentimenti e atteggiamenti degli utenti nei confronti di qualche evento o tematica discussa.

Visto il corpus interamente in italiano è stato necessario individuare una libreria che fosse stata appositamente implementata per la lingua Italiana. Per questo motivo è stato deciso di utilizzare la libreria *open source* FEEL-IT[8] sviluppata da ricercatori dell'Università Bocconi in seno alla realtà *Hugging face*, un'azienda con la missione di democratizzare l'accesso ai sistemi di NLP, permettendo così un continuo sviluppo di tecnologie del mondo delle Intelligenze Artificiali.

La scelta di tale libreria è ulteriormente avvalorata dal tipo di dato su cui sono stati addestrati i modelli. Si tratta di un corpus di benchmark costituito da post di Twitter annotati con quattro emozioni base: *rabbia, paura, gioia, tristezza*. I tweet, in particolar modo, spaziano su temi molto attuali e spesso affini al nostro dataset in analisi, come il COVID-19 e la scuola.

Come definito in (Bianchi et al., 2021) il progetto ha visto la creazione di un corpus di dati per la predizione di sentimenti ed emozioni, successivamente utilizzato per addestrare un modello *BERT* (*Bidirectional Encoder Representation*) per l'italiano, *UmBERTo*, ottenendo così prestazioni ottime su diversi corpora di benchmark.

[8]Libreria disponibile e consultabile all'interno della seguente Repository su GitHub: `https://github.com/MilaNLProc/feel-it`

4.2 Applicazione di *FEEL-IT*

Come primo task è stata utilizzato il modello **Sentiment Classifier** di FEEL-IT con lo scopo di identificare la polarità dei tweet del corpus di riferimento così da meglio comprendere il generale atteggiamento di **positività** o **negatività** espresso dagli utenti per mezzo dei tweet in merito alla Didattica a Distanza. Oltre ad uno studio sul corpus intero, si è voluto approfondire ulteriormente l'indagine confrontando i sentimenti scaturiti nelle tre fasi distinte già studiate precedentemente.

Mediante l'ausilio della seguente rappresentazione grafica (Figura 11) è stato possibile stabilire il peso che ciascun tipo di sentimento ha in relazione ai tweet di riferimento. In particolare analizzando la totalità del corpus risulta immediatamente evidente il profondo distacco tra i **5006** tweet ottimistici e i **12328** tweet considerati pessimistici. Non è affatto sorprendente quindi che la gran parte degli utenti, autori dei tweet in questione, abbia sentimenti negativi nei confronti della tematica affrontata.

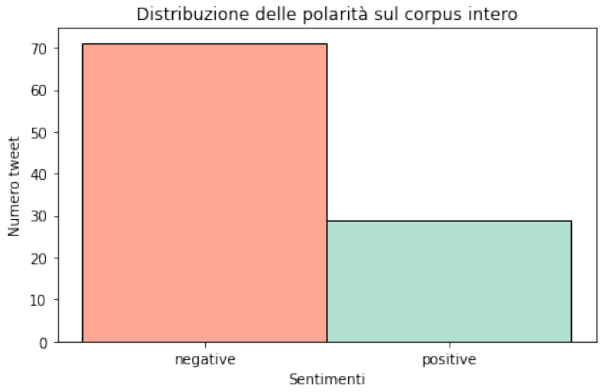


Figura 11: Percentuali di tweet negativi e positivi nell'intero corpus

Tuttavia lo studio delle polarità sui diversi periodi (Tabella 5) mostra una maggiore variabilità, che conferma quanto già evidenziato nelle precedenti analisi: nel periodo di primo lockdown il distacco tra sentimenti negativi e positivi va decisamente ad attenuarsi con circa il **54%** di tweet negativi e **43%** di tweet positivi; nel periodo immediatamente successivo invece si nota un improvviso cambio di rotta, con un **79%** di tweet negativi e solo un **21%** di tweet positivi. Questo stesso trend viene poi confermato nel periodo finale marzo 2021-marzo 2022, con un **73%** di tweet pessimistici e un **27%** di tweet ottimistici.

	Tweet Positivi	**Tweet Negativi**	**Tweet Totali**
Marzo 2020 - Maggio 2020	801	1016	1880
Ottobre 2020 - Gennaio 2021	540	1989	2529
Marzo 2021 - Marzo 2022	2141	5800	7941
Corpus intero	5006	12328	17334

Tabella 5: Polarità dei tweet del corpus

In un secondo momento si è deciso di sfruttare la capacità del modello **Emotion Classifier** di FEEL-IT di individuare in maniera più granulare le emozioni che emergono dai tweet, come

paura, **gioia**, **tristezza** e **rabbia**, anche in questo caso dapprima studiando l'intero corpus e poi spostandoci verso sottoinsiemi di esso. Come affermato in (Bianchi et al, 2021) esiste un enorme differenza tra l'essere arrabbiati o l'essere spaventati o tristi, sentimenti difatti tutti classificati inizialmente come negativi.

	Tweet Gioia	**Tweet Rabbia**	**Tweet Paura**	**Tweet Tristezza**
Marzo 2020 - Maggio 2020	687	571	109	450
Ottobre 2020 - Gennaio 2021	467	1180	198	684
Marzo 2021 - Marzo 2022	1973	3513	580	1875
Corpus intero	4536	7281	1250	4277

Tabella 6: Distribuzione delle emozioni nei tweet del corpus

Anche in questo caso, se si indaga sull'intero corpus a nostra disposizione (Figura 12) notiamo quanto il sentimento di **rabbia**, circa **7281 tweets**, negativo per antonomasia, predomini su tutti gli altri, seguito poi dall'emozione di gioia, quasi a pari merito con il sentimento di tristezza, e infine dal sentimento di paura in percentuale nettamente minore.

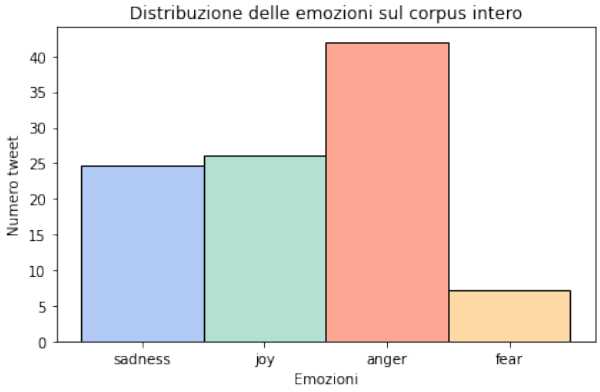
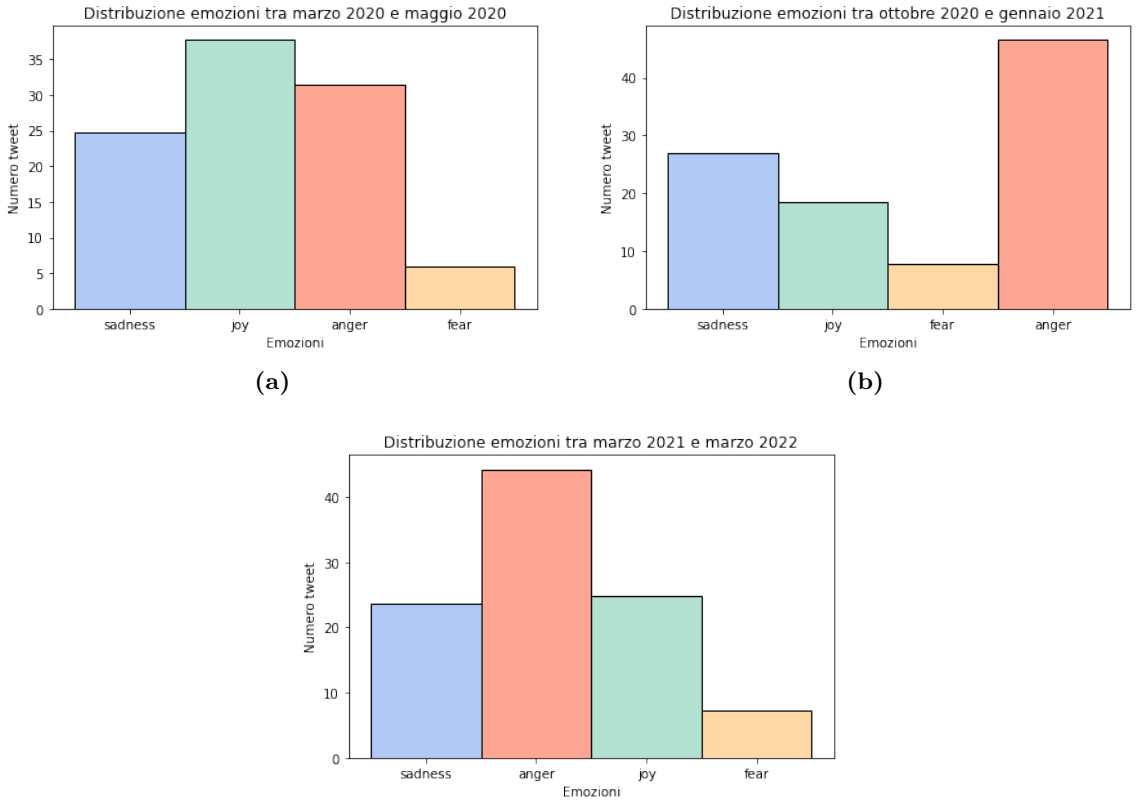


Figura 12: Distribuzione delle emozioni sul corpus totale

Analizzando tutti e tre i periodi (Figura 13) notiamo un trend piuttosto attinente con quanto già studiato: mentre nel periodo di primo lockdown è possibile osservare un numero maggiore di tweet pervasi dal sentimento di gioia rispetto agli altri, in entrambi i periodi successivi si assiste invece ad una inversione di tendenza, con un numero estremamente alto di tweet con sentimento di rabbia e numeri piuttosto simili di tweet indicati come felici o tristi.

Per approfondire ulteriormente la questione ed avere una prospettiva ancor più completa, abbiamo affiancato a tali analisi le **rappresentazioni *Word Cloud*** più interessanti (Figura 14) dalle quali è possibile evincere le parole più frequentemente usate e la loro accezione negativa/positiva. Interessante è stato notare nella rappresentazione riferita al periodo marzo 2021-marzo 2022 (Figura 14a) la presenza di alcune parole fortemente negative legate al mondo scolastico non per forza digitale come la stessa *scuola*, ma si è verificato anche un netto aumento della frequenza con cui si è discusso su Twitter di altre tematiche legate al COVID-19, come *vaccinati*, *quarantena*, *lockdown* e *casa*. Si può dedurre che in quel periodo, venisse meno la necessità di discutere della didattica a distanza, ormai quasi soppiantata dalla didattica digitale integrata, ma anzi che vi fosse un maggiore interesse verso la campagna vaccinale che si stava



(a) **(b)**

(c)

Figura 13: Distribuzione emozioni nei tweet durante il primo lockdown, (13a) il periodo immediatamente successivo ottobre-gennaio (13b) e i restanti mesi (13c)

aprendo in questa fase, seguita dall'obbligatorietà che si voleva imporre nel mondo lavorativo come, nel caso della scuola, al corpo docenti.

(a) **(b)**

Figura 14: Rappresentazione Word Cloud delle parole più frequenti sia con accezione negativa relative al periodo marzo 2021-marzo 2022 (14a) che con accezione positiva relative al periodo di primo lockdown (14b)

5 Topic Modeling

5.1 Il modello *BERTopic*

Twitter è la piazza digitale delle persone, dove possono esprimere liberamente le proprie opinioni e preoccupazioni senza eccessivi vincoli. Una volta ottenute le informazioni sul generale

atteggiamento negativo nei confronti della didattica a distanza, si è cercato di intercettare e sviscerare il più possibile gli argomenti e le tematiche di maggior rilievo che più attanagliano il popolo dell'internet in merito proprio a questo nuovo modo di fare scuola.

L'ultimo task affrontato consiste quindi nell'applicazione di una tecnica di elaborazione automatica del testo impiegata per l'estrazione degli argomenti presenti all'interno di risorse testuali, raggruppandoli in clusters o gruppi di argomenti; tale tecnica prende il nome di **Topic Modeling**. Tale analisi è stata condotta mediante l'utilizzo di una particolare tecnica di *Topic Embedding* derivata da BERT, ossia il **BERTopic** [9].

5.2 Applicazione di *BERTopic*

Il modello BERTopic è stato applicato su due sottoinsiemi del corpus originale, uno relativo ai tweet classificati nella sezione precedente (Sezione 4) come negativi e l'altro come positivi. Tale distinzione dovrebbe permettere una migliore comprensione di ciò che rende la didattica a distanza secondo gli utenti di Twitter un fallimento o un successo come modello educativo. Da ciascun sottoinsieme sono stati estratti i venti argomenti più rilevanti; l'interazione tra gli argomenti sarà il criterio discriminante per valutare la bontà dei clusters individuati.

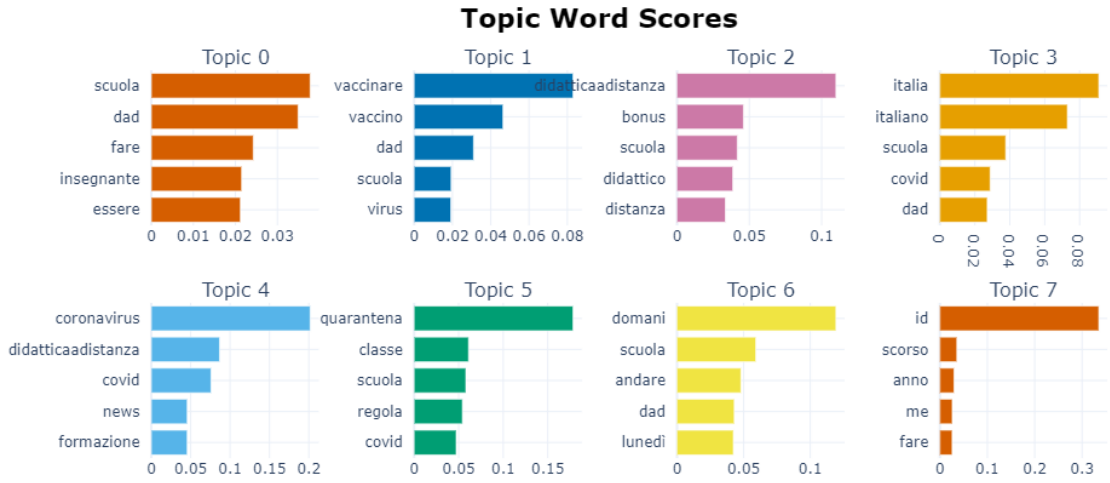


Figura 15: Distribuzione delle parole più probabili nei top 7 argomenti più discussi nei tweet negativi

La raffigurazione mediante *barcharts* (Figura 15) mostra gli argomenti maggiormente discussi nei tweet negativi nonché le parole chiave più probabili e significative per ciascun argomento. Come si può notare, gli argomenti più trattati, otre alla didattica a distanza in generale, sono in primis i ***vaccini***, nonché il senso di ***precarietà***, che ha caratterizzato tutti gli ambiti della nostra vita durante questo periodo di pandemia, e la ***quarantena*** in stretto legame con la possibilità di seguire le lezioni a distanza o in presenza. Tutto ciò è probabilmente correlato a quanto evidenziato nella precedente analisi effettuata sul corpus in merito al periodo marzo 2021-marzo 20202 (Sezione 4), dove si era notato un aumento di termini che fanno riferimento ai vaccini e alla quarantena imposta. Tra i tweet classificati come positivi invece emergono argomenti legati alle ***videolezioni*** e ***smartworking*** in generale, nonché alle tecnologie ad essi collegate come ***tablet*** e ***PC***. Ciò risulta assai interessante se si pensa come in (Bazzoli et al, 2021) si afferma

[9]Progetto disponibile e consultabile all'interno della seguente Repository su GitHub: `https://github.com/MaartenGr/BERTopic`

piuttosto come tali strumenti di supporto in determinate fasce della popolazione abbiano rappresentato più che altro un impedimento per molti studenti e docenti.

In seguito ai risultati ottenuti si è voluto verificare quale fosse l'andamento temporale degli argomenti se si considera l'intero corpus in esame. Il risultato di tale operazione, osservabile in figura 16, mostra come la frequenza dei tweets riferiti ad uno specifico tema sia influenzata dalla fase in cui si trova la pandemia; gli argomenti sono dunque in stretta relazione con gli eventi che si stanno verificando. In particolare, se si osservano l'argomento legato ai *vaccini* (topic 0), questo risulta essere sempre più di spicco man mano che procede la campagna vaccinale e la conseguente obbligatorietà imposta dal governo. L'argomento *DAD* (topic 2) dal canto suo riscuote notevole interesse nel periodo relativo al primo lockdown, ma col passare del tempo, a seguito dell'introduzione della DDI, perde sempre più rilevanza, in maniera quasi speculare all'argomento relativo alla *presenza* (topic 8)

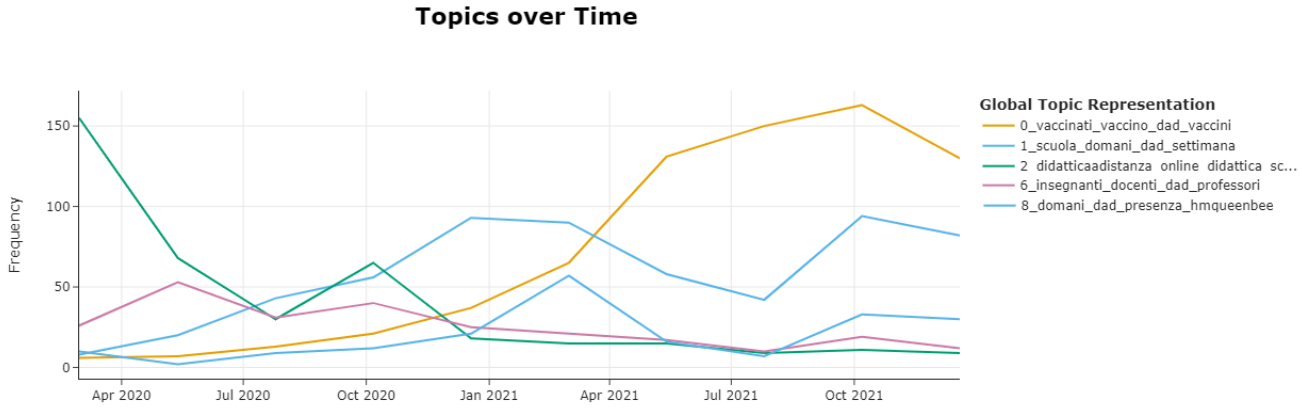


Figura 16: Andamento temporale degli argomenti nei tweet

6 Conclusioni

Tale progetto ha riguardato l'utilizzo di tecniche di linguistica computazionale e di due modelli di machine learning per lo svolgimento dei task di *Sentiment analysis* e *Topic modeling* su un corpus di tweets estratti relativi alla tematica della didattica a distanza. Si è deciso di operare le seguenti analisi sia sul corpus nella sua interezza, che su parti di esso così da analizzare con maggior dettaglio il picco di pubblicazione di tweets relativo al primo periodo di lockdown, differenziandolo dalle fasi successive della pandemia.

La prima parte del lavoro ha riguardato la creazione tramite tecnica di *scraping* di un dataset, successivamente ripulito da informazioni ridondanti o inutili. A partire da tale dataset è stato così ottenuto il corpus di riferimento costituito da risorse testuali su cui sono stati eseguiti tutti i passi richiesti dalla tradizionale pipeline del *Natural Language Processing*. Si parla quindi di sentence splitting, tokenizzazione, lemmatizzazione, *Part-Of-Speech tagging* e analisi semantica tramite *Named-entity Recognition*. A partire da questa prima fase, sono emersi risultati piuttosto interessanti in merito alle parole e agli hashtag più frequenti che hanno mostrato un'iniziale fase di interesse nei confronti della DAD che poi è andato scemando, lasciando lo spazio al altri temi ad esso scollegati.

La seconda parte di questa relazione è stata fondamentale in quanto, sin dall'inizio, l'intento di tale lavoro era quello di studiare i sentimenti, le emozioni e le tematiche che trapelavano dal tweet scritti dai vari utenti in merito al metodo educativo della didattica a distanza; si è

provveduto quindi alla presentazione ed analisi dei risultati ottenuti grazie al modello *FELL-IT* per la *Sentiment Analysis* e del modello *BERTopic* per il *Topic Modeling*.

Lo studio ha mostrato una forte presenza di opinioni molto sbilanciate verso una visione più negativa, pervase da un generale senso di rabbia; in particolare, la maggioranza di tali tweet riguardavano le vaccinazioni, l'obbligatorietà imposta e il senso di precarietà. Si è tuttavia notato un maggior senso di positività nei tweet relativi alle prime fasi della pandemia tanto da far emergere un maggior sentimento di gioia. Ciò potrebbe essere dovuto al senso di novità nei confronti della DAD nonché di vicinanza nel popolo italiano per poter fronteggiare la pandemia in generale. Infatti tutto ciò col passare del tempo ha lasciato spazio alle problematiche sopraccitate, probabilmente dovute ad un sistema scolastico di per sé caratterizzato da forti mancanze dal punto di vista infrastrutturale, nonché a gravi diseguaglianze che pervadono la Penisola. Si può quindi affermare senza timore di essersi schierati da una parte piuttosto che da un'altra e che tale argomento ha fatto emergere delle tematiche che vanno al di là della disputa interna tra didattica a distanza e in presenza.

Possibili sviluppi futuri potrebbero riguardare lo studio di un confronto diretto tra argomenti piuttosto correlati come *didattica a distanza*, *didattica in presenza* e *mista* (o DDI). Si potrebbe quindi estendere la ricerca al nuovo periodo scolastico 2022-2023, facendo riferimento al solo mondo universitario, dove la didattica a distanza potrebbe continuare a coesistere con la didattica in presenza.

Riferimenti bibliografici

[1] Alban Conto, C., Akseer, S., Dreesen, T., Kamei, A., Mizunoya, S. & Rigole, A. (2020) *COVID-19: Effects of School Closures on Foundational Skills and Promising Practices for Monitoring and Mitigating Learning Loss*. Centro di Ricerca Innocenti dell'UNICEF, Firenze.

[2] Bazzoli, N., Barberis, E., Carbone, D. & Dagnes, J. (2021). *La didattica a distanza nell'Italia diseguale. Criticità e differenze territoriali durante la prima ondata Covid-19*. Rivista Geografica Italiana.

[3] Bianchi F., Nozza D. & Hovy D. (2021). *FEEL-IT: Emotion and Sentiment Classification for the Italian Language*, 76-83.

[4] Bird, S., Klein, E. & Loper E. (2009). *Natural Language Processing with Python*. O'Reilly.

[5] Ferritti M. (2020). *Scuole chiuse, classi aperte Il lavoro di insegnanti e docenti al tempo della didattica a distanza*. Connessioni tra ricerca e politiche pubbliche, 64-76.

[6] Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*.

[7] Johansson, V. (2008), *Lexical diversity and lexical density in speech and writing: a developmental perspective*, 61-79. Lund University.

[8] Lamba M. & Margam M. *Sentiment Analysis*, 191–211.

[9] Lenci, A., Montemagni, S. & Pirrelli, V. (2016). *Testo e computer*. Carocci editore.

[10] Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato J. & Gómez-Berbís, J. M. (2013). *Named Entity Recognition: Fallacies, challenges and opportunities, Computer Standards & Interfaces*, 482-489.

[11] Mascheroni, G., Saeed, M., Valenza, M., Cino, D., Dreesen, T., Zaffaroni, L. G. & Kardefelt-Winther D. (2021). *La didattica a distanza durante l'emergenza COVID-19: l'esperienza italiana*. Centro di Ricerca Innocenti dell'UNICEF, Firenze.

[12] Monella, P. (2020). *Metodi digitali per l'insegnamento classico e umanistico*. EDUCatt.

[13] Montemagni, S. (2013). Tecnologie linguistico-computazionali e monitoraggio della lingua italiana. *Studi Italiani di Linguistica Teorica e Applicata (SILTA)*, 145-172.

Sitografia

[1] *Hugging Face: State of the Art Natural Language Processing in ten lines of TensorFlow 2.0 — The TensorFlow Blog*. Consultato il 29 agosto 2022, in `https://blog.tensorflow.org/2019/11/hugging-face-state-of-art-natural.html`

[2] *Sentiment Analysis and Emotion Recognition in Italian (using BERT)*. Consultato il 28 agosto 2022, in `https://towardsdatascience.com/sentiment-analysis-and-emotion-recognition-in-italian-using-bert-92f5c8fe8a2`

[3] *MilaNLProc/feel-it-italian-sentiment*. Consultato il 25 agosto 2022, in `https://huggingface.co/MilaNLProc/feel-it-italian-sentiment`

[4] *BERTopic*. Consultato il 29 agosto 2022, in `https://github.com/MaartenGr/BERTopic`

[5] *Natural Language Toolkit — NLTK 3.7 documentation*. Consultato il 25 agosto 2022, in `https://www.nltk.org/`

[6] *Twint project*. Consultato il 20 agosto 2022, in `https://github.com/twintproject/twint`

[7] *spaCy - Industrial-strength Natural Language Processing in Python*. Consultato il 20 agosto 2022, in `https://spacy.io/`