Sentiment Analisys

Elaborazione di post Twitter tramite Vader

BANCHINI FEDERICO N46003517
AVETA CARMINE N46003385
BENFENATI DOMENICO N46003380
GALASSO ALESSANDRO N39

9 giugno 2019

Sentiment Analisys

Elaborazione di post Twitter tramite Vader

Sommario

Introduzione

Twitter, esattamente come Facebook, è un social network che è fonte di produzione di una grande mole di informazioni. In particolare, quest'ultime non sono tecnicamente "open data", ma è possibile effettuare analisi sul contenuto dei post, partendo da una parola chiave.

Lo scopo della relazione è quello di focalizzare l'attenzione circa tematiche che riguardano l'aspetto dell'analisi che si occupa di stabilire quali sentimenti sono involontariamente inseriti all'interno di un tweet: questa specifica branchia è detta appunto **analisi sentimentale dei tweet,** nota anche con il nome di **opinion mining**.

La sentiment analisys è un campo dell'elaborazione del linguaggio naturale che si occupa di costruire sistemi per l'identificazione ed estrazione di opinioni dal testo; in particolare si è pensato di sfruttare questo tipo di analisi per svolgere delle ricerche circa il gradimento del popolo sul web riguardo il presidente degli Stati Uniti d'America Donald Trump, utilizzando l'omonimo hashtag.

La struttura dell'articolo è suddivisa in tre parti: la prima parte descrive velocemente come recuperare i dati da Twitter, la seconda introduce il Machine Learning di cui la Sentiment Analysis fa parte e, infine, la terza descrive i risultati ottenuti analizzando 1 giorno di Twitter su #trump.

Il primo passo da fare è stato scaricare in modo massivo una consistente quantità di tweet, seguendo le indicazioni fornite sulla pagina developer di Twitter. Il programma realizzato sfrutta le librerie per scaricare real-time tutti i tweet con uno specifico hashtag. In particolare, per questa analisi abbiamo scelto di analizzare l'hashtag **#trump** ed abbiamo usufruito di un database di tweet che fa riferimento al periodo in cui Trump era molto in voga tra i bookmakers del web, ossia tra il giorno 28 Maggio 2017 e il giorno 29 dello stesso mese. La raccolta che comprende un totale di circa **100.000 tweet.**

Raccolta Dati

La base dati reperita dal web, è stata analizzata attraverso un algoritmo di **Machine Learning.** Lo strumento di sentiment utilizzato è il **Vader**.

Vader(*Valence Aware Dictionary and sEntiment Reasoner*) è un tool per l'analisi sentimentale di testi, specializzato nell'identificazione di contenuti quali post o messaggi inviati tramite social media. Il tool è inglobato nella libreria *nltk (Natural Language Tool Kit)* di Python.

Vader riesce a riconoscere frasi che contengono i seguenti punti chiave:

- Tipiche negazioni *("not good"…)*
- Utilizzo convenzionale di punteggiatura *("Good!!")*
- Utilizzo convenzionale di formattazioni testuali per dare enfasi
- Interpretazione di slang, emoticon codifcate e non, acronimi e diciture popolari *("lol", "😊", ":)", ...)*

La Sentiment Analysis funziona un po' come tutti gli algoritmi di Machine Learning relativi alla tipologia "*Supervised Learning*" e si basa sui seguenti step:

- Costruzione di un lessico di base**,** con frasi distinte tra positive e negative, cosa che è possibile implementare tramite Vader.
- Definizione di un algoritmo di machine learning che "impara" le regole necessarie a classificare una frase in positivo o negativo sulla base del lessico appena definito. Anche questo è compito di Vader.
- Download di nuove frasi da elaborare
- Applicazione dell'algoritmo alle nuove frasi che si occuperà, in autonomia, di decidere se la frase è "positiva" oppure "negativa".

Così facendo la ricerca tra i testi può essere effettuata in un documento dinamico, che si va a comporre di tutti i tweet che è possibile raccogliere circa l'argomento indicato fino alla data attuale.

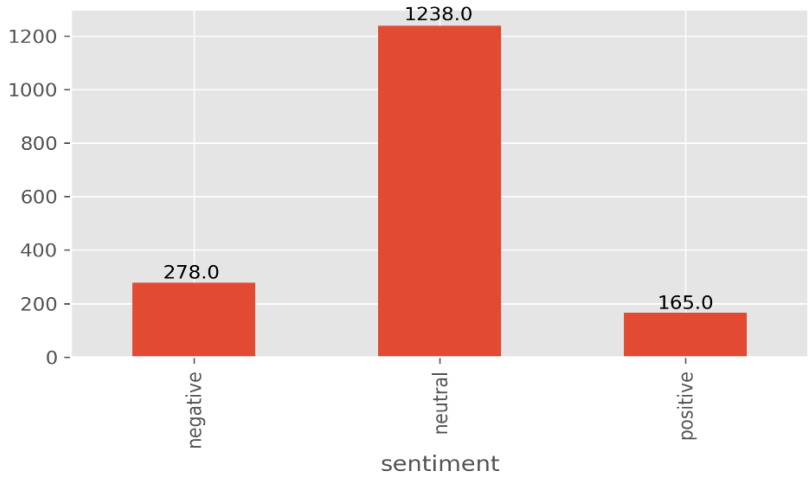
Per realizzare gli ultimi due punti, non possiamo fare affidamento su Vader, ma si è creato uno script ad hoc per l'analisi relativa ai tweet contenenti l'hashtag #**trump** e per l'elaborazione tramite algoritmo di machine learning.

Analisi dei sentiment

La base di dati, come detto, è stata ottenuta tramite un dataset prefabbricato, il quale si compone di tweet che contengono l'hashtag #trump, reperiti a cavallo tra il 28 maggio 2017 e il 29 maggio 2017. Per comprendere la scelta di queste date, basti ricordare che in quei giorni Trump fu indagato per rischio "impeachment" a causa dello scambio illecito di informazioni tra Stati Uniti e Russia, il cosiddetto **Russia Gate**. Inoltre, durante il G7 tenutosi a Taormina lo stesso mese, Trump fu accusato dalla cancelliera tedesca Angela Merkel di essere "inaffidabile", dato che egli si era rifiutato di cedere agli accordi sul clima di cui si è discusso durante la riunione. Come si evince, quindi, quel periodo è stato ricco di commenti sul web riguardanti proprio il presidente stesso, date le mani calde del popolo del web circa il presidente stesso.

Classificazione dei tweet

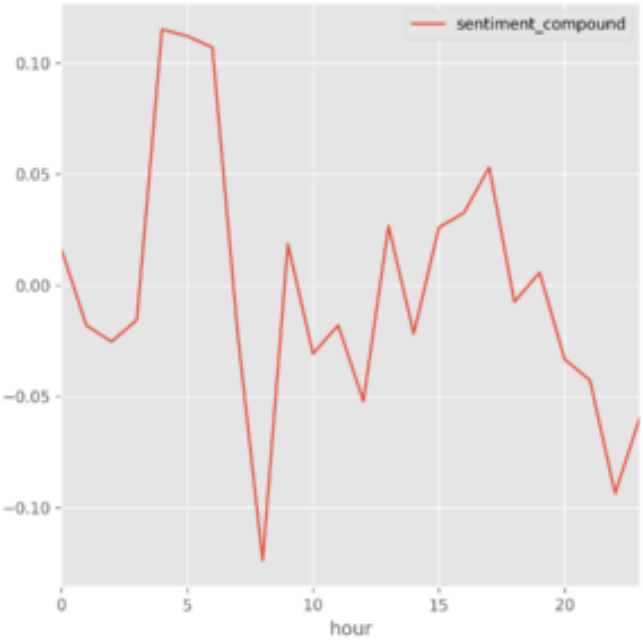
La prima analisi è stata quella di effettuare una distinzione tra tweet positivi e negativi. Ciò che si è ottenuto è descritto nel grafico che segue.



Come mostra il grafico, emerge una notevole quantità di tweet "neutri", per i quali l'algoritmo non ha evidenziato un particolare sentiment positivo o negativo. Tra i contenuti analizzati si evince una poca differenza tra i tweet positivi e quelli negativi. Ciò sta a significare che, nonostante le notizie negative di quel periodo, Trump risulta molto popolare tra gli americani.

Andamento temporale dell'umore

Altra analisi possibile è l'andamento del sentiment durante tutto l'arco temporale di raccolta dei dati, per verificare come l'umore degli utenti varia a seconda del momento della giornata. Purtroppo, non è possibile ottenere un grafico con delle specifiche di orario di dettaglio, dato che magari sarebbe più giusto effettuare la seguente analisi per un intervallo temporale più ampio. Di seguito vengono mostrati i risultati tramite grafico.



Analisi geografica dei tweet

Altre informazioni che possiamo reperire son di carattere più generale, come ad esempio la posizione geografica della provenienza del tweet. In questo caso ciò che è possibile ottenere sono tre tipo di informazioni sulla posizione:

- Le coordinate: è interessante come dai circa 100.000 tweet raccolti siano emersi solo un centinaio con questa informazione, sintomo del fatto che non tutto il bacino di utenza di Twitter utilizza la geolocalizzazione sul proprio dispositivo durante una sessione di Twitter.
- La location: ogni utente può inserire informazioni circa la zona in cui si trova, senza necessariamente reperire la posizione geografica attuale. Emerge dall'analisi che i commenti negativi provengono maggiormente dalla zona di New York, mentre i commenti più critici hanno origine tra il Texas e la California.
- La timezone: grazie al fuso orario del dispositivo di rete su cui viene utilizzato Twitter, è possibili risalire alla zona geografica in cui si trova l'utente. La ricerca ha evidenziato che la quasi totalità di tweet proviene dall'America.

Infine, è interessante ricercare le parole più frequenti che sono state utilizzate nei tweet, dividendo i termini rilevanti per i tweet positivi da quelli negativi: ne è uscito fuori che per i tweet positivi gli utenti prediligono la parola "TrumpTaxPlan", mentre per quelli negativi abbiamo "spied" e "Merkel".

Conclusioni

A livello generale, l'analisi effettuata porta a due riflessioni principali:

- La qualità dell'analisi effettuata, così come gli algoritmi di machine learning, non sono ancora così maturi da permettere un'analisi approfondita del sentimento umano, ma sembra che la strada intrapresa sia quella giusta.
- Analizzando il dato nudo e crudo, si possono notare discrepanze tra l'opinione pubblica generale e ciò che gli utenti divulgano tramite social media; infatti, sebbene in quel periodo non ci siano state notizie positive riguardanti Trump, il numero di tweet in suo favore non era poi così diverso da quello di tweet negativi.

Ciò che si vede nell'analisi è che, preso in esempio un tweet tra tutti quelli contenuti nel file analizzato, esso ricorre più volte. Di seguito si riporta il testo del suddetto tweet: **TRUMP SHARES MY VALUES GOD FAMILY COUNTRY STRONG WORK ETHIC GREAT EXAMPLE 4 OUR KIDS 2 EMULATE.**

È evidente che tale tipo di tweet, ovviamente con sentiment positivo, sembra un **tweet di propaganda** volto ad influenzare l'opinione pubblica, che si ritrova coinvolto in una serie di retweet. È chiaro che ciò condiziona oltre che la nostra analisi, anche il comportamento delle persone, un po' come accade per le fake news di cui siamo spesso vittime. In questo caso, ovviamente, il post non è falso, ma è semplice propaganda politica, che ha effetti sugli utenti dei social che ancora non si comprendono.

Il problema della manipolazione dell'informazione non ha ancora trovato una valida soluzione, ma al momento ciò che si può fare è semplicemente tentare di arginare le manipolazioni, e non farsi influenzare solo da ciò che qualcuno trova interessante scrivere su un social network.