LA RETE VI ANALIZZA… Come funziona l'analisi dei testi sulla rete

Consideriamo le analisi del testo, le cui applicazioni pratiche possono riguardare la reputazione di un partito politico o di un'azienda, la valutazione dei risultati di una campagna di marketing, la comprensione del profilo di chi compra una particolare marca…
Prendiamo per esempio una tecnica chiamata "**Sentiment Analysis**".
Ci si riferisce al termine inglese, visto che gli sviluppi sono portati avanti in paesi anglosassoni.
Si tratta di una tecnica di analisi del testo che rileva un'opinione positiva o negativa all'interno del testo stesso e che viene svolta in automatico dai computer con algoritmi sofisticati, anche di intelligenza artificiale.
Le persone oggi hanno interazioni quotidiane sulla rete attraverso social network, blog, forum e molto altro. Queste interazioni possono riguardare gli ambiti più svariati, dalla politica fino al business, passando per il sociale.

Comprendere le emozioni è essenziale per le aziende poiché i clienti esprimono i loro pensieri e sentimenti in maniera molto aperta. L'analisi automatica dei feedback consente ai marchi di ascoltare attentamente i propri clienti e di personalizzare prodotti e servizi per soddisfare le loro esigenze.
Alcune aziende misurano lo stato emotivo dei loro dipendenti raccogliendo dati sui social network interni. Questa tecnica potrebbe avere applicazioni in campo medico, per esempio identificando le persone affette da depressione.

Nel campo dell'analisi del testo, **la comprensione del linguaggio** è uno dei problemi più complessi per l'intelligenza artificiale.
All'interno di un testo scritto esistono parecchi segnali emotivi che i computer possono riconoscere anche senza capire il significato delle parole.

Gli strumenti più avanzati sono in grado di fornire informazioni tramite il **conteggio delle parole** più utilizzate nei commenti o l'analisi delle emoji ad esse collegate.
Il computer conta il numero di parole positive e sottrae il numero di parole negative.

Un rilevamento migliore può essere ottenuto soppesando le parole, il cui valore è normalmente stabilito da esperti in carne e ossa.
Il conteggio delle parole presenta alcuni problemi, per esempio ne ignora l'ordine.
Per ottenere risultati più precisi ed affidabili si usano gli algoritmi di apprendimento automatico, per insegnare ad un programma come riconoscere i rapporti tra le parole. Queste associazioni possono offrire indizi sul significato o sul sentimento di fondo di una frase.

Nel campo dell'analisi del linguaggio**, si stima che l'80% dei dati nel mondo non sia strutturato** e organizzato in modo predefinito. La maggior parte di questi dati proviene da testi, come e-mail, chat, social media, sondaggi, articoli e documenti. Questi testi sono in genere difficili, lunghi e costosi da analizzare e capire. I sistemi di analisi consentono di dare un senso a questi testi non strutturati, automatizzando i processi e risparmiando ore di elaborazione manuale.

Gli esseri umani **non osservano criteri chiari ed univoci** per valutare il contenuto di un testo. È un compito soggettivo, fortemente influenzato da esperienze personali, pensieri e credenze. Utilizzando un sistema standardizzato di analisi, è possibile applicare gli stessi criteri a tutti i dati.

Anche la Banca d'Italia e altre Banche Centrali Europee stanno impiegando questo tipo di analisi per

"prevedere" gli indicatori di stabilità politico-monetaria, fino ad ora studiati attraverso complessi modelli econometrici.

Il processo utilizzato allo scopo si chiama **Text Mining,** che raggruppa tutte le tecnologie grazie alle quali è possibile trarre informazioni da documenti non strutturati.
Sono oggi **disponibili parecchi software anche gratuiti** per questo processo, però di uso non immediato.
Ad esempio, NLTK è un potente pacchetto Python che fornisce un insieme di diversi algoritmi di linguaggi naturali.
L'obiettivo principale è quello di riportare le informazioni in modo strutturato. Questa tecnica trova numerose applicazioni, tra le più comuni c'è la creazione di database quantitativi.
Nell'ottica delle tecniche utilizzate dobbiamo considerare che in un insieme di documenti **non tutti potrebbero avere lo stesso formato**. Potremmo trovarci di fronte a documenti scritti e salvati in un formato proprietario (Word, Excel...), in semplice formato testo o codice ASCII, documenti scannerizzati e quindi sotto forma di immagini ecc…

La capacità di identificare le parti di un documento risiede nella possibilità di selezionarle.
Bisogna partire con una elaborazione del testo per renderlo analizzabile. Queste operazioni di pulizia servono a rimuovere le anomalie per poter sintetizzare i concetti e ridurre le dimensioni.
Una doverosa precisazione: **i sistemi più avanzati di Mining sono riferiti alla lingua inglese,** quindi i software sono molto più efficienti per questa lingua.
Per l'italiano esistono, ma sono un po' meno efficienti: il più noto è lo Snowball (https://snowballstem.org/algorithms/).
Nel giudicarne la funzionalità dobbiamo però considerare la complessità morfologica dell'italiano.
Prima di poter effettuare il Mining di un documento, **occorre passare per una serie di fasi** che portano da un documento non strutturato ad uno strutturato.
Le fasi principali del processo sono: **Tokenizzazione**, **Filtraggio**, **Lemmatizzazione**, **Stemming**, **Pattern Recognition**, **Part-of-Speech Tagging, Document-term Matrix.**

Tokenizzazione
È il primo passo nell'analisi dei testi ed è il processo di scomposizione di un paragrafo di testo in pezzi più piccoli, come parole o frasi. Si parte con l'isolare le singole frasi, per poi separare le varie parole. Si elimina il rumore, inteso come caratteri speciali, punteggiatura, numeri. Occorre inoltre prestare attenzione alle lettere minuscole e maiuscole, normalmente riportando tutto in minuscolo.
Questi compiti, banali per gli umani, non sono immediati per un computer: ad esempio la stessa parola in minuscolo e maiuscolo potrebbe essere considerata come due parole diverse.

Filtraggio
Si passa poi al **Filtraggio,** con l'eliminazione delle"**Stop words**", parole che, data la loro elevata frequenza in una lingua, sono di solito ritenute poco significative. Fra queste si trovano articoli, preposizioni e congiunzioni, parole generiche o verbi molto diffusi. Non esiste un elenco ufficiale delle Stop Words italiane ed ogni software ha il suo proprio elenco. Come curiosità, le Stop Words vengono utilizzate nello smascherare testi copiati ed incollati, che vengono facilmente individuati e scartati.
Vi sono però casi in cui la rimozione delle Stop words può portare a difficoltà nella comprensione del testo, quindi va valutata accuratamente.

Lemmatizzazione
I metodi di lemmatizzazione si propongono di mappare le forme verbali alla loro forma infinita e i sostantivi alla loro forma singolare, eseguendo un'analisi morfologica completa per individuare più accuratamente la radice: il "**lemma**". L'obiettivo è quello di raggruppare le varie forme flesse di una parola in modo che possano essere analizzate come una sola

entità.
È necessario che la forma di ogni parola sia nota, quindi ogni termine sia riconosciuto come verbo oppure sostantivo. Essendo questo processo di riconoscimento delle parti del discorso pesante, sia per il tempo di esecuzione, sia per la risoluzione degli errori, per facilitare le cose vengono applicati metodi di stemming, anche se meno precisi.

Stemming
Lo **Stemming** è il processo di riduzione delle parole alla loro radice.
Lo scopo principale è quello di ridurre drasticamente il numero di caratteri da trattare. Questa riduzione ha un doppio effetto sull'analisi del testo: il minor numero di caratteri agevola la successiva fase di addestramento e accorpando parole con radice simile -e quindi presumibilmente con significato simile- può risultare maggiormente affidabile un'analisi statistica.
Non è semplice realizzare un software che riesca in modo corretto ad effettuare lo stemming di un documento, in quanto fortemente dipendente dalla lingua e dal contesto. Esempi di stemming, considerando la lingua inglese, si hanno eliminando la "s" finale dai nomi, il suffisso "ing" dai verbi, ecc.

Pattern recognition
Una "**Bag of Words**" (borsa di parole, BoW) descrive la presenza delle parole all'interno di un documento. Comprende due cose: un vocabolario delle parole conosciute e un conteggio delle stesse.

Si chiama "borsa" di parole, perché ogni informazione sull'ordine o la struttura delle parole nel documento viene scartata. Il modello riguarda solo la presenza o meno di parole note nel documento, non la posizione.
Attraverso la"Bag of Words", la rappresentazione di un oggetto si ottiene contando il numero di occorrenze di ogni parola nell'oggetto.
Il modello "BoW" è utilizzabile per estrarre funzionalità dal testo da algoritmi di apprendimento automatico.

Part-of-Speech Tagging
La fase successiva è chiamata "**Part-Of-Speech Tagging**" o "**POS-tagging**", in italiano la traduzione potrebbe essere "etichettatura delle parti del discorso".
In questa fase, si etichettano le parole individuate nella fase di analisi lessicale, senza eseguire una vera e propria analisi sintattica, ma ricorrendo in genere ad informazioni statistiche o a regole, associando ad ogni parola tutte le possibili alternative lessicali e grammaticali, provvedendo poi alla disambiguazione dei casi ambigui.

Creazione della Document-term Matrix
Giunti a questo punto, si ha una collezione di elementi (stilemi) che rappresentano l'informazione estratta dai testi. Si procede costruendo una "**Document-term Matrix"**, una matrice matematica che descrive la frequenza dei termini che ricorrono in una raccolta di documenti.
Nella matrice, le righe corrispondono ai documenti rielaborati e le colonne corrispondono ai termini.
La matrice è utilizzabile come una comune matrice di regressione.

Si hanno dunque tutte le componenti per la stima di un modello di classificazione supervisionata che sia in grado di prevedere le classi delle variabili basandosi sulle informazioni estratte dai testi tramite **algoritmi di Machine Learning**.
Si usa normalmente il metodo "supervisionato", che è la metodologia per ricavare dai dati linguistici il risultato atteso, partendo da un modello generale ("train") che "allena" il sistema e sottoponendogli poi i nuovi dati ("test"), a cui viene applicato il modello appreso.
Da tutto ciò, otteniamo documenti statistici che riescono a rilevare in maniera abbastanza affidabile le diverse espressioni e quindi i relativi contenuti.
Queste indagini hanno bisogno di vaste quantità di documenti per evitare distorsioni nella valutazione, infatti vengono prevalentemente utilizzate da grandi aziende o da istituti specializzati.

Saranno gli algoritmi a sostituire i sondaggisti?
Prevedere l'esito delle elezioni sta diventando sempre più difficile, ma con la tecnologia si possono capire le intenzioni di voto anche solo analizzando le conversazioni sui social media.
Nelle ultime elezioni americane, alcuni metodi basati sull'intelligenza artificiale e sull'analisi dei contenuti postati sui social media sono stati in grado di scattare una fotografia elettorale aderente alla realtà; potrebbero quindi in futuro almeno integrare i sondaggi tradizionali.

Le immagini contenute nell'articolo sono tratte da:
https://ecommercefastlane.com/,
https://www.cogitotech.com/,
https://www.extrasys.it/it/red,
https://laptrinhx.com