Università degli studi di
Milano-Bicocca

Social Media Analytics

Final Project

Un data analyst nel mondo social

Authors:
Matteo Carcano - 873258 - m.carcano7@campus.unimib.it
Marco Donzella - 829358 - m.donzella1@campus.unimib.it
Andrea Marinoni - 799690 - a.marinoni14@campus.unimib.it

12 febbraio 2022

1 Introduzione

Il tema della valutazione del proprio lavoro, del comprendere la propria community al meglio è sempre più centrale nel successo e nel perseguimento dei propri obiettivi sui social network. Il report tratterà l'analisi del canale YouTube Luke Barousse [1]. Luke è un ragazzo che si occupa di produrre contenuti e tutorial riguardanti il mondo data science. Il suo canale ha ormai più di 130,000 iscritti ed è in espansione; l'idea del progetto è quella di aiutarlo a meglio comprendere la sua community, in modo da migliorare e adeguare i suoi contenuti. Si è consapevoli di quanto questo sia un argomento complesso da trattare e di come non bastino le sole analisi effettuate nel report per fare a Luke una visione completa della situazione, ma sicuramente possono aiutarlo ad avere una prima infarinatura della sua situazione.

L'analisi proposta si divide in due macro aree, la prima si occuperà di community detection dove verranno costruiti dei grafi per comprendere come siano suddivisi gli utenti che usufruiscono del canale. Questa analisi verrà fatta dapprima su tutti i video e poi solo su un argomento trattato all'interno del canale.

La seconda parte di comprensione del canale YouTube invece tratterà gli spinosi argomenti della Sentiment Analysis e della Emotion Analysis. L'obiettivo è quello di scoprire se i contenuti soddisfino gli utenti di Luke, e se si, quali in maggior modo. Così da poter capire i video più interessanti prodotti e studiarne le caratteristiche che hanno convinto la community a commentarli entusiasti.

2 Dataset

Il dataset utilizzato consiste nella raccolta di commenti estratti dai video del canale YouTube di *Luke Barousse* tramite API key fornita da YouTube [2]. In particolare, sono stati raccolti 4975 commenti effettuati da 4017 utenti distribuiti tra i 65 video presenti nel canale. Oltre al testo dei commenti sono state raccolte ulteriori informazioni come l'autore del commento, il titolo del video commentato e i rispettivi ID per un totale di 12 variabili:

- comment;
- comment_id;
- author_url;

- author_name;
- reply_count;
- like_count;
- date;
- vidid;
- total_reply_counts;
- vid_title;
- just_name.

3 Social Network Analysis

Per comprendere le interazioni degli utenti con il canale YouTube, si è optato per l'utilizzo della teoria dei grafi. La teoria dei grafi è una branca della matematica che si occupa di reti di punti collegati da linee. Il termine grafo non si riferisce a grafici di dati come il nome potrebbe far sembrare, ma rappresenta il nome di una visualizzazione composta da un insieme di nodi e di archi. L'utilità dei grafi risiede nella loro capacità di modellare molte situazioni diverse. Al giorno d'oggi se si compie un'analisi riguardante i social media bisogna passare obbligatoriamente per la teoria dei grafi. In questo report verrà utilizzata per meglio comprendere le community del canale youtube, studiandone la struttura e la distribuzione. Un primo approccio con la teoria dei grafi è stato compiuto quasi a livello esplorativo, per comprenderne la struttura. Per prima cosa si è costruito un grafo dove gli autori dei commenti e i video rappresentano i nodi e sono stati collegati tra di loro. Successivamente si sono calcolate alcune metriche come il grado e la centralità dei vari nodi. Calcolando il grado dei nodi, ovvero il numero di archi incidenti, si è in grado di scoprire i video che hanno ottenuto più commenti e quindi più o meno engagement. In Figuara 2 vengono riportati i gradi dei video, mentre in Tabella 1 vengono riportati i primi 5 e ultimi 5 video per numero di commenti. Dalla Figura 2 si può inoltre notare come i video con meno commenti siano i primi pubblicati e che dopo il picco di 996 commenti i video iniziano ad avere più interazione tra gli utenti.

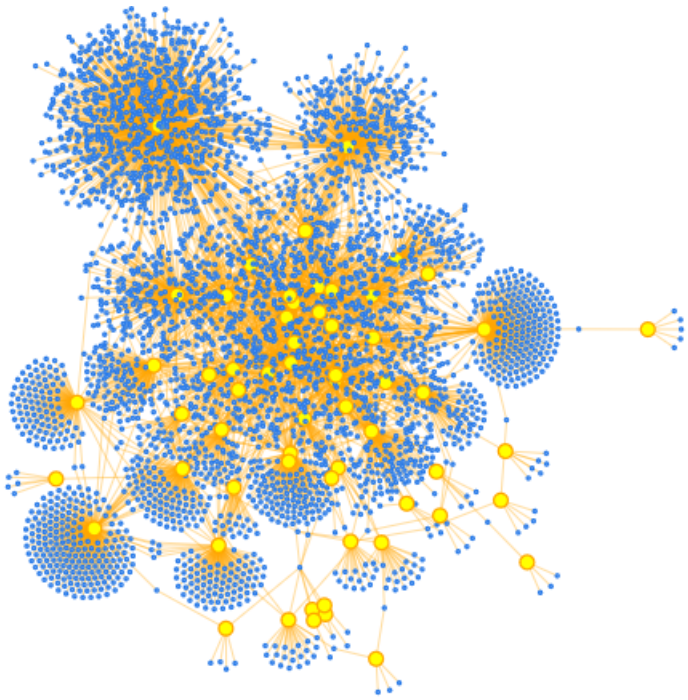

Figura 1: Grafo ottenuto collegando gli utenti ai video che hanno commentato. Per una maggiore interpretabilità i nodi dei video sono rappresentati in giallo, mentre i nodi che rappresentano gli autori sono in blu. Sono stati inoltre eliminati i video esterni al grafo per una maggiore leggibilità.

Come si può vedere dalla Figura 1 molti utenti commentano un solo video creando così community differenti. Per cercare queste community sono state utilizzate due tecniche: l'algoritmo Girvan-Neumann e un approccio greedy basato sulla modularità.
L'algoritmo di Girvan-Neumann elimina iterativamente gli archi caratterizzati dal numero più elevato di sentieri più corti tra i nodi che passano attraverso loro. Rimuovendo tali archi uno alla volta, la rete viene divisa in parti più piccole, dette appunto communities. Come si può vedere in Figura 3 si ottengono sei community, dove 4 si riferisocno ai video meno commentati, mentre non trova community "all'interno" del grafo.
L'approccio greedy, invece, porta a risultati più interessanti di prima, in quanto l'algoritmo individua 29 diverse community all'interno della rete, di

Video	Grado
Become a DATA ANALYST with NO degree?!? The Google Data Analytics Professional Certificate	996
Google vs IBM Data Analyst Certificate - BEST Certificate for Data Analysts	376
Get a JOB w/ Google Data Analytics Certificate?!? (ft. Certificate Holders)	257
Windows on the M1 Mac - What are your options?	236
STOP using Spreadsheets for Everything!	226
How To Use Tableau Desktop Controls - Tableau Tutorial P.2	2
Calculated Fields in Tableau (Formulas & IF Statements) - Tableau Tutorial P.6	2
Top Non-technical Skills for Business Intelligence	1
What is Business Intelligence (BI)? #shorts	1
Conditional Format Tables in Tableau (Like Excel!) - Tableau Tutorial P.5	1

Tabella 1: Tabella rappresentante i 5 video con grado più alto e i 5 video con grado più basso.

dimensionalità diverse.

È interessante notare come i nodi video appartenenti ad una stessa comunità siano spesso caratterizzati da un filone logico/descrittivo (ad esempio una serie di video riguardanti il linguaggio di programmazione Python). Per questo motivo si sono considerati i video all'interno della stessa community e si è costruito un nuovo grafo dove i nodi sono gli utenti collegati tra loro se hanno commentato lo stesso video; in questo modo gli utenti che hanno commentato uno o più video risultano collegati da un arco caratterizzato da un peso proporzionale al numero di video con cui entrambi hanno interagito. Una volta risposto alle domande preliminari e aver compreso la rete si è deciso di effettuare l'analisi per grafi vera e propria, ponendo come nodi, gli autori dei commenti e come archi i video, e pesando gli archi in caso di più video in comune. Per un motivo computazionale e anche di puro interesse, si è deciso con questa analisi di restringere il campo al topic Python del canale. Più in particolare si sono considerati i video con tema Python [3].

Per eseguire questo grafo è stato generato un nuovo dataset con un video

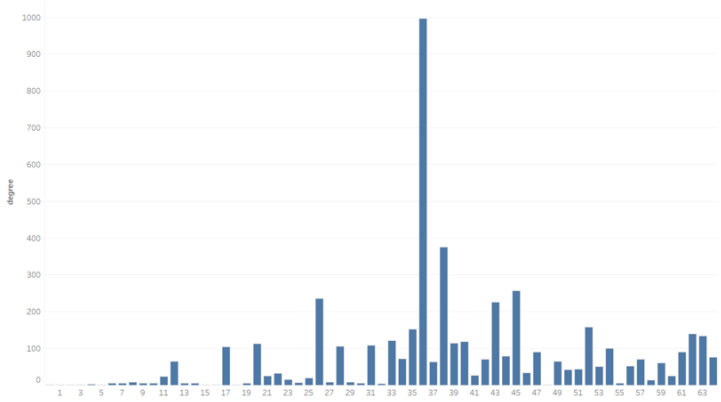

Figura 2: Grado per i nodi video. I video sono ordinati rispetto alla data di pubblicazione.

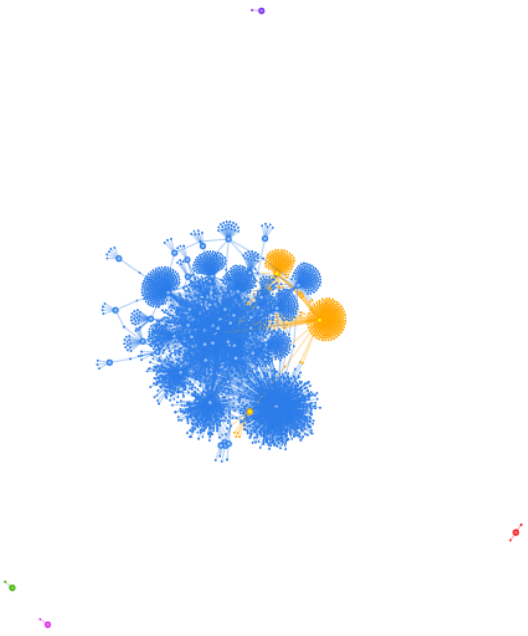

Figura 3: Community individuate dall'algoritmo di Girvan-Neumann.

per ogni riga ed una colonna authors, variabile caratterizzata da una lista contenente tutti gli autori che hanno commentato il relativo video. Nel dataset

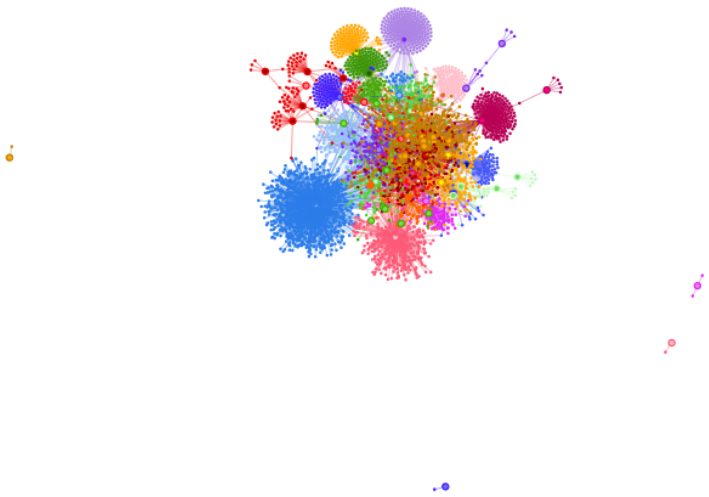

Figura 4: Community individuate con il metodo greedy.

sono presenti solamente i video appartenenti alle playlist sopra menzionate e gli autori che hanno interagito con essi, scrivendovi almeno un commento. Il grafo ricavato da tale dataset e costruito come spiegato in precedenza, risulta caratterizzato da 96 nodi e 702 archi.

Dopo aver fatto ciò, è stata condotta un'analisi di community detection tramite il metodo Girvan-Newman, per vedere se fossero presenti eventuali comunità all'interno della rete. Dallo studio sono state trovate cinque comunità rappresentate in Figura 5, due delle quali sono caratterizzate da una numerosità molto più elevata rispetto alle altre tre. Interessante scoprire come all'interno di un canale youtube e all'interno di una playlist specifica in cui il topic dovrebbe essere solo uno, si formino più comunità. Analizzando i video si scopre come le comunità rappresentino dei sotto argomenti del canale.

La comunità evidenziata in blu è caratterizzata da archi che riguardano una serie di video informativi di Python applicato su vari dispositivi (come, ad esempio, il mac M1); la comunità evidenziata in verde è caratterizzata da video tutorial riguardanti i cicli for, while e le funzioni/gli oggetti principali del linguaggio di programmazione).

Infine, sono stati calcolati i valori di modularità ed assortatività della rete, così da verificare che lo studio condotto non sia dettato da una distribuzione randomica delle comunità e da poter trarre delle conclusioni dal loro

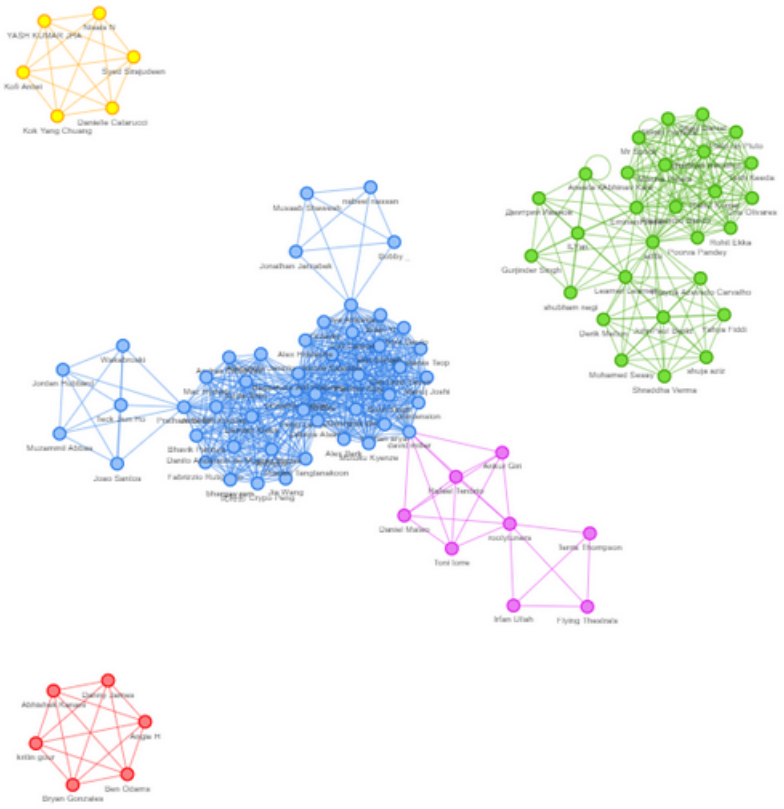

Figura 5: Community individuate con il metodo greedy per il grafo riguardante gli autori che hanno commentato i video a tema Python.

comportamento.

Il valore del coefficiente di modularità della rete è pari a 0.477; questo suggerisce che la configurazione a cluster trovata, si distingue da una possibile composizione randomica della rete. Inoltre, essendo il valore positivo, si evincono dei legami positivi all'interno della rete (ovvero il numero di archi presenti nella rete è maggiore del numero di archi attesi).

Il valore del coefficiente di assortatività della rete è pari a 0.482. Trattandosi di un valore positivo, si può evincere che la rete sia assortativa e, quindi, i gradi dei nodi della rete sono correlati tra loro positivamente. Questo è probabilmente dovuto al fatto che è presente una community fedele al canale (gli iscritti) che partecipano attivamente, commentando spesso i video di nuova uscita. Una rete assortativa per altro è tipica dei social network, e la rete presa in esame non fa eccezioni.

4 Social Content Analysis

Lo scopo della Social Media Analysis è trasformare informazioni in conoscenza, rendere quantificabili aspetti qualitativi che di solito non lo sono. I testi e i commenti dei social media, resi quantificabili, vengono ampiamente utilizzati nella ricerca in tutte le discipline, comprese le scienze politiche, le comunicazioni, il giornalismo e gli affari. L'impatto dei social media sul panorama politico, nonché l'interazione tra i media tradizionali e i social media, sono argomenti di attualità. I dati dei social media possono anche fungere da barometro per monitorare i cambiamenti di atteggiamento nei confronti di questioni degne di nota o controverse.

I risultati della ricerca sui social possono essere sfruttati nel mondo degli affari. Possono essere utilizzati sia come mezzo per comprendere le esigenze dei clienti sia per sviluppare strategie di comunicazione e pubblicità orientate. La ricerca sui social può anche servire a determinare l'esposizione di un'azienda all'interno del mercato ed essere utilizzata per monitorare i concorrenti. Per compiere la ricerca sul canale YouTube di Luke Barousse si sono utilizzati due differenti tipologie di analisi, la Sentiment Analysis e la Emotion Analysis.

4.1 Sentiment Analysis

La sentiment Analysis è una tecnica di elaborazione del linguaggio naturale (NLP) utilizzata per determinare se i dati a disposizione sono positivi, negativi o neutri. L'analisi del sentiment viene spesso eseguita su dati testuali per aiutare le aziende a monitorare il sentimento del marchio e del prodotto nel feedback dei clienti e comprenderne le esigenze. L'obiettivo è verificare l'efficacia comunicativa, capire come gli utenti interagiscono e se sono soddisfatti o meno dei contenuti. L'analisi del sentimento si concentra sulla polarità di un testo per capire se esso risulta positivo, negativo oppure neutro. Per il caso di studio si sono utilizzate due tipologie di sentiment analysis differenti, per avere uno sguardo più completo della situazione. Entrambi gli algoritmi selezionati non necessitano di un preprocessing, sono perfettamente in grado di lavorare su testi, con il vincolo che non siano eccessivamente complessi e lunghi. I commenti ad un canale youtube rispecchiano esattamente questa necessità.

4.1.1 Vader

L'analisi del sentimento Vader è un metodo di analisi del testo efficace, sia che si tratti di un intero documento, di un paragrafo o di una frase. Vader (Valence Aware Dictionary for Sentiment Reasoning) è un modello utilizzato per l'analisi del sentiment testuale sensibile sia alla polarità in termini di positiva/negativa, sia all'intensità delle emozioni. È disponibile nel pacchetto NLTK e può essere applicato direttamente a dati di testo senza preprocessing.

La Sentiment Analysis Vader si basa su un dizionario che unisce le caratteristiche lessicali alle intensità emotive e le trasforma in punteggi di sentiment. Esso può essere ottenuto sommando l'intensità di ogni parola nel testo. Vader è inoltre un algoritmo abbastanza intelligente da riuscire anche ad interpretare il contesto delle parole e a categorizzarle correttamente anche se sembrano di significato opposto (es: non mi piace, viene ritenuta negativa) L'analisi Vader prende una stringa e restituisce un dizionario di punteggi in ciascuna delle quattro categorie:

- negativo: calcola la percentuale che un testo sia negativo da 0 a 1;
- neutro: calcola la percentuale che un testo sia neutro da 0 a 1;
- positivo: calcola la percentuale che un testo sia positivo da 0 a 1;
- composto: calcolo normalizzato dei 3 indici sopra, range compreso tra -1 (negativo) e $+1$ (positivo).

	comment	compound	neg	neu	pos
0	What is the average salary?	0.0000	0.000	1.000	0.000
1	Hi. What do you recommend For stata ana spss? ...	0.4329	0.000	0.840	0.160
2	to talk about my self , i started university w...	0.8791	0.062	0.788	0.150
3	#1 skill is learning how to identify what the ...	0.0772	0.000	0.942	0.058
4	Buying an exorbitantly priced Mac to end up ru...	-0.3883	0.107	0.843	0.050

Figura 6: Esempio dei risultati della Vader analysis applicata ai commenti.

4.1.2 TextBlob analysis

TextBlob è una libreria Python per Natural Language Processing (NLP), che supporta analisi e operazioni complesse su dati testuali. Essa utilizza la libreria NLTK per svolgere i suoi compiti. NLTK offre un facile accesso a molte risorse lessicali e consente principalmente agli utenti di lavorare con categorizzazione, classificazione.

Per gli approcci basati sul lessico, un sentimento è definito dal suo orientamento semantico e dall'intensità di ogni parola nella frase. Ciò richiede un dizionario predefinito che classifichi le parole negative e positive. Dopo aver assegnato punteggi individuali a tutte le parole, il sentiment finale viene calcolato da operazioni di pooling come l'utilizzo di una media di tutti i sentimenti.

TextBlob restituisce la polarità e la soggettività di una frase. La polarità è compresa tra $[-1, 1]$, -1 definisce un sentimento negativo e 1 definisce un sentimento positivo. Le parole di negazione sono comunque valutate e comprese ed invertono la polarità.

	comment	textblob_polarity
1344	"Experiencal work" awesome	1.0
2609	Awesome video. I just started the course	1.0
171	9:07 In my role of procurement..... is this ...	1.0
1553	Superb advice 😎	1.0
1147	Great insight from everyone!	1.0
2506	Best Guide on the internet. Thank You!	1.0

Figura 7: Esempio dei risultati ottenuti con la TextBlob analysis.

4.1.3 Analisi esplorativa

L'analisi esplorativa dei dati viene utilizzata per analizzare le due diverse sentiment e riepilogarne gli andamenti. Serve per verificare che stia operando correttamente seguendo andamenti simili, ma anche per comprendere la sentiment dei commenti. Nella prima esplorazione si sovrappongono l'indice di polarità di TextBlob e l'indice compound effettuato tramite Vader (Figura 8). Essendo entrambi compresi in un range $[-1, +1]$, sono confrontabili e riportabili in un grafico con assi uguali tra loro.

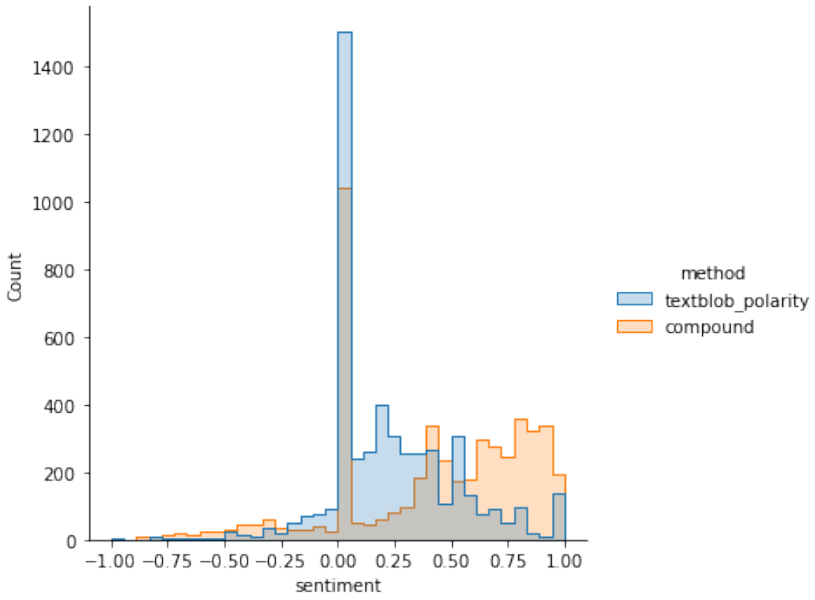

Figura 8: Confronto tra l'indice di polarità di TextBlob e l'indice compound effettuato tramite Vader.

Si nota immediatamente la presenza di un grande accumulo di commenti con sentiment 0.0. Ovviamente si tratta dei commenti con sentiment neutra, ed entrambe le librerie sembrano trovarne in misura simile. Analizzando il canale dal punto di vista della sentiment, è evidente come la maggior parte dei commenti risultano essere positivi, evidentemente i contenuti di Luke Barousse stanno soddisfando gli utenti che interagiscono.

La seconda analisiesplorativa proposta in Figura 9 riguarda l'andamento nel tempo dell'indice compound della libreria Vader e l'indice di polarità della libreria TextBlob.

Esattamente come nell'esplorazione precedente l'andamento positivo della sentiment è evidente per entrambi gli indici, si alternano commenti molto positivi a commenti neutri, con solo qualche eccezione negativa. Le serie storiche mostrano andamenti costanti nel tempo ad eccezione dell'ultimo periodo. Si evidenzia infatti un aumento della variazione, simbolo che i nuovi contenuti proposti nel canale polarizzino di più gli utenti.

Nell'ultima analisi esplorativa proposta in Figura 10 si vuole rappresentare in ordine per l'indice compound della libreria Vader i video che hanno ottenuto il miglior sentiment. Si pensa che questa esplorazione possa portare

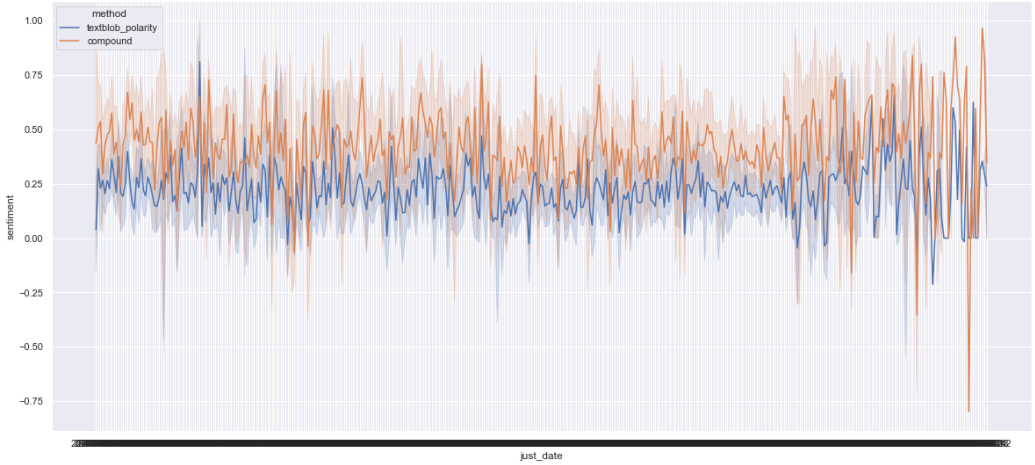

Figura 9: Andamento nel tempo dell'indice compound della libreria Vader e l'indice di polarità della libreria TextBlob.

il proprietario del canale a meglio comprendere i propri contenuti video e capire quali hanno avuto più successo e quali invece sono stati ritenuti scadenti dalle community del suo canale.

4.2 Emoticon analysis

In alcuni casi l'analisi del sentiment potrebbe non comprendere a sufficienza cosa prova effettivamente il cliente. L'Emotion Analysis è il processo di identificazione e analisi delle emozioni celate e nascoste espresse nei dati testuali. Sia l'Emotion che la Sentiment vengono utilizzate per scoprire e quantificare il modo in cui le persone rispondono ai nuovi prodotti, che si tratti di misurare la risposta a una nuova serie TV o evento o nel caso specifico, video di youtube, Emotion e Sentiment Analysis possono aiutare i creatori di contenuti a rispondere con offerte su misura. Può sembrare che Emotion e Sentiment siano la stessa cosa, ma le due tecniche presentano alcune differenze importanti. Entrambe hanno il desiderio di comprendere meglio le esigenze del consumatore, ma ci riescono in modi differenti. L'analisi emotiva utilizza un sistema complesso per comprendere le risposte dei consumatori, analizza in modo approfondito una gamma completa di emozioni e sensibilità umane. Le statistiche sul sentimento invece monitorano indici positivi o negativi semplificati.

vid_title	compound
Dimensions Vs Measures (Blue Vs Green Data) - Tableau Tutorial P.3	0.842350
Python Functions for Data Science / Data Analysis - P.5	0.838717
M1 vs Intel Mac for Python #shorts	0.830200
Create Stacked Bar Chart (and any other visuals EASILY!) w/ Show Me! - Tableau Tutorial P.4	0.808750
Parameters (Create & Use in Calculated Fields and/or Visuals) - Tableau Tutorial P.7	0.768333
...	...
Install VS Code with Python for Data Science / Data Analysis - P.3	0.331717
Python for M1 Mac vs Intel (SPOILER: M1 is 2x faster)	0.323996
Become a DATA ANALYST with NO degree?!? The Google Data Analytics Professional Certificate	0.308248
Install Python for Data Science on Mac & Windows (PC) with Anaconda - P.1	0.139120
M1 Chip is as FAST as M1 Max!!! 😱 (13" Mac Air Vs. 14" Mac Pro) #shorts	0.103780

Figura 10: Elenco dei commenti ordinati secondo l'indice compound.

L'Emotion Analysis misura le differenze nei sentimenti che esprimono vari spettatori o acquirenti, utilizzando emoji e analisi del testo. Questo è uno sguardo approfondito alle emozioni e alle intensità vissute mentre vengono provate.

L' analisi emotiva esamina anche le motivazioni e gli impulsi di uno spettatore, acquirente o lettore e si traducono facilmente in azioni.

Anche scoprire una risposta confusa può rivelare che i tuoi contenuti sono troppo complicati e devi fornirne di più chiari. D'altra parte, se la sensazione dominante è la noia, dovrai ravvivare le cose con umorismo o contenuti creativi. Si sono quindi analizzate le emozioni per ogni singolo commento e le si sono salvate per aumentare le informazioni a disposizione.

	emotion	comment
0	{'anticipation': 1, 'joy': 1, 'positive': 1, '...	What is the average salary?
1	{'positive': 3, 'trust': 2, 'anticipation': 1}	Hi. What do you recommend For stata ana spss? ...
2	{'positive': 14, 'anticipation': 8, 'trust': 1...	to talk about my self , i started university w...
3	{'positive': 2, 'anticipation': 1}	#1 skill is learning how to identify what the ...
4	{'positive': 7, 'joy': 4, 'trust': 5, 'anticip...	Buying an exorbitantly priced Mac to end up ru...

Figura 11: Emotion dei primi 5 commenti.

5 Conclusioni

Purtroppo non è stato possibile effettuare l'analisi attraverso la teoria dei grafi su tutto il canale per motivi computazionali. Gli esiti dell'analisi sul topic Python, hanno evidenziato la presenza di 5 community, le quali rappresentano degli aspetti diversi trattati da Luke Barousse riguardo python. Il valore del coefficiente di modularità della rete 'e pari a 0.477; dunque la configurazione a cluster trovata, si distingue da una possibile composizione randomica della rete. Il valore del coefficiente di assortatività della rete è pari a 0.482. La rete risulta quindi assortativa e perciò i gradi dei nodi della rete sono correlati tra loro positivamente. Per quanto riguarda la Sentiment Analysis, è evidente come il canale di Luke Barousse, produca contenuti apprezzati, tutti i video hanno una sentiment superiore allo 0. Bisognerebbe analizzare nello specifico i video con migliori risultati e comprenderne il perché del successo così da riprodurli. Un campanello d'allarme sorge nell'ultimo periodo, dove i contenuti stanno diventando molto polarizzanti, ci sono alternanza di commenti molto positivi e la presenza di qualche commento negativo. Nel complesso si può giudicare il canale Luke Barousse, un canale apprezzato e che continuerà a crescere nel tempo.

Riferimenti bibliografici

[1] Canale youtube luke barousse. [Online]. Available: https://www.youtube.com/c/LukeBarousse

[2] Api youtube. [Online]. Available: https://developers.google.com/youtube/v3

[3] Video tema python. [Online]. Available: https://www.youtube.com/c/LukeBarousse/playlists