UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

SOCIAL MEDIA ANALYTICS

Effetto Reddit

Analisi delle differenze strutturali e comportamentali tra subreddit

Teresa Cigna – 813925 - t.cigna@campus.unimib.it

Chiara di Domenico – 815463 – c.didomenico@campus.unimib.it

Federico De Servi – 812166 – f.deservi1@campus.unimib.it

Gennaio 2021

Contenuti

Sommario

Oggigiorno i social media sono diventati strumenti centrali nella diffusione di informazioni a livello globale. In questo progetto viene considerato un social in particolare: Reddit, comunemente noto come la prima pagina di internet. Reddit è un sito Internet di social news, intrattenimento e forum, dove ciascun iscritto (redditor) può creare contenuti testuali o ipertestuali in una serie di subreddit specifici per ciascun argomento. Il progetto si pone l'obiettivo di analizzare le differenze tra le varie comunità prese in esame *(r/Coronavirus, rr/Science, r/Television, r/Gaming, r/Politics)*, sfruttando tecniche quali la *Sentiment Analysis* e studiando il modo in cui l'informazione si diffonda per ciascuno dei macro argomenti scelti, analizzando delle caratteristiche dei vari subreddit e sfruttando tecniche quali la *Community Detection con Infomap*.

1 Introduzione

Internet è un vasto spazio virtuale dove informazioni, idee e opinioni si diffondono rapidamente ed efficacemente. Nell'ultimo decennio, in particolare negli Stati Uniti, Reddit è diventato la prima pagina di internet; il sito web dove informarsi e dove confrontarsi con altri utenti. Il suo punto di forza riguarda proprio l'essere l'aggregatore di così tanti e diversi topic, chiamati "subreddit", da rendere quasi impossibile non trovarne uno di proprio interesse.

Fondato da Steve Huffman e Alexis Ohanian nel 2005, è cresciuto rapidamente nel decennio successivo, portandolo nell'Ottobre del 2020 ad essere il diciassettesimo sito web più visitato al mondo e il settimo negli stati uniti. La sua base di utenti iscritti risulta essere proveniente da tutto il mondo ma, al momento della stesura di questo paper, gran parte di essa è composta da utenti degli Stati Uniti, i quali rappresentano la maggioranza.

Chiaramente, risulta evidente a chiunque ne faccia un uso quantomeno sporadico, come ciascun subreddit abbia caratteristiche uniche, e di come gli utenti e le conversazioni in esso generate abbiano caratteristiche differenti. Basta entrare nei subreddit riguardanti politica e scienza per notare immediatamente differenze abissali nel numero di commenti per post (assai maggiori nel primo che nel secondo) e nel modo in cui gli utenti parlino tra di loro.

Lo scopo di questo progetto è quindi quello di indagare se effettivamente tali differenze tra subreddit risultino effettivamente riscontrate o se siano meramente frutto dei bias cognitivi degli utenti che, come noi, navigano il sito alla ricerca di news e spinti dalla curiosità di sapere l'opinione del resto degli utenti che fanno parte delle rispettive comunità di riferimento.

2 Caso di studio

Il progetto prevede l'analisi delle reti sociali generate dai commenti di migliaia di redditors appartenenti a 5 diversi topic (*r/Coronavirus, rr/Science, r/Television, r/Gaming, r/Politics*). Come anticipato nel paragrafo precedente, Reddit è un sito Internet di social news, intrattenimento, e forum, dove gli utenti registrati (redditors) possono pubblicare contenuti sotto forma di post. Gli utenti, inoltre, hanno la possibilità di commentare ciascun post, di rispondere a commenti di altri utenti e infine di attribuire un voto "su" o "giù" (comunemente chiamati "upvote" e "downvote") a ciascun commento. Tali valutazioni determinano, posizione e visibilità dei vari post sulle pagine del sito. I contenuti del sito sono organizzati in aree di interesse chiamate subreddit.

L'analisi degli utenti Reddit è avvenuta con l'obiettivo di rispondere alle seguenti domande di ricerca:

- Vi sono differenze strutturali tra i subreddit relativi ad argomenti diversi?
- Come circola l'informazione all'interno di diversi subreddit?
- Le interazioni tra utenti sono principalmente positive o negative? Possiamo notare particolari differenze tra i subreddit?
- Come possono essere spiegate tali differenze?

3 Dati

3 .1 Raccolta dati con API Reddit

I dati sono stati ottenuti il giorno 22/12/2020 utilizzando le API rese disponibili da Reddit.

Si è iniziato specificando quali fossero i subreddit da cui attingere i dati. Per ognuno di essi la raccolta è stata effettuata un post per volta, partendo dal post occupante la prima posizione nella pagina e proseguendo verso il basso, finché non si fosse raggiunto un numero di utenti almeno pari al migliaio per subreddit. Ricordiamo che in ognuno di essi i post vengono ordinati dall'algoritmo di Reddit in base a quanto esso sia recente, al numero di commenti ricevuti, e al numero di upvotes e downvotes.

Si è ottenuto così un dataset che presenta i seguenti campi:

- **author**: l'autore del commento
- **body**: il testo del commento
- **created at**: data e ora in cui è stato postato il commento
- **id**: id del commento
- **parent id**: id del commento a cui l'autore sta rispondendo
- **replies**: numero di risposte al commento
- **subreddit**: nome del subreddit di appartenenza
- **subreddit id**: id relativo al subreddit di appartenenza
- **answers_to**: nome dell'utente a cui l'autore del commento sta rispondendo

Una volta ottenute tali informazioni è stato necessario processarle al fine di renderle adatte alla creazione dei relativi network.

3.2 Preprocessing

Per quanto riguarda il preprocessing le uniche operazioni effettuate e richieste dalla natura dei dati collezionati sono state le seguenti:

- Rimozione di tutti quei commenti che presentano valori nulli all'interno del campo "author". La presenza di questi valori è spiegata dall'eventualità che un utente abbia cancellato il suo commento subito dopo averlo postato, causando le API a restituire tale valore.
- Eliminazione delle cosiddette "self connections", ovvero utenti i quali rispondono ad un loro stesso precedente commento. Questo comportamento, per quanto non comune all'interno dei datasets, avrebbe creato problemi durante la creazione dei network.

3.3 Creazione dei network

Per predisporre i dati alla creazione dei network, si è creato un dataset di interazioni tra utenti all'interno di ogni subreddit, raggruppando il precedente per i campi "author" e "answers_to". Si è quindi ottenuto un dataset con i seguenti attributi:

- **author**: autore di uno o più commenti
- **answers_to**: utente a cui author ha risposto tramite uno o più commenti
- **comment_ids**: id dei commenti scritti dall'autore per rispondere all'utente in questione

Ciascuno di questi datasets è stato poi fornito ad iGraph, una libreria per l'analisi e la manipolazione di reti, e tramite essa si sono creati i grafi relativi a ciascuno dei subreddit come grafi diretti.

4 Contenuto dei subreddit - Named Entity Recognition

I subreddit considerati identificano già di per sé il macro argomento trattato nei commenti al loro interno, ma indicando categorie molto ampie si è deciso di esaminare, per ogni subreddit, i titoli e il corpo dei post analizzati in questo progetto in modo da poter essere più consapevoli nell'interpretazione dei risultati.

Nello specifico si è deciso di utilizzare la Named Entity Recognition per poter estrarre le entità nominate. Per questo task si è deciso di utilizzare Dandelion in quanto è risultato lo strumento maggiormente in grado di estrarre i contenuti chiave dei post. Babelfy, infatti, non è stato ritenuto in grado di evidenziare i contenuti chiave in quanto le entità estratte da un breve testo (il titolo) risultavano essere la quasi totalità delle parole presenti, rendendo quindi lo strumento poco utile allo scopo preposto.

In Dandelion è possibile impostare una confidenza minima di entità estratte. Una confidenza alta porta a un minor numero di entità estratte ma con maggior sicurezza che queste ultime siano corrette, una confidenza alta invece porta all'estrazione di un maggior numero di entità ma con una bassa precisione. Si è deciso di impostare il valore della confidenza a 0.7

(default = 0.6) in quanto si è ritenuto essere un buon compromesso tra numero di entità estratte e precisione delle stesse (prediligendo maggiormente quest'ultimo aspetto).

Science
['nazi germany' 'public health']
['climate change' 'pollution' 'the lancet']
['adolescence' 'career' 'conscientiousness' 'contentment' 'education'
'neuroticism' 'positive psychology' 'social status' 'trait theory']
['facebook' 'fox news' 'media consumption' 'twitter']
['mars' 'oxygen']
['hydroxychloroquine' 'interferon' 'lopinavir' 'remdesivir']
['photon' 'quantum computing' 'quantum supremacy']
['osteoarthritis' 'placebo' 'turmeric']
['neuroimaging' 'peer pressure' 'risk aversion']
['behavior' 'happiness' 'health' 'research' 'value (ethics)']

Esempio di output derivante dalla NER 1

Si può dire che gli argomenti trattati dai post utilizzati (e il relativo numero di commenti) sono:

- **Science**: salute in Germania (1890); cambiamento climatico e inquinamento (1723); risk aversion (467); adolescenza, istruzione e carriera (355); ossigeno su Marte (207); fotoni, quanti (138); felicità (135); remdesivir (antivirale) (48); osteoartrosi (18); media (15);
- **Politics**: Donald Trump, Joe Biden, Elezioni, Wisconsin (4331);
- **Television**: Amazon Studios, Pulp Fiction, Lord of the Rings (1165); John Mulaney (1086); Star Wars (445); Criminal Minds (245); Superstore(160); Black Sails (121); Game of Thrones (119); Hawkeye (65); Jason Sudeikis (Ted Lasso) (23); Cinderella (19); Battlebots (13);
- **Coronavirus**: parallelismo morti Coronavirus e 11/9/2001 (4263); Coronavirus (622); vaccino (144); texas; virologia, zoonosi (127);
- **Gaming**: Game of the Year (458); Percezione dei videogiochi tra genitori e figli (377); Final Fantasy VII (368); Minecraft (233); Dark Souls (52); meme (7);

Mentre per i primi 4 subreddit la NER è stata utile a identificare gran parte degli argomenti trattati dai post, per gaming non è stata altrettanto efficace, producendo solo 3 risultati su 9 post identificando unicamente: meme; Final Fantasy VII; Minecraft.

Indagando si è notato che la maggior parte dei post in questione erano poco argomentati a livello testuale, in quanto si riferivano ad immagini autoesplicative. In questi casi quindi si è reso necessario scaricare le immagini in questione per capirne il contenuto e integrare gli argomenti identificati dalla NER. Inoltre la NER ha fallito nell'identificare le entità dei post con più commenti all'interno di Gaming e Coronavirus. Si è pertanto ritenuto opportuno ispezionare manualmente i titoli di questi 2 post per poterne identificare gli argomenti trattati.

5 Reti sociali

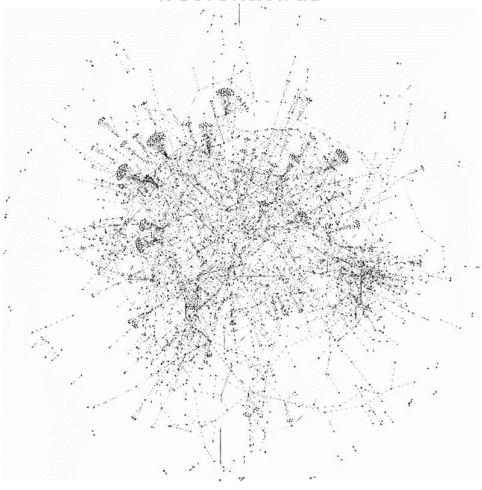
Una rete sociale, nota col termine inglese social network, è una struttura sociale costituita da attori sociali e dall'insieme di interazioni diadiche e relazioni sociali presenti tra di essi. Le interazioni possono essere monodirezionali o bidirezionali. Si pensi a Facebook; una relazione di amicizia in questo caso è bidirezionale, se A è amico di B è vero anche l'inverso. Discorso diverso per altri mezzi, quali Instagram, in cui una relazione di follow non è necessariamente bidirezionale. Se A segue B, non necessariamente è vero l'inverso.

Le reti sociali da noi analizzate in questo lavoro rientrano proprio in quest'ultima categoria. Questo perché il fatto che l'utente A risponda all'utente B con un commento non implica che avvenga il viceversa. Per questo motivo le reti da noi create sono reti dirette, in cui ogni nodo rappresenta un utente, ogni arco rappresenta uno o più commenti creati dall'utente A in risposta all'utente B ed il peso di ogni arco rappresenta il numero di tali commenti di risposta.

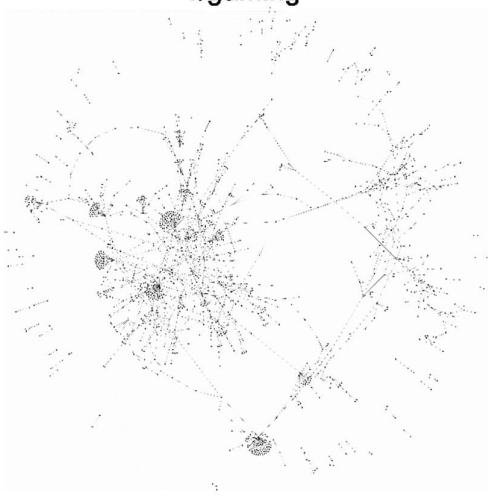
Per visualizzare questi grafi ci siamo avvalsi di Gephi, un software open source per l'analisi e la visualizzazione delle reti sociali. Prima di far ciò è stato però necessario scegliere un layout, ovvero un algoritmo che scegliesse come disporre nodi e archi nello spazio 2d. Si è scelto quindi di adottare il layout "Yifan Hu", per la sua elevata efficienza e per la sua capacità di mantenere chiarezza e di porre enfasi sulla struttura del grafo anche in presenza di medio-alta numerosità di dati.

Si riportano le visualizzazioni risultanti. In ciascuna di esse un punto indica un nodo della rete, ovvero un utente. Vi è un arco direzionato dall'utente A all'utente B se il primo ha risposto in un commento al secondo.

r/coronavirus



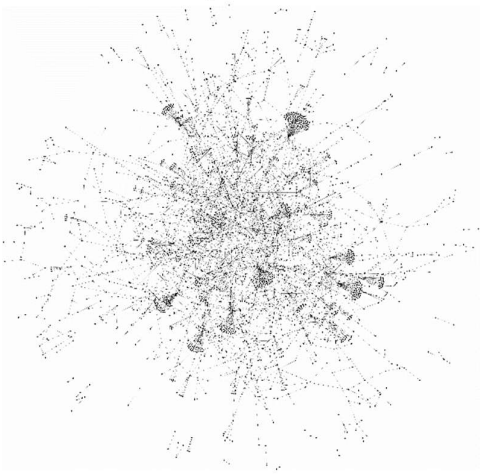
r/gaming



r/science

r/television

r/politics



5.1 Analisi delle reti

5.1.1. Hubs

Definiamo innanzitutto cosa si intende per grado di un nodo, in-degree e out-degree. Il grado di un nodo rappresenta il numero di connessioni incidenti allo stesso. Queste connessioni possono essere poi suddivise, nel caso di una rete sociale monodirezionale, in connessioni uscenti (per cui si calcola l'out-degree) e connessioni entranti (per il calcolo dell'in-degree). Un hub è un nodo il quale presenta un elevato out-degree ed il quale è connesso a nodi che presentano un basso grado. Si è scelto di selezionare soltanto i nodi con hub score maggiore di 0.5.

	Gaming	Coronavirus	Science	Television	Politics
# Hub score > 0.5	1	3	55	1	94

Come si vede dalla tabella, il numero di hubs in *r/Politics* e *r/Science* è molto numeroso, specialmente se confrontato con il numero di hubs presenti negli altri subreddit. Essi sono utenti che rispondono ad un alto numero di altri utenti, i quali hanno invece poche interazioni con altri. Questo può indicare come *r/Politics* e *r/Science* siano particolarmente interconnessi al loro interno.

5.1.2 Authorities

Le Authorities rappresentano il concetto opposto degli Hubs precedentemente definito, ovvero nodi i quali presentano un elevato in-degree ed i quali sono connessi a nodi che presentano un basso grado. Calcolato l'Authority score per ciascun nodo presente nei networks, si sono selezionati soltanto coloro che presentano un valore relativamente alto (maggiore di 0.5) e si è ottenuto che in tutti e 5 i subreddit, un solo nodo soddisfa tale requisito.

	Gaming	Coronavirus	Science	Television	Politics
# Authority score > 0.5	1	1	1	1	1

Questi nodi rappresentano utenti che generano commenti che creano spunti di conversazione particolarmente sentiti dagli altri utenti, e che quindi presentano un elevato numero di risposte.

5.1.1 Assortativity

Il coefficiente di assortativity è un coefficiente che assume valori positivi quando i nodi nella rete hanno la tendenza a connettersi maggiormente con nodi a loro simili, mentre assume valore negativo quando accade l'opposto.

	Gaming	Coronavirus	Science	Television	Politics
Assortativity	-0,16	-0,19	-0,17	-0,14	-0,16

Come si vede dalla tabella, nei network i nodi non presentano la tendenza a connettersi con nodi simili. Vi è anzi una lieve tendenza al respingimento tra nodi simili. Nel caso specifico degli hubs, questo comportamento potrebbe essere spiegato dal fatto che trattandosi di utenti che rispondono ad una grande quantità di commenti differenti, raramente essi danno vita a nuovi spunti di conversazione, e conseguentemente interagiscono in minor misura tra loro.

5.1.1. Reciprocity

La reciprocity calcola la proporzione di mutual connections, ovvero la percentuale di connessioni bidirezionali all'interno del grafo. In tutti e 5 i casi si può notare come il valore sia basso, confermando la presenza di tale comportamento all'interno di tutti i network in esame e confermando la correttezza della scelta di creare tali reti come reti monodirezionali.

	Gaming	Coronavirus	Science	Television	Politics
Reciprocity	0,32	0,29	0,33	0,25	0,22

6 Community detection

6.1 L'algoritmo Infomap

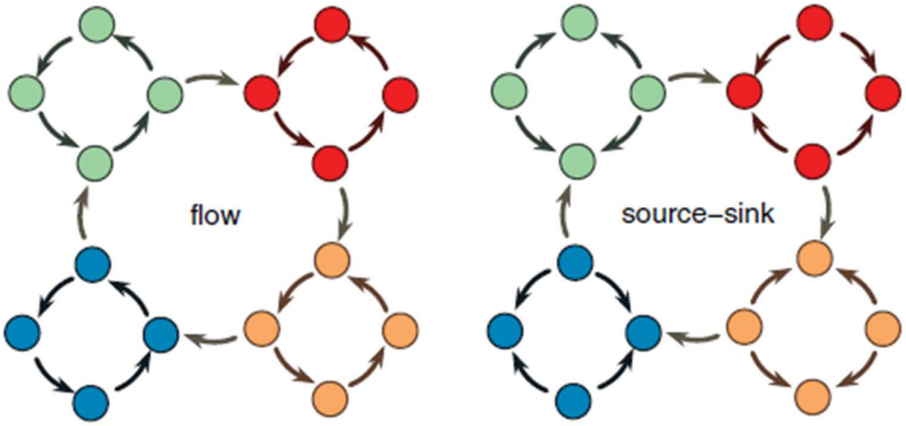
Per la community detection è stato utilizzato l'algoritmo *Infomap* in quanto utile a capire il modo in cui circola l'informazione all'interno della rete.

Infomap è un metodo introdotto da Rosvall e Bergstrom (2008). L'algoritmo trova le partizioni che minimizzano la cosidetta Map-equation (si veda successivamente la descrizione della stessa). Si tratta di un algoritmo a 2 livelli in cui i nodi vicini vengono uniti. Inizialmente ogni nodo è assegnato alla sua partizione, successivamente, in un ordine casuale, ogni nodo viene spostato nella partizione più vicina che porta ad una maggior diminuzione della map equation. Se nessun movimento porta alla diminuzione della funzione obiettivo, il modulo rimane nella sua partizione originale. Questa operazione viene ripetuta partendo da un ordine casuale differente finché non viene prodotta più alcuna mossa. A questo punto la rete è ricostruita in base a quest'ultima configurazione e viene ripetuto il procedimento precedente, ma ora i nodi appartenenti alla stessa partizione sono costretti a muoversi congiuntamente. Si procede finché non avviene più alcuna riduzione.

La Map equation sfrutta la dualità tra trovare strutture a cluster all'interno dei network e minimizzare la lunghezza del random walk all'interno delle stesse. Il random walk è un oggetto matematico, il quale descrive un percorso che consiste in una successione di step randomici in un determinato spazio. In questo caso, esso si muove casualmente tra oggetti all'interno del network. Più una connessione ha peso alto, più è probabile che il random walker usi quella connessione per arrivare all'oggetto. L'obiettivo è formare cluster in cui il random walker stia il più tempo possibile (i.e. i pesi delle connessioni interne al cluster devono essere maggiori di quelle tra oggetti appartenenti a cluster differenti).

$$L(M) = w \curvearrowright log(w \curvearrowright) - 2\sum_{k=1}^{K} w_k \curvearrowright log(w_i \curvearrowright) - \sum_{i=1}^{N} w_i log(w_i) + \sum_{k=1}^{K} (w_k \curvearrowright + w_i) log(w_k \curvearrowright + w_k)$$

Questa definizione è basata sul concetto di entropia o informazione media del contenuto o densità dell'informazione. Dalle seguenti immagini è inoltre possibile capire venga valutata la map equation in reti apparentemente simili:



$L = 3.33$ $L = 4.58$

La differenza tra le 2 immagini consiste nella direzione degli archi. E' possibile notare come nella prima rete la map equation (L) abbia una valore minore, e quindi migliore, rispetto alla seconda.

6.2 Risultati

I risultati ottenuti sono i seguenti. Riportiamo in tabella per ogni topic le 5 communities più grandi trovate, il relativo numero di utenti compresi e il diametro della stessa.

Gaming - 194		Coronavirus - 256		Science - 217		Television - 306		Politics - 286	
# Users	Diameter	# Users	Diameter	# Users	Diameter	# Users	Diameter	# Users	Diameter
81	4	117	7	49	3	68	4	127	3
73	4	65	8	46	8	60	6	97	7
66	7	56	3	43	4	46	5	86	5
44	5	50	6	41	3	45	6	76	5
31	6	45	2	41	5	45	3	73	9

Si può notare come *r/Politics* e Coronavirus siano i network contenenti le communities più grandi, le quali, nonostante la loro elevata dimensionalità, presentano comunque un diametro particolarmente basso. Questo indica ulteriormente che tali communities siano particolarmente interconnesse al loro interno. Si può quindi affermare che all'interno dei subreddit *r/Politics* e r/Coronavirus l'informazione circoli più facilmente e più efficacemente.

7 Analisi dei commenti - Sentiment Analysis

Per l'analisi testuale ci si è concentrati sulla sentiment espressa nei commenti, al fine di individuare le differenze di mood presenti tra i 5 subreddit considerati.

7.1 Tecniche di Sentiment Analysis

Si dà, prima di tutto, una breve introduzione sulle tecniche di sentiment utilizzate, in quanto necessaria per comprendere al meglio le scelte che si sono fatte durante la fase di preprocessing:

- **Afinn:** si basa su una lista di più di 3300 parole che assumono polarità tra -5 e 5 che viene assegnata in relazione al fatto che una parola sia considerata negativa, positiva o neutra.
- **NRCLex**: è un lessico che restituisce un dizionario contenente dieci chiavi, ove ognuna di queste chiavi rappresenta un'emozione (es. positive, negative, joy, fear, ecc.) e il valore associato è la somma delle parole che figurano per quell'emozione.
- **VADER (Valence Aware Dictionary and Sentiment Reasoner)**: è un lessico e uno strumento di sentiment analysis basato specificatamente sui sentimenti espressi nei social media e che restituisce un punteggio tra -1 e 1.
- **Bing-Liu**: questo lessico contiene una lista di circa 6800 parole inglesi sia positive che negative. Per poter utilizzare questo lessico si è creata una funzione che sommasse il sentiment score delle varie parole all'interno di una frase e, in caso di frasi precedute da una negazione, esso considera il sentiment score inverso fino al successivo segno di punteggiatura.
- **TextBlob**: è una libreria python open source. Questo algoritmo contiene due implementazioni di sentiment analysis, PatternAnalyzer (basato sulla libreria pattern) e NaiveBayesAnalyzer (un classificatore NLTK addestrato su un corpus di recensioni di film). L'implementazione utilizzata è stata PatternAnalyzer che si basa sugli aggettivi positivi e negativi più comunemente utilizzati, assegnando al testo un valore tra -1 e 1.
- **IBM (Watson Natural Languages Understanding)**: è un tool di intelligenza artificiale avanzata che si basa sulle ultime innovazioni del machine learning. Analizza il testo per estrarre i metadati dal contenuto, quali concetti, entità, parole chiave, categorie, sentimenti, emozioni, relazioni e ruoli semantici utilizzando la comprensione del linguaggio naturale. Per lo scopo del progetto, si è estratta solo la polarità della sentiment, ovvero un punteggio tra -1 e 1.

7.2 Preprocessing

Nonostante tutte le tecniche presentate non necessitino di un particolare preprocessing, si è deciso di apportare alcune modifiche ai testi in modo che i vari tool potessero performare al meglio.

A questo scopo si sono scelte diverse tipologie di preprocessing che si potessero adattare a ciascun analyzer.

7.2.1. Sostituzione termini ingannevoli

Come prima cosa, per il subreddit 'gaming' si è reso necessario sostituire alcuni nomi di giochi al cui interno apparivano termini negativi (*e.g. 'No Man's Sky', 'Undead Nightmares', 'Red Dead Redemption', 'Mortal Kombat'*) in quanto falsavano il risultato della sentiment.

A questo scopo, tramite la NER, si sono estratte le entità relative ai commenti del suddetto subreddit e si sono identificati i titoli di giochi che potessero rientrare in questa categoria. Si è poi arricchita questa lista con titoli non identificati dalla NER come "No Man's Sky".

Non essendo interessati all'oggetto/soggetto specifico del commento ma all'emozione espressa, si è quindi scelto di sostituire questi titoli con *'this game'*.

7.2.2. Eliminazione link e simboli

Sono stati eliminati link e simboli *(e.g. \n, \r, -, <, >, [,], ecc.)* mantenendo però le emoticons in quanto potenzialmente utili per l'individuazione della sentiment specialmente da parte di VADER.

Es.

"\r So look at this! https://www.fabulousthing.it I think it is fabulous! :D \n\n"

↓

"So look at this! I think it is fabulous! :D"

7.2.3. Lemmatizzazione, rimozione stopwords e gestione delle negazioni

Preprocessing per Afinn, NRC, Bing-Liu, e TextBlob

In questo caso oltre alle operazioni precedenti, si è optato per:

- Lemmatizzazione: consiste nella riduzione delle parole alla loro forma base. Si rivela utile per le tecniche strettamente basate sul lessico in quanto permettono di avere una corrispondenza tra le parole all'interno del testo e le parole contenute nel lessico di riferimento.
- Rimozione di stopwords inglesi: consiste nell'eliminazione di tutte quelle parole non ritenute significative *(i.e. articoli, preposizioni ecc.)*. In questo caso si è scelto però di mantenere le negazioni, ritenute importanti specialmente per la tecnica che fa uso del lessico di Bing-Liu come spiegato in precedenza; togliendo la negazione infatti, una frase come *'this is not good'* avrebbe il significato opposto.

Es.

“He studied a lot to act in this film, but he’s not good”

↓

“Study act film, not good”

7.2.4. Tokenizzazione

Per poter eseguire tutte le operazioni precedenti è stato necessario effettuare una tokenizzazione iniziale del testo; tuttavia tutte le sentiment, tranne quella di Bing Liu, richiedono in input una stringa, non un lista di token, pertanto, al termine delle operazioni precedentemente descritte, la lista di token pre-processati è stata ritrasformata in stringa.

Es.

“So i think it is fabulous!”

↓

“[think, fabulous, !]”

↓

“think fabulous!“

7.2.5. Punteggiatura

Si è scelto di non eliminare la punteggiatura in quanto ritenuta utile sia per VADER che per gestire correttamente le negazioni definendo il punto oltre il quale la sentiment delle parole deve tornare ad essere considerata normalmente (i.e. non in maniera inversa).

7.3 Analisi

Per ogni tecnica di sentiment analysis utilizzata, si descrive di seguito il testo su cui si è applicata la tecnica e le funzioni implementate per avere uno score normalizzato quindi con un range compreso nell’intervallo (-1 ; 1) che permetta di confrontare tra di loro le varie tecniche di sentiment al fine di scegliere l’analyzer che performa meglio in termini di F1_score.

Vader e IBM sono state applicate ai commenti a cui è stato applicato il preprocessing semplice.

Invece, Afinn, NRCLex, Bing-Liu e TextBlob, oltre ad essere applicate ai testi con preporocessing semplice, sono stati applicati anche a commenti su cui è stato applicato il preprocessing preposto.

Per normalizzare gli score si sono implementate due funzioni.

Si è creata una funzione per cui, passando come parametro la lista degli score, restituisse:

- -1 ovvero negativo se il valore è minore di -0.1;
- 0 ovvero neutro se il valore è compreso tra -0.1 e 0.1;
- 1 ovvero positivo se il valore è maggiore di 0.1.

Questa funzione è stata applicata a tutte le tecniche di sentiment tranne NRCLex, per la quale si è creata una funzione a cui passando come parametro iniziale il dizionario degli score, essa divide internamente le emozioni tra positive e negative e somma i relativi valori.

Per ogni commento si sono confrontati i valori delle emozioni positive con i valori delle emozioni negative, al fine di restituire:

- -1 ovvero negativo se il valore delle emozioni negative è minore del valore delle emozioni positive;
- 0 ovvero neutro se il valore delle emozioni negative è uguale al valore delle emozioni positive;
- 1 ovvero positivo se il valore delle emozioni positive è maggiore del valore delle emozioni negative.

7.4 Valutazione

Al fine di scegliere la tecnica di sentiment migliore, si è effettuato un controllo manuale in doppio cieco di 60 commenti per ogni subreddit considerato, assegnando ad ogni commento uno score pari a -1 per i commenti negativi, 0 per i commenti neutri e 1 per i commenti positivi. Prima si sono confrontati gli score assegnati dai due componenti che sono risultati per lo più simili tra loro e per tale motivo si sono ridotte da due valutazioni differenti ad un'unica valutazione. Dopo si è calcolato l'F1_score, utilizzando come GroundTruth la valutazione data dai componenti del gruppo, per capire effettivamente quale tecnica di sentiment classifichi meglio.

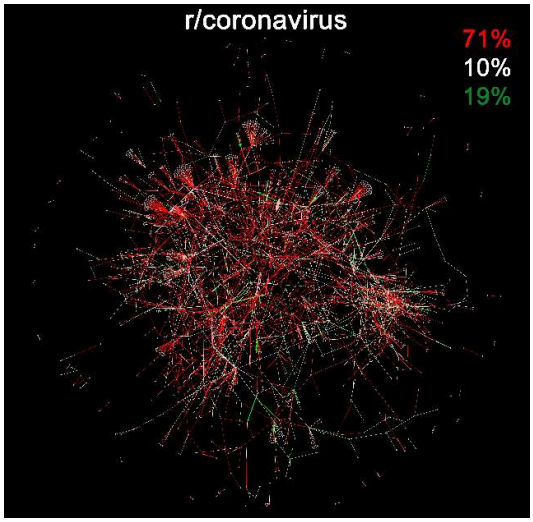
Per tutti e cinque i subreddit si trova al primo posto la tecnica di IBM con un F1_score maggiore di 0. Si è poi notato che la tecnica Afinn raggiunge in tre dei cinque subreddit un F1_score migliore di tutte le altre tecniche restanti e negli altri due subreddit lo score di Afinn non si discosta tanto dallo score delle tecniche che lo precedono.

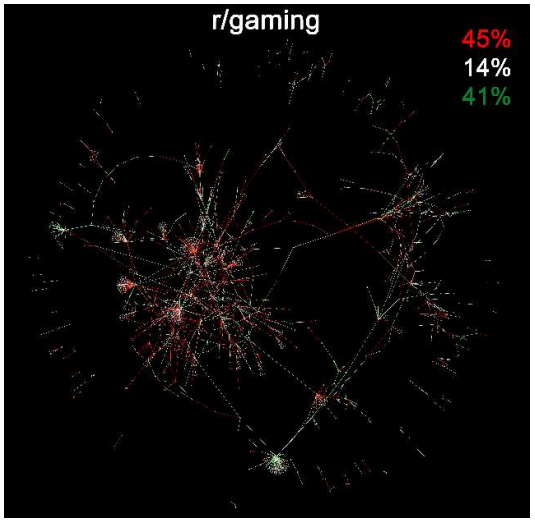
Per questo motivo, si è deciso di considerare entrambe le tecniche di sentiment (IBM e Afinn) ma si è dato loro un peso diverso, ovvero 0.75 per IBM e 0.25 per Afinn. Questi pesi sono stati dati in quanto al momento della valutazione ci si è resi conto che, nelle poche eventualità in cui IBM avesse classificato in modo errato, Afinn lo avesse fatto correttamente correttamente.

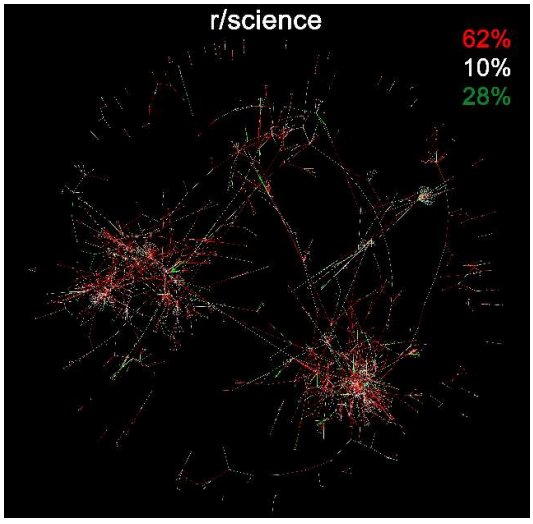
Quindi, si è optato per l'utilizzo di questi pesi in modo che IBM prevalga su tutti, essendo effettivamente la tecnica che raggiunge in ogni subreddit l'F1_score maggiore, ma laddove IBM sbagli, lo score di IBM possa essere mitigato dallo score di Afinn.

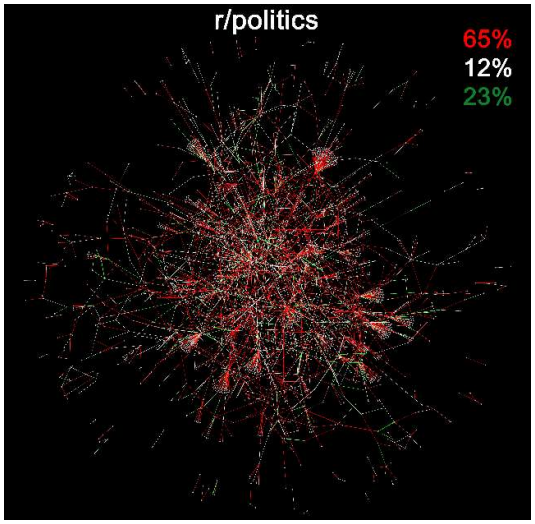
7.5 Risultati

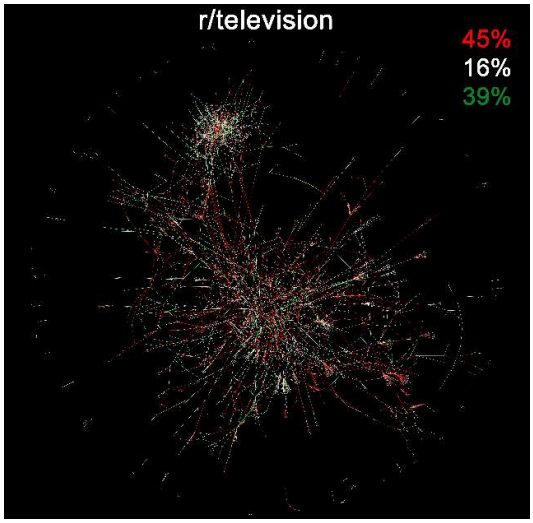
Riportiamo alcuni grafici rappresentanti il network con nodi e archi, i quali sono stati colorati di diverse gradazioni di rosso, verde o bianco in base al risultato ottenuto dalla sentiment analysis. Inoltre, si mostra la percentuale di commenti positivi, negativi e neutri in alto a destra di ognuno di essi.











Come si può vedere dalle immagini, la sentiment all'interno dei subreddit *r/Coronavirus, r/Science, r/Politics* è per lo più negativa. Invece, non risulta esserci una sentiment preponderante all'interno di *r/Gaming* e *r/Television*.

8 Conclusioni e futuri sviluppi

Il progetto si è incentrato sulla ricerca e sull'identificazione di differenze tra i subreddit considerati, che ricordiamo essere stati (*r/Coronavirus, rr/Science, r/Television, r/Gaming, r/Politics*).

Per quanto riguarda il punto di vista strutturale, si può affermare che le principali e sostanziali differenze riguardano il numero di hub. In particolar modo, si è notato chiaramente come *r/Politics* e science risultino avere un numero di hubs particolarmente elevato rispetto agli altri subreddit. Ciò significa che contengono molti più utenti che rispondono ad un alto numero di altri utenti, i quali hanno invece poche interazioni con altri. Questo può indicare come i subreddit *r/Science* e *r/Politics* siano particolarmente interconnessi al loro interno.

Per quanto riguarda invece la circolazione dell'informazione al loro interno, si è sfruttato l'algoritmo infomap, il quale ha fornito risultati interessanti. Si è notato come *r/Politics* e Coronavirus fossero i network contenenti le communities più grandi, le quali presentavano comunque un diametro particolarmente basso. Questo indica ulteriormente che tali communities sono molto interconnesse al loro interno, nonostante l'elevata dimensionalità. Possiamo quindi affermare che all'interno dei subreddit *r/Politics* e r/Coronavirus l'informazione circoli più facilmente e più efficacemente.

Come abbiamo già visto precedentemente, dal punto di vista del contenuto delle interazioni tra gli utenti e del loro comportamento, *r/Politics, r/Coronavirus* e *r/Science* risultano caratterizzate da una sentiment prevalentemente negativa, cosa che non accade in *r/Gaming* e *r/Television*, i quali non risultano avere una sentiment preponderante.

Pare che quindi, argomenti che possiamo definire più frivoli e leggeri abbiano una sentiment relativamente neutra, mentre argomenti che più suscitano emozioni forti e coinvolgimento psicologico ed emotivo come *r/Politics* presentino utenti che esprimono emozioni negative.

Questo non stupisce particolarmente, e conferma in gran parte l'esperienza quotidiana che tutti notano nelle discussioni relative a temi caldi come politica e all'ancor più controverso tema del r/*Coronavirus*.

Futuri sviluppi su questo progetto possono sicuramente essere la ricerca di fake news all'interno dei subreddit che hanno presentato un information flow maggiore, al fine di verificare se possa essere spiegato dal fatto che le notizie e le discussioni circolanti in tali subreddit siano in gran parte notizie negative e sensazionaliste, al limite delle fake news (le quali ricordiamo non vengono censurate dal social) e che proprio in quanto tali, esse abbiano un impatto più profondo sugli utenti, i quali inconsciamente o meno, condividono e diffondono tali informazione maggiormente.

Bibliografia

[1] Ludvig Bohlin, Daniel Edler, Andrea Lancichinei, and Martin Rosvall, Community detection and visualization of networks with the map equation framework, https://www.mapequation.org/assets/publications/mapequationtutorial.pdf

[2] M. Rosvall, D. Axelsson, C.T. Bergstrom, The map equation, https://www.mapequation.org/assets/publications/EurPhysJ2010Rosvall.pdf

[3] Niklas Junker, Community Detection with Louvain and Infomap, https://www.statworx.com/de/blog/community-detection-with-louvain-and-infomap

[4] Shaham Farooq, Generating A Twitter Ego-Network & Detecting Communities, https://towardsdatascience.com/generating-twitter-ego-networks-detecting-ego-communities-93897883d255

[5] Wikipedia, Random Walk, https://en.wikipedia.org/wiki/Random_walk

[6] Wikipedia, Reddit, https://en.wikipedia.org/wiki/Reddit

[7] B. Liu, Sentiment Analysis and Opinion Mining, 2012

[8] Himanshu Sentiment Analysis with AFINN Lexicon, https://medium.com/@himanshu_23732/sentiment-analysis-with-afinn-lexicon-930533dfe75b