17 FEB 2016

BIG DATA: UNA QUESTIONE DI METODO

Scritto da A.G..

Fonti, analisi, integrazione. La ricerca sui big data, quelli veri, si trova ancora a confrontarsi con aspetti chiave della propria metodologia. Lo si è visto in una recente presentazione tenuta da Paola Cerchiello lo scorso 4 febbraio all'Università di Pavia, in occasione del lancio del gruppo di ricerca su "Analisi dei big data in economia e finanza", coordinato dal Prof. Paolo Giudici: l'occasione era raccontare i risultati, apparsi su una rivista scientifica internazionale, di uno studio sulla rilevazione e l'utilizzo delle informazioni finanziarie "pubblicate" su Twitter.

IL NODO DELLA METODOLOGIA

Risultati a cui è stato dedicato un tempo tutto sommato marginale, alla fine dell'incontro. Il pubblico era infatti composto in stragrande maggioranza di ricercatori e addetti ai lavori, parte di quel nascente ecosistema di startup e innovazione in costruzione a Pavia. E quindi le due ore di vivace confronto si sono concentrate non sui risultati ma su un altro aspetto, più rilevante in questa fase: il metodo di ricerca.

PRIMO ASPETTO: LE FONTI

Perché sembra facile, parlare di big data e di Twitter: una quantità sconfinata di cinguettii e dati (informazioni, forse, meno) su praticamente ogni aspetto, dal quotidiano delle persone agli scandali finanziari. E allora il primo nodo da sciogliere per i ricercatori è stato schiettamente metodologico: come li seleziono, questi big data? Come scelgo le fonti rilevanti per l'analisi, in un mare di tweet? Come raccontato dalla dottoressa Paola Cerchiello, ricercatrice all'Università di Pavia, si è cercato un parametro per valutare la rilevanza dei twittatori. Una fonte individuata in una lista di profili Twitter segnalati dal Financial Times come particolarmente autorevoli sulle tematiche finanziarie.

COME VALUTARE L'AUTOREVOLEZZA

E già su questo primo punto la sala ha rumoreggiato. La classifica del Financial Times, autorevolezza della testata a parte, è non solo una fonte esterna, ma ha anche adottato come criterio di selezione il numero di follower di ciascun profilo. Troppo poco, forse, anche considerando il fatto che nel frattempo alcuni di quei profili sono stati disattivati o hanno smesso di produrre contenuto rilevante. Comunque, a partire dalla classifica del FT sono stati elaborati ulteriori parametri per valutare la rilevanza di un follower: in estrema sintesi, in analogia con la valutazione dell'indice di impatto utilizzato per le pubblicazioni scientifiche (ben nota al pubblico dei ricercatori), viene valutato il numero di re-tweet (le "citazioni") dei contenuti (le "pubblicazioni") di ciascun utente.

COME ANALIZZARE IL CONTENUTO

Seconda questione: l'analisi del contenuto. E qui emerge tutta la multidisciplinarietà dell'analisi dei big data: dopo le competenze statistiche necessarie a elaborare un "indice di impatto" dei twittatori, si passa a dimensioni fortemente linguistico-testuale. Come valutare, ad esempio, il "sentiment" (cioè il valore positivo, negativo o neutro) di ciascun post e di ogni parola pubblicata? Come giudicare la coerenza di ogni tweet con l'oggetto dell'analisi? Un lavoro reso più complicato dalla mancanza di una lista di parole rilevanti fondamentale nei problemi di classificazione testuale. L'oggetto dell'analisi (la reazione dei tweeter alla riforma delle Banche popolari; per la cronaca, ha prevalso il sentiment negativo) come detto è stato poco discusso.

INTEGRARE BIG DATA E INFORMAZIONI GIÀ DISPONIBILI

Più rilevante che si sia arrivati a definire mappe che rappresentano le occorrenze delle parole frequenti, individuando cluster di termini che ricorrono insieme e definendo sentiment precisi per ciascun cluster. Perché a partire da queste mappe si è toccato il tema dell'integrazione. Confrontando l'andamento del sentiment con quello delle azioni delle banche (quelle quotate, ovviamente). E individuando correlazioni tra diverse banche: come in un classico modello di "contagio", eventi negativi riferiti a una banca hanno un impatto negativo (in termini di tweet e valore in borsa) su una seconda banca o più. Da qui il tentativo di aggiustare il rischio di default di ciascuna banca (rappresentato dalla valutazione di una nota società di rating) in base alla correlazione di ogni singola banca alle altre, tenendo conto dell'effetto a cascata di eventuali eventi negativi.

IL BANKING È INFORMAZIONE

Come detto, conta il metodo, non i risultati. «Il filone di ricerca in corso a Pavia – commenta **Mario Bonelli, Presidente di RES** – è collegato alla nostra iniziativa RES Academy, che unisce l'attività di ricerca in ambito big data al sostegno a start up innovative nel settore. Con l'Università di Pavia e il gruppo di lavoro del Prof. Giudici si è creato sin da subito un'intesa particolare. Ci abbiamo creduto come RES e ci abbiamo investito come RES Academy. Oggi, a poco più di sei mesi dall'inizio della collaborazione, possiamo già vedere i frutti di questa proficua collaborazione. Il dottorando che sovvenzioniamo con una borsa di studio di tre anni, Gancarlo Nicola, insieme al team di R&D di RES, sta già realizzando un prototipo per l'analisi testuale dei big data in ambito finanziario che presto potrà diventare un prodotto licenziabile. Sono convinto che questa sia la strada giusta per tutti coloro che vogliano creare un futuro tangibile e duraturo, sia che si tratti di un'impresa, sia che si tratti di un giovane che vuole entrare nel mondo del lavoro».

«L'intera attività bancaria si basa sulle informazioni – aggiunge **Enzo Rocca, Vice Direttore Generale di Credito Valtellinese** – dall'erogazione del credito al denaro stesso, trasformato in dati. Guardiamo quindi con grande interesse a queste attività di ricerca: come nel caso dei rating, ci vorranno anche persone con competenze integrate di informatica, statistica ed economia e finanza, per guardare ai dati e valutare le opportunità enormi che racchiudono».

Le ultime notizie su: Big Data, RES Group