Che cos’è la stilometria

Qualche settimana fa, il *New York Times* ha pubblicato un articolo in cui riporta che due gruppi separati di informatici hanno trovato prove più che convincenti sull’identità di “Q”, l’autore anonimo di un post pubblicato nel 2017 sulla piattaforma per la condivisione di immagini anonime 4chan, poi diventato il leader della teoria del complotto conosciuta come QAnon, popolarissima nell’estrema destra americana.

Nell’articolo si parla in realtà di due autori (uno sarebbe subentrato all’altro a un certo punto) i cui nomi stavano già circolando da tempo: Paul Furber, informatico, giornalista sudafricano e uno dei primi commentatori online a richiamare l’attenzione sui primi messaggi di “Q”; e Ron Watkins, oggi candidato dei Repubblicani in Arizona per le elezioni di metà mandato di fine 2022 e sul quale si erano fino ad ora concentrati i maggiori sospetti.

Gli studi, scrive il *New York Times*, hanno fornito le prime prove empiriche a conferma di quelle che fino ad ora erano solo ipotesi. E le hanno fornite attraverso la stilometria, cioè l’analisi di un testo per determinarne la paternità o la maternità in mancanza di indizi certi. La stilometria ha una sua storia, sebbene piuttosto recente, e trova applicazione in ambito accademico e letterario, ma anche in quello legale e forense.

Per attribuire un testo a un autore o a un’autrice esistono diversi metodi che, come ha sintetizzato in un articolo di qualche anno fa per Treccani il linguista Michele Cortelazzo, fanno riferimento direttamente al testo o ricorrono a dati che non riguardano il suo contenuto. Tra questi ultimi possono esserci la data di ritrovamento dell’opera, l’analisi della grafia in caso di manoscritti, l’individuazione di coincidenze di nomi, luoghi, tempi e temi tra l’opera in esame e le altre

opere del presunto autore, l'analisi approfondita della casa editrice o delle circostanze in cui è stato pubblicato, e così via. Quando però le informazioni di questo genere sono assenti, resta solo il testo e lo stile con cui è stato composto.

Lo stile – che è il risultato delle scelte ricorrenti e più o meno consapevoli fatte da chi scrive – può essere analizzato con una metodologia qualitativa, e dunque tramite la stilistica, oppure quantitativa, tramite la stilometria. La stilometria ha iniziato a diffondersi dalla seconda metà dell'Ottocento interessando non soltanto studiosi di ambito letterario ma anche matematici, statistici e informatici. Inizialmente, i metodi di attribuzione quantitativa venivano considerati come poco credibili, ma nel tempo sono stati migliorati e in certi casi ritenuti così affidabili da poter essere usati come prove anche in tribunale.

Stilometria significa "misurazione dello stile", o meglio, come precisa Cortellazzo, misurazione della «similarità» quantitativa esistente tra due o più testi.

Parte dal presupposto che ogni autore o ogni autrice abbia delle caratteristiche peculiari, dei tratti costitutivi, un complesso di abitudini che costituiscono una specie di "DNA autoriale", il cosiddetto *idioletto*, e che non sono per lui o per lei del tutto manipolabili: agiscono cioè a livello inconscio. Con la stilometria questi indicatori stilistici, che rendono lo stile di qualcuno diverso da quello di qualcun altro, possono essere quantificati.

Le basi della stilometria furono stabilite dal filosofo polacco Wincenty Lutosławski, che a fine Ottocento disse di aver utilizzato questo metodo per stabilire la cronologia dei dialoghi di Platone: una questione che era stata a lungo oggetto di speculazioni e analisi, che però fino a quel momento si erano basate soprattutto sulle idee che il filosofo aveva sviluppato nei suoi testi (e dunque sul contenuto, non

sullo stile con cui erano stati scritti). Da Lutosławski in poi, molti studiosi migliorarono e crearono altre teorie legate a questo metodo, ma lo studio più significativo – che aiuta anche a capire in che cosa consista la stilometria – fu realizzato negli anni Sessanta da due statistici americani: Frederick Mosteller e David Wallace.

Mosteller e Wallace cercarono di comprendere, attraverso l'uso di metodi statistici, chi potessero essere gli autori dei *Federalist Papers*, una raccolta di ottantacinque articoli e saggi pubblicati nel Diciottesimo secolo con lo scopo di convincere i membri dell'assemblea dello Stato di New York a ratificare la Costituzione degli Stati Uniti d'America. Tutti gli scritti erano firmati con lo pseudonimo "Publius", e anche se era noto che fossero stati scritti da Alexander Hamilton, John Jay e James Madison, dodici degli ottantacinque articoli vennero attribuiti sia a Madison che ad Hamilton. Mosteller e Wallace riuscirono a rintracciare l'autore dei dodici articoli contesi.

Per farlo, si resero conto che l'uso di parole come "war" non era utile a distinguere tra i possibili autori, mentre considerando le parole "while" o "upon" diventava più chiaro come uno dei due amasse usarle e un altro no. Ad esempio, la parola "upon" compariva con una media di 3,24 ogni 1.000 parole negli articoli di Hamilton, ma solo di 0,23 negli articoli di Madison. Basandosi dunque su un'analisi esclusivamente quantitativa di questo tipo arrivarono ad attribuire i dodici articoli a Madison: attribuzione che trovò successivamente pieno consenso fra gli storici.

Nella loro ricerca, Mosteller e Wallace considerarono il testo come un insieme di parole in cui ciascuna aveva una frequenza di occorrenza e studiarono tale occorrenza non attraverso le parole-contenuto ad alto contenuto semantico (come nomi, verbi o aggettivi), ma attraverso le parole-funzione: parole cioè a basso contenuto semantico, appartenenti a una classe chiusa, che non veicolano propriamente un significato, ma che hanno

principalmente una funzione sintattica e grammaticale.

Secondo Mosteller e Wallace sono parole-funzione le preposizioni, le congiunzioni, i pronomi, gli articoli: rivelano le relazioni strutturali all'interno di una frase, sono indipendenti dall'argomento, sembra che siano utilizzate inconsciamente e dato che sono difficili da percepire sono anche più difficili da falsare.

Le parole ad alto contenuto semantico spesso dipendono e sono determinate dall'argomento del testo. Le parole-funzione, invece, sono indipendenti dall'argomento trattato.

In una serie di articoli di autrici o autori vari che si occupano di interruzione di gravidanza, ad esempio, sarà molto probabile trovare parole come "scelta", "diritto", "autodeterminazione", "obiettori", verbi come "decidere", "abortire" e aggettivi come "chirurgico", "volontario", "farmacologico" e così via. Identificare e quantificare quelle parole, che sono parole-contenuto, potrebbe non essere molto utile per stabilire chi abbia scritto quello specifico testo, soprattutto se i testi noti a disposizione per il confronto non si occupano dello stesso argomento. Si è dunque scoperto che possono essere molto più utili le parole che apparentemente possono apparire meno significative, perché sono indipendenti dal contenuto.

Inoltre, queste parole vengono spesso utilizzate in maniera inconsapevole.

Commentando gli studi di stilometria per scoprire o confermare l'identità di "Q", uno degli autori identificati, Paul Furber, ha giustificato la somiglianza tra il proprio stile e quello dei post di "Q" dicendo che i messaggi di quest'ultimo «hanno preso il sopravvento sulle loro vite» e che «tutti, lui compreso, hanno iniziato a parlare come lui». Jean-Baptiste Camps, uno degli studiosi che hanno condotto analisi su "Q", ha spiegato che essere influenzati da uno stile è certamente

possibile, ma che l’uso delle parole-funzione si costruisce attraverso tutta una vita, fin da quando si impara a parlare: si tratta dunque di qualcosa di molto profondo e unico che, pertanto, è molto complicato sia da acquisire che da dissimulare.

La stilometria, che nel tempo ha generato diverse teorie e metodi, può considerare un testo non solo come un insieme di parole, ma anche come una peculiare unità di misura: gli n-grammi.

Gli n-grammi non corrispondono a nessun criterio di divisione tradizionale del testo: non tengono conto, ad esempio, di quando una parola inizia o finisce, né considerano la parola come gerarchicamente superiore rispetto ai singoli segni. Non corrispondono nemmeno a una convenzionale unità di misura: possono essere composti da caratteri, lettere o sillabe. A seconda del contesto, negli n-grammi rientrano non solo le lettere ma anche gli spazi e i segni di interpunzione. A n può essere assegnato un qualsiasi valore a discrezione di chi compie lo studio e delle sue specifiche valutazioni.

Poniamo che nella nostra analisi consideriamo n-grammi i singoli caratteri di una parola, ma anche gli spazi e i segni di interpunzione fra loro. Un esempio di divisione in 8-grammi (segnati di seguito in grassetto) della frase “la stilometria, spiegata bene” può essere,

“**la stilo**metria, spiegata bene”

“l**a stilom**etria, spiegata bene”

“la **stilomet**ria, spiegata bene”

“la s**tilometr**ia, spiegata bene”

“la st**ilometri**a, spiegata bene”

“la sti**lometria**, spiegata bene” e così via.

Contando quante volte ciascun n-gramma compare in un testo viene costruito un indicatore che esprime numericamente la distanza o la similarità tra due testi.

In uno degli studi fatti per individuare l’identità di “Q”, oltre alle parole-funzione sono stati ad esempio considerati alcuni particolari 3-grammi di caratteri. Jean-Baptiste Camps ha spiegato che i 3-grammi «sono ampiamente utilizzati negli studi statistici, anche di più delle sole parole-funzione, perché hanno la capacità di catturare i morfemi grammaticali», cioè l’elemento più piccolo di una parola che contiene del significato. Nella parola “casa” ci sono per esempio due morfemi: cas-, una radice che trasmette l’idea di “casa”, e -a, che ne veicola il genere femminile.

Il problema della stilometria, fino a poco tempo fa, era che questo lavoro di misurazione era estremamente faticoso, perché veniva fatto a mano. Ma a partire dalla metà degli anni Novanta i metodi sono stati migliorati: non solo perché si sono cominciate ad avere a disposizione enormi quantità di testi in formato elettronico, ma anche per lo sviluppo dell’informatica che ha permesso di lavorare su enormi volumi di dati in tempi rapidi.

La stilometria ha avuto nel tempo, e continua ad avere, diversi ambiti di applicazione. Come ricorda il *New York Times*, viene spesso citato il caso dell’attentatore soprannominato Unabomber – l’FBI utilizzò una forma di stilometria per dimostrare che si trattava di Ted Kaczynski – ma anche il caso letterario del poliziesco *The Cuckoo’s Calling*, firmato da un certo Robert Galbraith.

Quando il libro venne pubblicato, nell’aprile del 2013, si sapeva che Robert Galbraith era uno pseudonimo. Un giornalista del *Sunday Times* – edizione domenicale del quotidiano britannico – ricevette un suggerimento iniziale sulla reale

identità di chi avesse scritto il romanzo e cioè che fosse in realtà la scrittrice J.K. Rowling, autrice della saga di Harry Potter, e contattò il professor Patrick Juola, informatico, esperto di analisi dei testi e di stilometria, per avere delle conferme.

Juola utilizzò il Java Graphical Authorship Attribution Program (JGAAP), un programma da lui sviluppato che attraverso un'analisi di tipo matematico è in grado di trovare la similarità fra più autori o autrici analizzando una grande quantità di parole e caratteristiche del testo. Juola paragonò così *The Cuckoo's Calling* con l'ultimo romanzo scritto da J.K. Rowling, *The Casual Vacancy*, e con altri libri di gialliste britanniche contemporanee. Fu eseguita l'analisi delle parole-funzione e furono condotti due test sugli n-grammi. Infine, incrociando i risultati, fu elaborato un indice di similarità.

Rowling compariva in tutti i risultati, risultando sempre come l'autrice più probabile o come seconda possibilità. Juola spiegò come il risultato non fosse la prova definitiva del fatto che Rowling avesse scritto il romanzo, ma disse che dall'analisi risultava la candidata più probabile. Il 14 luglio 2013, comunque, il *Times* pubblicò la notizia che la vera autrice di *The Cuckoo's Calling* era J.K. Rowling, che a sua volta lo ammise.

Quello dell'analisi quantitativa di un testo è un ambito in cui la ricerca e il dibattito sono ancora in corso e tra i linguisti sembra prevalere la tesi che l'uso di una valutazione quantitativa nei casi di attribuzione d'autore abbia valore soprattutto in un'ottica confermativa: quando cioè esiste già un sospettato.

Per quanto riguarda gli studi su "Q", uno è stato condotto dalla società svizzera OrphAnalytics, e l'altro da due linguisti computazionali francesi. Entrambe le analisi hanno trovato conferme che gli stili di scrittura siano quelli di Furber e Watkins, riuscendo anche a stabilire il passaggio

della gestione dei post dall’uno all’altro, e con un grado di affidabilità che va dal 93 al 99 per cento. Gli esperti che ha intervistato il *New York Times* hanno detto a loro volta di aver trovato entrambi i risultati credibili e convincenti.

Read More